Decoupled and Memory-Reinforced Networks: Towards Effective Feature Learning for One-Step Person Search

Chuchu Han[1], Zhedong Zheng[2,3], Changxin Gao[1*], Nong Sang[1], Yi Yang[2]

[1] Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation,
Huazhong University of Science and Technology, Wuhan, China
[2] ReLER, University of Technology Sydney, Australia [3] Baidu Research, China
{hcc, cgao, nsang}@hust.edu.cn, zhedong.zheng@student.uts.edu.au, yi.yang@uts.edu.au

Abstract

The goal of person search is to localize and match query persons from scene images. For high efficiency, one-step methods have been developed to jointly handle the pedestrian detection and identification sub-tasks using a single network. There are two major challenges in the current one-step approaches. One is the mutual interference between the optimization objectives of multiple sub-tasks. The other is the sub-optimal identification feature learning caused by small batch size when end-to-end training. To overcome these problems, we propose a **d**ecoupled and **m**emory-**r**einforced network (DMRNet). Specifically, to reconcile the conflicts of multiple objectives, we simplify the standard tightly coupled pipelines and establish a deeply decoupled multi-task learning framework. Further, we build a memory-reinforced mechanism to boost the identification feature learning. By queuing the identification features of recently accessed instances into a memory bank, the mechanism augments the similarity pair construction for pairwise metric learning. For better encoding consistency of the stored features, a slow-moving average of the network is applied for extracting these features. In this way, the dual networks reinforce each other and converge to robust solution states. Experimentally, the proposed method obtains 93.2% and 46.9% mAP on CUHK-SYSU and PRW datasets, which exceeds all the existing one-step methods.

Introduction

Person search aims at localizing and identifying a query person from a gallery of uncropped scene images. This task is generally decomposed into two sub-tasks, *i.e.*, pedestrian detection, and person re-identification (re-ID) (Zheng et al. 2019). Based on this, two-step and one-step methods have been developed. Two-step methods sequentially process the sub-tasks with two separate networks, where a detector is applied to raw images for localization and a followed re-ID network extracts identification features from the detected person images (Zheng et al. 2017; Lan, Zhu, and Gong 2018; Chen et al. 2018; Han et al. 2019; Chang et al. 2018; Wang et al. 2020). In contrast, one-step methods learn person localization and identification in parallel within a single network, exhibiting higher efficiency (Xiao et al. 2017, 2019; Munjal et al. 2019; Yan et al. 2019; Dong et al. 2020b,a; Chen et al. 2020; Zhong, Wang, and Zhang 2020). Given an uncropped input image, one-step models predict the bounding boxes and the corresponding identification features of all the detected persons.

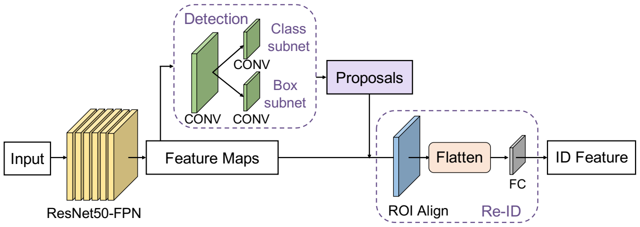


Figure 1: The inference of the proposed one-step framework.

Although significant progress has been made in the one-step person search, there are two crucial issues that have not been fully solved by previous works. The first issue is that coupling the two sub-tasks in a shared network may be detrimental to the learning of each task. Specifically, popular one-step methods based on the Faster R-CNN (Ren et al. 2015) supervise the shared Region-of-Interest (RoI) features with multi-task losses, *i.e.*, regression loss, foreground-background classification loss, and identification loss. The competing objectives of these sub-tasks make the RoI features difficult to optimize, as pointed in (Chen et al. 2018). The second issue lies in the constrained small batch size under the end-to-end fashion, caused by limited GPU memory. It leads to sub-optimal identification feature learning since metric learning requires vast informative similarity pairs. Previous works tackle this issue by maintaining an exponential moving average (EMA) feature proxy for every identity, *i.e.*, a look-up table. However, when an identity is infrequently visited, its feature proxy could be outdated as the weights of the model evolve. It is unclear that this strategy could be scaled to larger datasets with numerous identities.

In the paper, we rethink the decoupling and integration of pedestrian detection and identification in the one-step person search framework. Considering that RoI features contain the detailed recognition patterns of detected persons, they can be specific to the re-ID task. On the other hand, bounding box regression and foreground-background classification do not have to rely on the fine-grained RoI features

[*] Corresponding author: cgao@hust.edu.cn.

in light of the success of one-stage detectors. Based on these insights, we take the one-stage detector as our base network instead. As shown Fig. 1, foreground-background classification, regression, and re-ID subnets are branched from the layers of the feature pyramid network (FPN), which contain rich visual information and could burden multiple types of task-specific feature encoding. The fine-grained RoI features extracted from FPN are only fed into the re-ID subnet for transformation. We demonstrate that this new design makes the two sub-tasks substantially decoupled and facilitate the learning for both tasks. Specifically, the decoupled network with RetinaNet backbone (Lin et al. 2017b) achieves 6.0% improvements on mAP compared to the popular baseline with Faster R-CNN.

To further boost the identification feature learning, we build a memory-reinforced feature learning mechanism. Inspired by the recent unsupervised contrastive learning study (He et al. 2020), we memorize the feature embeddings of the recently visited instances in a queue-style memory bank for augmenting pairwise metric learning. The memorized features are consistently encoded by a slow-moving average of the network and are stored in a queue-style bank. The dual networks reinforce each other and converge to robust solution states. Experimental evidence proves that our mechanism is more effective than the look-up table.

The resulting model is called decoupled and memory-reinforced network (DMRNet). Our network is easy to train because of the task decoupling in the architecture. The inference of our framework (shown in Fig. 1) is also very simple and efficient. In experiments, we validate the effectiveness of our methods on different one-stage detectors. Our DMRNet surpasses the state-of-the-art one-step method (Chen et al. 2020) by 1.1% and 2.9% mAP on the CUHK-SYSU and PRW datasets, respectively.

Our contributions can be summarized in three-folds:

- We propose a simplified one-step framework that decouples the optimization of pedestrian detection and identification. In particular, RoI features are only specific to the re-ID task, promoting the performance of both sub-tasks.
- We introduce a memory-reinforced mechanism for effective identification learning. A slow-moving average of the network is incorporated for consistently encoding features in a queue-style memory bank. This reinforced training makes the identification features highly discriminative.
- Our model is easy to train and efficient to use. It surpasses the previously best one-step methods and matches the accuracy of two-step methods.

Related Work

Person search. Person search aims at matching a specific person among a great number of whole scene images, which has raised a lot of interest in the computer vision community recently (Xiao et al. 2017; Zheng et al. 2017; Chen et al. 2018; Lan, Zhu, and Gong 2018; Chang et al. 2018). In the literature, there are two approaches to deal with the problem.

Two-step methods (Zheng et al. 2017; Lan, Zhu, and Gong 2018; Chen et al. 2018; Han et al. 2019; Chang et al. 2018; Wang et al. 2020) separate the person search task into two sub-tasks, the pedestrian detection, and person re-ID, trained with two independent models. Zheng *et al.* (Zheng et al. 2017) first make a thorough evaluation on various combinations of different detectors and re-ID networks. Chen *et al.* (Chen et al. 2018) consider the contradictory objective problem existing in person search, and extract more representative features by a two-steam model. Han *et al.* (Han et al. 2019) develop an RoI transform layer that enables gradient backpropagated from re-ID network to the detector, obtaining more reliable bounding boxes with the localization refinement. Wang *et al.* (Wang et al. 2020) point out the consistency problem that the re-ID model trained with hand-drawn images are not available. They alleviate this issue by producing query-like bounding boxes as well as training with detected bounding boxes.

One-step methods (Xiao et al. 2017, 2019; Munjal et al. 2019; Yan et al. 2019; Dong et al. 2020b,a; Chen et al. 2020; Zhong, Wang, and Zhang 2020) develop a unified model to train the pedestrian detection and person re-ID end-to-end. Generally, this manner is more efficient with fewer parameters. Xiao *et al.* (Xiao et al. 2017) employ the Faster R-CNN as the detector, and share base layers with the person re-ID network. Meanwhile, an Online Instance Matching (OIM) loss is proposed to enable a better convergence with large but sparse identities in the classification task. To incorporate the query information into the detection network, Dong *et al.* (Dong et al. 2020a) propose a Siamese network that both takes scene images and cropped person patches as input. With the guidance of the cropped patches, the learned model can focus more on persons. As pointed out by (Chen et al. 2018), pedestrian detection focuses on learning the commonness of all persons while person re-ID aims to distinguish the differences among multiple identities. Chen (Chen et al. 2020) solves this problem by disintegrating the embeddings into norm and angle, which are used to measure the detection confidence and identity similarity. However, this method ignores the effect of regression loss, and excessive contexts still hamper the feature learning. Different from (Chen et al. 2020), we identify that the inherently defective module design is the core cause of the conflict and hinders effective feature learning.

Pedestrian detection. Pedestrian Detection plays a crucial role in the person search framework. In recent years, with the advent of Convolutional Neural Network (CNN), the object detection task is soon dominated by the CNN-based detectors, which can be broadly divided into two categories: the one-stage manner (Lin et al. 2017b; Redmon et al. 2016; Liu et al. 2016) and two-stage manner (Girshick 2015; Ren et al. 2015; Dai et al. 2016; He et al. 2017). Due to the high efficiency, the one-stage manner has attracted much more attention recently. YOLO (Redmon et al. 2016; Redmon and Farhadi 2017) directly detects objects though a single feed-forward network with extremely fast detection speed. RetinaNet (Lin et al. 2017b) solves the problem of class-imbalance by the focal loss, which focuses on learning hard examples and down-weight the numerous easy negatives. The two-stage manner is composed of a proposal generator and a region-wise prediction subnetwork ordinarily. Faster R-CNN (Ren et al. 2015) proposes a region proposal network (RPN). It greatly reduces the amount of computation while shares the characteristics of the backbone network. Lin *et al.* (Lin et al. 2017a) design a top-down architecture with lateral connections for building multi-level semantic feature maps at multiple scales, which is called Feature Pyramid Networks (FPN). Using FPN in a basic detection network can assist in detecting objects at different scales. Recent anchor-free detectors have raised more interest. FCOS (Tian et al. 2019) employs the center point of objects to define positives, then predict the four distances from positives to object boundary. Reppoints (Yang et al. 2019) first locate several self-learned keypoints and then predict the bound the spatial extend of objects. Without excessive hyper-parameters caused by anchors, these methods are more potential in terms of generalization ability.

Proposed Method

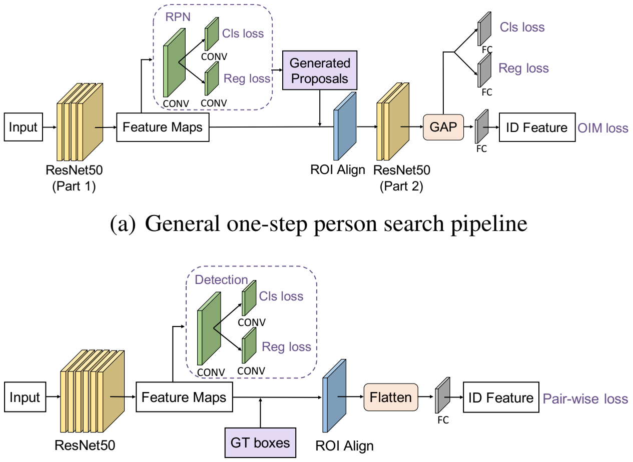
In this section, we describe the structure of the decoupled one-step person search network and present the memory-reinforced feature learning mechanism for identification.

Decoupled One-Step Framework

General one-step pipeline. The first and most representative framework for one-step person search is proposed by (Xiao et al. 2017), and it is widely adopted in the following research work (Xiao et al. 2019; Munjal et al. 2019; Yan et al. 2019; Dong et al. 2020b,a; Chen et al. 2020; Zhong, Wang, and Zhang 2020). This pipeline is based on a Faster R-CNN detector (Ren et al. 2015), as illustrated in Fig. 2(a). For the re-ID module, the features are supervised by OIM loss. Together with the detection losses in RPN head and RoI head, the whole network is trained end-to-end.

However, there exist contradictory objectives when supervising the shared RoI features with multi-task losses. For the person search task, the detector only requires to distinguish person or background, rather than the multi-classification task in object detection. Thus, the foreground-background classification loss in the RoI head is unnecessary, even seriously affect the optimization. Evidently, foreground-background classification pursues to learn the universality of all the persons while person re-ID aims at distinguishing different persons. Moreover, the regression loss requires more information around the box boundary, while excessive contexts harm the fine-grained features for identification.

Decoupled one-step pipeline. Although (Chen et al. 2020) reconciles the conflict by factorizing embeddings into magnitude and direction for foreground scoring and re-ID, respectively, we identify that the inherently defective module design is the core cause of this issue and hinders the effective feature learning of the one-step models.



(a) General one-step person search pipeline

(b) Our decoupled one-step person search pipeline

Figure 2: Comparisons between general training pipeline and ours. (a) General one-step person search pipeline. Multi-task losses are applied on shared RoI features. (b) Our decoupled one-step person search pipeline. The RoI features are specific to the re-ID task.

In this paper, we mainly focus on learning representative RoI features for identification, instead of the multi-task losses under a shared feature space. This decoupling is based on the following considerations. First, since the RoI features contain the detailed recognition patterns of detected persons, they can be specific to the re-ID task. Second, bounding box regression and foreground-background classification do not have to rely on the fine-grained RoI features in light of the success of one-stage detectors, *e.g.*, RetinaNet (Lin et al. 2017b), FCOS (Tian et al. 2019) and Reppoint (Yang et al. 2019). Based on some simplifications, we introduce the one-stage detector as our base network instead. Here we take the RetinaNet for example. As Fig. 2(b) shows, ResNet50 with a feature pyramid network (FPN) is used as the shared backbone. A class subnet and a box subnet based on FPN are employed to performs foreground-background classification and bounding box regression on each location. We add the RoI align on FPN to extract fine-grained embeddings for person re-ID. Since FPN layers include rich semantic information while RoI features contain specific content, this design makes the two sub-tasks substantially decoupled. Moreover, we only employ the ground truth bounding boxes to extract RoI features for re-ID training, without the usage of the predicted proposals from the regression subnet. This simplification further reduces dependencies between regression and identification. We experimentally show that using the largely reduced but accurate training bounding boxes could result in slightly better performance.

Memory-Reinforced Feature Learning

Effective feature learning is challenging for the one-step person search. Due to the limited batch size caused by GPU memory constraints in the end-to-end fashion, it may suffer from a large variance of gradients when directly use the softmax loss or triplet loss. Previous works (Xiao et al. 2017) use the Online Instance Matching (OIM) loss that maintains an EMA feature proxy for every identity, *i.e.*, a look-up table. Nevertheless, due to the limited batch size, the fea-

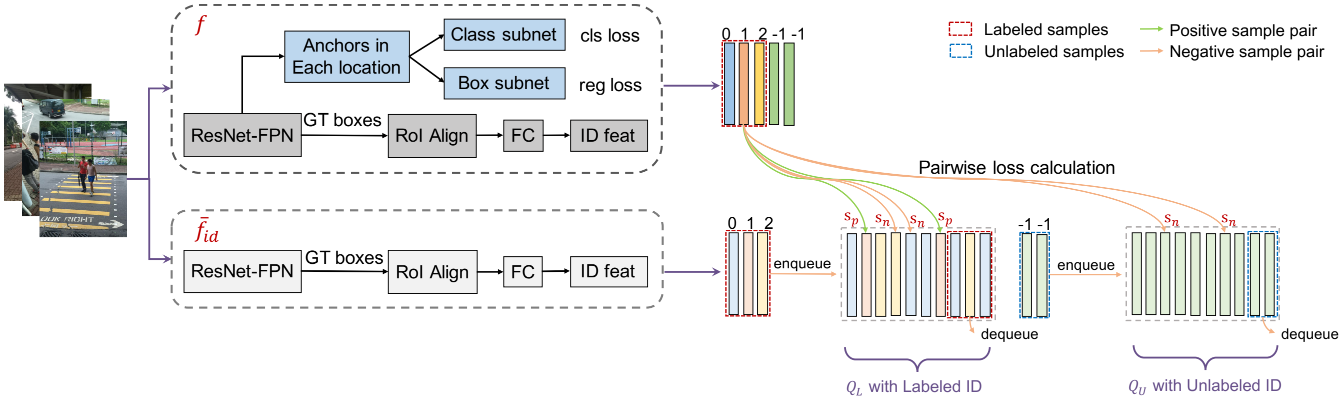


Figure 3: An overview of Decoupled and Memory-Reinforced Networks (DMRNet). $\mathbf{f}$ is our decoupled person search network that trained by a SGD optimizer. $\bar{\mathbf{f}}_{id}$ is a slowly-updating network counterpart, which is utilized to consistently encode the re-ID features in the training stage. Given an input image, $\mathbf{f}$ extracts the labeled pedestrian features, termed as anchors. The features extracted by $\bar{\mathbf{f}}_{id}$ are employed to update the labeled and unlabeled queues, respectively. Thus, multiple positive and negative similarity pairs can be built between anchors and queued embeddings, supervised by a pairwise loss.

ture proxy could be outdated as the weights of the model evolve. It is unclear that this strategy could be scaled to larger datasets with numerous identities. To keep the consistency of the comparing feature embeddings, we propose a memory-reinforced method for effective feature learning. Inspired by (He et al. 2020; Tarvainen and Valpola 2017), a slowly-updating network counterpart is incorporated for yielding a consistent queue-style feature memory bank.

Queue-style memory bank. Instead of keeping the class proxy embedding within a look-up table, we maintain a queue-style memory bank. It only keeps the features of recently visited instances, avoiding features being outdated. Moreover, it decouples the memory bank size from the number of identities. This is more flexible to set the size as a hyper-parameter.

An slow-moving average of the network. To make the stored features encoded more consistently, we introduce a slow-moving average of the network for generating features in the memory bank. We denote our decoupled network as $\mathbf{f}$, where its parameters θ are updated by the back-propagation. The slow-moving average of the network is denoted by $\bar{\mathbf{f}}_{id}$. Its parameters $\bar{\theta}$ are updated by EMA at each iteration:

$$\bar{\theta} \leftarrow m\bar{\theta} + (1-m)\theta, \tag{1}$$

where m is the momentum factor. With a large momentum, the parameters $\bar{\theta}$ are updated slowly towards θ, making little difference among encoders from different iterations. This ensures the consistency of the encoded features in the memory bank. Note that $\bar{\theta}$ is only used for extracting identification embeddings, without detection subnets. $\bar{\mathbf{f}}_{id}$ requires no gradient and brings little overhead at each iteration.

Pairwise loss for re-ID feature learning. We use a pairwise loss for supervising the re-ID feature learning. The foundation of pairwise loss is to construct positive and negative pairs for metric learning.

In this paper, we maintain a queue $Q_l \in \mathbb{R}^{L\times d}$ containing the features of L labeled persons, and a queue $Q_u \in \mathbb{R}^{U\times d}$ containing the features of U unlabeled persons, where d is the feature dimension. Suppose the input image contains one labeled person with class-id i and several unlabeled ones. The embedding of the labeled person encoded by $\mathbf{f}$ is viewed as an anchor x_a. The embeddings of labeled and unlabeled persons extracted by $\bar{\mathbf{f}}_{id}$ are used to update the Q_l and Q_u, respectively. As Fig.3 shows, these newest embeddings are enqueued while the outdated ones are dequeued, maintaining the queue length fixed. Assuming that there are K positive samples in Q_l sharing the same identity with x_a, and the rest J ones in Q_l and Q_u are viewed as negative samples, the cosine similarities are denoted as $\{s_p^i\}(i = 1, 2, ..., K)$ and $\{s_n^j\}(j = 1, 2, ..., J)$, respectively. To make every s_p^i is greater than every s_n^j, we utilize the following loss function (Sun et al. 2020):

$$L = log[1 + \sum_{i=1}^{K}\sum_{j=1}^{J} \exp(\gamma(s_n^j - s_p^i))] \tag{2}$$

where γ is a scale factor. We note that this loss formulation is the natural extension of OIM loss in the case of multiple positive similarity pairs. By the supervision of the pairwise loss, $\mathbf{f}$ and $\bar{\mathbf{f}}$ reinforce each other and their parameter spaces converge to robust solution states.

Experiments

In this section, we first describe the datasets and evaluation protocols, after which the implementation details are elaborated. Then, we conduct comprehensive ablation studies and analysis to explore the effects of different components. We further compare our method with state-of-the-art methods.

Datasets and Settings

CUHK-SYSU dataset. CUHK-SYSU (Xiao et al. 2017) is a large scale person search dataset consisting of street/urban scenes shot by a hand-held camera and snapshots chosen from movies. There are $18,184$ images and $96,143$ annotated bounding boxes, containing $8,432$ labeled identities, and the unlabeled ones are marked as unknown instances. The training set contains $11,206$ images and $5,532$ identities, while the testing set includes $6,978$ gallery images and $2,900$ probe images.

PRW dataset. PRW (Zheng et al. 2017) is extracted from the video frames that are captured by six spatially disjoint cameras. There are a total of $11,816$ frames with the $43,110$ annotated bounding boxes. Similar to CUHK-SYSU, it contains unlabeled identities and labeled identities ranged from 1 to 932. In training set, there are $5,704$ frames and 482 identities, while the testing set includes $6,112$ gallery images and $2,057$ query images from 450 different identities.

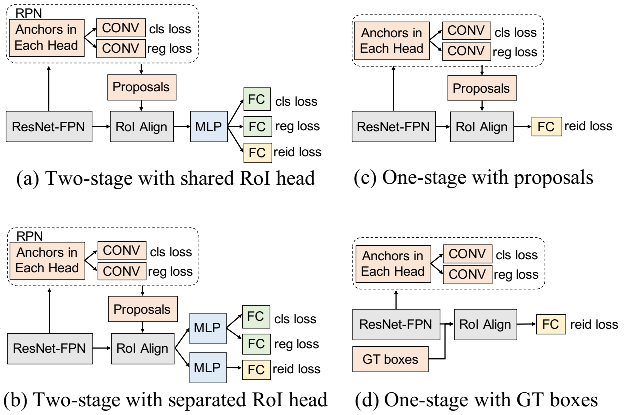
Evaluation protocols. Our experiments adopt the same evaluation metrics as previous work (Xiao et al. 2017; Munjal et al. 2019). One is widely used in person re-ID, namely the cumulative matching cure (CMC). A matching is considered correct only if the IoU between the ground truth bounding box and the matching box is larger than 0.5. The other is the mean Average Precision (mAP) inspired by the object detection task. For each query, we calculate an averaged precision (AP) by computing the area under the precision-recall curve. Then, the mAP is obtained by averaging the APs across all the queries.

Implementation Details

For the detection network, we use the latest PyTorch implementation of RetinaNet (Lin et al. 2017b) and Reppoint (Yang et al. 2019) released by OpenMMLab [1] (Chen et al. 2019). Actually, our framework is compatible with most detectors. The queue sizes L and U are set to 4096 and 4096 for CUHK-SYSU while 1024 and 0 for PRW. The momentum factor m is set to 0.999, and the scale factor γ is set to 16. The batch size is 3 due to the limitation of GPU memory. We use the batched Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9. The weight decay factor for L2 regularization is set to 5×10^{-4}. As for the learning rate strategy, we use a step decay learning rate schedule with a warm-up strategy, and our model is trained for 12 epochs totally. The base learning rate is 0, which warms up to 1×10^{-3} in the first 500 iterations, then decays to 1×10^{-4} and 1×10^{-5} after 8 and 11 epochs. All experiments are implemented on the PyTorch framework, and the network is trained on an NVIDIA GeForce GTX 1080 Ti. We also use PaddlePaddle to implement our method and achieve similar performance.

Ablation Study

In this section, we conduct detailed ablation studies to evaluate the effectiveness of each component. First, we explore the effect of different network designs. Second, we analyze two loss mechanisms under different sizes of memory banks. Third, we exhibit the performance of our proposed method under different settings.



(a) Two-stage with shared RoI head

(c) One-stage with proposals

(b) Two-stage with separated RoI head

(d) One-stage with GT boxes

Figure 4: Comparisons on different network designs. (a) is the general person search pipeline with a shared RoI head. (b) eases the coupling with separated RoI heads for detection and re-ID. (c) discards the detection losses in RoI head and the RoI features are specific for identification. (d) removes the selected proposals and only uses the GT boxes.

Why is one-step inferior than two-step? To investigate what causes the poor performance in one-step person search, we conduct several experiments to illustrate the comparisons among different network options, as shown in Fig. 4. Detailed results are shown in Tab. 1.

For fair comparisons, we incorporate FPN into the general one-step framework (Xiao et al. 2017) as our baseline (a), and this improves the performance by a large margin. When it comes to tangled sub-tasks (detection and re-ID) conflict in the one-step person search, it is natural to think about decoupling different tasks from the backbone. (b) employs separated RoI heads for detection and re-ID training. In Tab. 4, the results perform better than a shared RoI head manner on both re-ID and detection tasks. This indicates the severe coupling network harms the optimization on both sub-tasks when sharing feature space, and it can be mitigated with a simple head disentanglement.

To further eliminate the conflict, we only focus on identification feature learning instead of the multi-task loss under the shared RoI features. As presented in (c), a one-stage detector can be well incorporated and the RoI features are specific for person recognition. This manner surpasses the performance of (b) on both re-ID and detection tasks. It shows that the decoupling benefits the optimization on two sub-tasks. Note that the performance of separated trained detectors for one-stage (RetinaNet) or two-stage (Faster RCNN) is almost the same.

In (a)-(c), except for the ground truth boxes, the selected proposals (IoU>0.5) are also used to extract features for re-ID training. We further simplify the network by using only ground truth bounding boxes. Although the improvement is marginal, it saves much computational cost in training. Finally, based on our proposed memory-reinforced feature learning, the performance achieves 91.2%/92.5% on mAP/rank-1 on the CUHK-SYSU dateset.

[1] https://github.com/open-mmlab/mmdetection

Methods	Re-ID		Detect	Detect-S
	mAP	Rank-1	mAP	mAP
Faster+OIM	75.5	78.7	-	-
Faster(FPN) + OIM w/ (a)	84.3	84.6	86.9	92.2
Faster(FPN) + OIM w/ (b)	87.5	87.7	89.8	
Retina(FPN) + OIM w/ (c)	90.0	90.8	91.2	92.3
Retina(FPN) + OIM w/ (d)	90.3	91.0	91.4	
Retina(FPN) + DMR w/ (d)	91.2	92.5	91.3	

Table 1: Comparisons of different network designs on the CUHK-SYSU dataset. The performance of re-ID and detector trained in a single network is represented. Detector-S denotes the result of the separated trained detector.

Methods	Resolution	CUHK-SYSU		PRW	
		mAP	Rank-1	mAP	Rank-1
Retina+D	1333*800	90.3	91.0	36.1	73.6
Retina+DMR	1333*800	91.2	92.5	44.6	82.0
Retina+DMR	1500*900	91.6	93.0	46.1	83.2
Reppoint+D	1333*800	92.4	93.2	39.1	73.6
Reppoint+DMR	1333*800	92.9	93.7	46.0	83.2
Reppoint+DMR	1500*900	93.2	94.2	46.9	83.4

Table 2: The results on the CUHK-SYSU and PRW datasets with different detectors. D denotes the decoupled framework while DMR means our decouple and memory-reinforced network.

Effectiveness on different detectors. In order to evaluate the expandability of our method, we incorporate different detection networks into our framework, including RetinaNet (Lin et al. 2017b) and Reppoint (Yang et al. 2019). The separated trained detectors reach 92.3% and 93.1% on mAP, respectively. We show the person search results in Tab. 2 under different settings. When only perform the decoupled network, the results have achieved 90.3% and 92.4% rank-1 with RetinaNet and Reppoint, respectively. The performance is further promoted when employing the memory-reinforced method for training. This confirms the effectiveness and robustness of our method when extended to different detectors. Moreover, we show the experimental results under different resolutions. It is obvious that a larger image reaches higher performance.

Momentum factor. The performance of our method with different momentum factors is shown in Tab. 3. We obtain the optimal result when m is set to 0.999. This indicates a relatively large momentum facilitates learning discriminative identification features. When m is zero, it means the parameters of $\mathbf{f}$ and $\bar{\mathbf{f}}_{id}$ are identical. Surprisingly, with the least consistent encoding, our mechanism still slightly outperforms the look-up table by 0.3% mAP and 0.6% rank-1, showing the effectiveness of the queues.

m	0	0.5	0.9	0.99	0.999	0.9999
Rank-1	91.6	91.6	91.7	91.7	92.5	91.6
mAP	90.6	90.7	90.9	90.9	91.2	90.4

Table 3: The results with different momentum factors m on the CUHK-SYSU dataset.

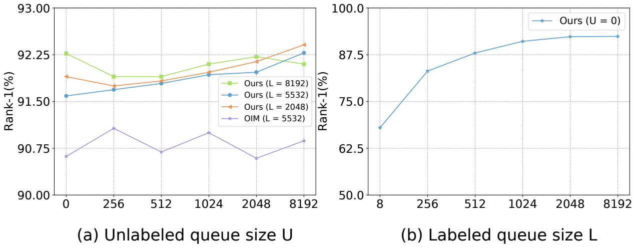


Figure 5: Comparison between OIM loss and our DMRNet with different sizes of memory bank. The numbers of labeled and unlabeled samples are denoted as L and U, respectively.

Different sizes of the memory bank. We analyze the effect of different memory bank sizes on two metric learning mechanisms, OIM loss and our memory-reinforced mechanism. They are implemented on the same network, as described in Fig. 4(d). L is the length of the look-up table or queue with labeled samples, and U is the length of the queue with unlabeled ones. The comparisons are shown in Fig. 5, from which we have the following observations.

- To explore the effect of unlabeled samples, we compare OIM (L=5532) with our method (L=2048/5532/8192) under different sizes of U. As shown in Fig. 5 (a), the performance of our method is constantly promoted as U increases when L=2048/5532. This shows that exploring more negative samples is better for optimization. The relatively large size of the labeled queue ($L = 8192$) cannot benefit from U. This is reasonable as a larger L has provided sufficient negative samples. For OIM loss, there is no significant improvement when U increases. Due to the lack of feature consistency, more sample pairs contribute little to the result.

- As Fig. 5(a) shows, when U is set to zero, our method benefits from a larger L. This is intuitive since more positive/negative sample pairs can be exploited.

- From Fig. 5(a)(b), it can be observed that when the two methods reach the same performance, our method is more efficient (L=2000, U=0) than OIM (L=5532, U=5000).

Comparisons with the State-of-the-Art Methods

In this section, we compare our proposed DMRNet with current state-of-the-art methods on person search in Tab. 4. The results of two-step methods (Chang et al. 2018; Chen et al. 2018; Lan, Zhu, and Gong 2018; Han et al. 2019; Wang et al. 2020) are shown in the upper block while the one-step methods (Xiao et al. 2017, 2019; Liu et al. 2017; Yan et al. 2019; Zhang et al. 2020; Munjal et al. 2019; Chen et al. 2020) in the lower block.

Methods	CUHK-SYSU		PRW	
	mAP	Rank-1	mAP	Rank-1
Two-Step Methods				
RCAA (Chang et al. 2018)	79.3	81.3	-	-
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
TCTS (Wang et al. 2020)	**93.9**	**95.1**	**46.8**	**87.5**
One-Step Methods				
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
CTXGraph (Yan et al. 2019)	84.1	86.5	33.4	73.6
DC-I-Net (Zhang et al. 2020)	86.2	86.5	31.8	55.1
QEEPS (Munjal et al. 2019)	88.9	89.1	37.1	76.7
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
Ours	**93.2**	**94.2**	**46.9**	**83.3**

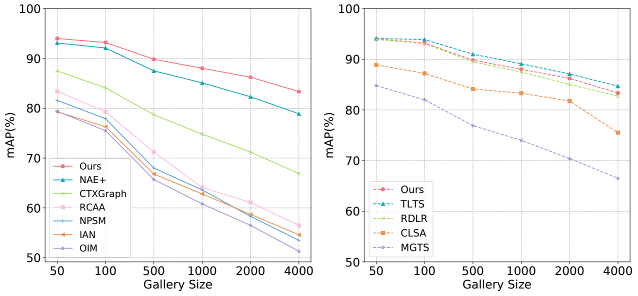
Table 4: Experimental comparisons with state-of-the-art methods on the CUHK-SYSU and PRW dataset.

Evaluation On CUHK-SYSU. The performance comparison between our network and existing competitive methods on the CUHK-SYSU dataset is shown in Tab. 4. When the gallery size is set to 100, our proposed DMRNet reaches 93.2%mAP and 94.2%rank-1. It can be seen that our method significantly outperforms all other one-step methods, as well as most two-step ones.

To evaluate the performance consistency, we also compare with other competitive methods under varying gallery sizes of $[50, 100, 500, 1000, 2000, 4000]$. Fig. 6 (a) shows the comparisons with one-step methods while (b) with two-step ones. It can be seen that the performance of all methods decreases as the gallery size increases. This indicates it is challenging when more distracting people are involved in the identity matching process, which is close to real-world applications. Our method outperforms all the one-step methods while achieving comparable performance to the two-step methods under different gallery sizes.

Evaluation On PRW. We further evaluate our method with the competitive techniques on the PRW dataset, shown in Tab. 4. We follow the benchmarking setting (Zheng et al. 2017) that the gallery contains all the 6112 testing images. Compare with the current state-of-the-art one-step method (Chen et al. 2020), it can be seen that our method outperforms it by 2.9%/2.2% on mAP and rank-1. Moreover, the mAP even surpasses the best two-step method (Wang et al. 2020) by a marginal improvement.

Runtime Comparisons. To compare the efficiency of our framework with other methods in the inference stage, we report the average runtime of the detection and re-ID for a panorama image. For a fair comparison, we test the models with an input image size as 900×1500, which is the same as other works (Chen et al. 2020; Munjal et al. 2019; Chen et al. 2018). Since the methods are implemented with different GPUs, we also report the TFLOPs. As shown in Tab. 5, upon normalization with TFLOPs, our framework is 5.73 times faster than the two-step method MGTS (Chen et al. 2018). Moreover, our method is more efficient than NAE+ (Chen et al. 2020), which is the current state-of-the-art one-step method.



(a) One-step methods

(b) Two-step methods

Figure 6: Comparisons with different gallery sizes on the CUHK-SYSU dataset. (a) and (b) shows the comparisons with one-step methods and two-step methods, respectively.

Methods	GPU	TFLOPs	Time
MGTS (Chen et al. 2018)	K80	8.7	1296
QEEPS (Munjal et al. 2019)	P6000	12.0	300
NAE (Chen et al. 2020)	V100	14.1	83
NAE+ (Chen et al. 2020)	V100	14.1	98
Ours	V100	14.1	66

Table 5: Runtime comparisons of different methods.

Conclusion

In this work, we propose a novel one-step method for person search, called the decoupled and memory-reinforced network. Extend from the one-stage detector, our multi-task learning framework substantially decouples the two subtasks. The RoI features are specific to identification, rather than supervised by multi-task losses. It also incorporates a slow-moving average of the network for yielding a consistently encoded queue-style feature memory bank. By mining informative features, our model could learn highly discriminative identification feature embeddings. Due to the massive simplification of the pipeline design, our model is easy to train and efficient to use. It sets a new state-of-the-art among one-step methods and outperforms a lot of existing two-step methods. We believe that our findings can encourage a shift in the framework of the one-step person search and drive more research on this field.

Acknowledgments

This work was supported by the Project of the National Natural Science Foundation of China No. 61876210, the Fundamental Research Funds for the Central Universities No.2019kfyXKJC024, and the 111 Project on Computational Intelligence and Intelligent Control under Grant B18024.

References

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 *European Conference on Computer Vision*.

Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; and Tai, Y. 2018. Person Search via A Mask-Guided Two-Stream CNN Model. In *European Conference on Computer Vision*.

Chen, D.; Zhang, S.; Yang, J.; and Schiele, B. 2020. Norm-Aware Embedding for Efficient Person Search. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155* .

Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*.

Dong, W.; Zhang, Z.; Song, C.; and Tan, T. 2020a. Bi-Directional Interaction Network for Person Search. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Dong, W.; Zhang, Z.; Song, C.; and Tan, T. 2020b. Instance Guided Proposal Network for Person Search. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Girshick, R. 2015. Fast r-cnn. In *IEEE International Conference on Computer Vision*.

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 *IEEE International Conference on Computer Vision*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *IEEE International Conference on Computer Vision*.

Lan, X.; Zhu, X.; and Gong, S. 2018. Person Search by Multi-Scale Matching. In *European Conference on Computer Vision*.

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

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017b. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*.

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

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*.

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

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Redmon, J.; and Farhadi, A. 2017. YOLO9000: better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition*.

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*.

Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; and Wei, Y. 2020. Circle loss: A unified perspective of pair similarity optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems* .

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *IEEE International Conference on Computer Vision*.

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

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

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

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

Yang, Z.; Liu, S.; Hu, H.; Wang, L.; and Lin, S. 2019. Reppoints: Point set representation for object detection. In *IEEE International Conference on Computer Vision*.

Zhang, L.; He, Z.; Yang, Y.; Wang, L.; and Gao, X.-B. 2020. Tasks Integrated Networks: Joint Detection and Retrieval for Image Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

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

Zheng, R.; Li, L.; Han, C.; Gao, C.; and Sang, N. 2019. Camera Style and Identity Disentangling Network for Person Re-identification. In *British Machine Vision Conference*.

Zhong, Y.; Wang, X.; and Zhang, S. 2020. Robust Partial Matching for Person Search in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition*.