DMRNet++: Learning Discriminative Features with Decoupled Networks and Enriched Pairs for One-Step Person Search

Chuchu Han, Zhedong Zheng, Kai Su, Dongdong Yu, Zehuan Yuan, Changxin Gao, Nong Sang, Yi Yang

Abstract—Person search aims at localizing and recognizing query persons from raw video frames, which is a combination of two sub-tasks, *i.e.*, pedestrian detection and person re-identification. The dominant fashion is termed as the one-step person search that jointly optimizes detection and identification in a unified network, exhibiting higher efficiency. However, there remain major challenges: (i) conflicting objectives of multiple sub-tasks under the shared feature space, (ii) inconsistent memory bank caused by the limited batch size, (iii) underutilized unlabeled identities during the identification learning. To address these issues, we develop an enhanced **d**ecoupled and **m**emory-**r**einforced network (DMRNet++). First, we simplify the standard tightly coupled pipelines and establish a task-decoupled framework (TDF). Second, we build a memory-reinforced mechanism (MRM), with a slow-moving average of the network to better encode the consistency of the memorized features. Third, considering the potential of unlabeled samples, we model the recognition process as semi-supervised learning. An unlabeled-aided contrastive loss (UCL) is developed to boost the identification feature learning by exploiting the aggregation of unlabeled identities. Experimentally, the proposed DMRNet++ obtains the mAP of 94.5% and 52.1% on CUHK-SYSU and PRW datasets, which exceeds most existing methods.

Index Terms—Person Search, Person Re-identification, Object Detection, Semi-Supervised Learning

1 Introduction

Person search [1], [2], [3], [4], [5] aims to retrieve a query person from a gallery of uncropped scene images captured by different cameras. This task consists of two sub-tasks, *i.e.*, pedestrian detection [6], [7] and person re-identification (re-ID) [8], [9]. It requires locating the persons within images first, and then matching the query with other persons for identifying the correct ones across different cameras. Compared with the pure person re-ID task, person search acts on the whole scene images, showing more potentials in real-world applications such as video analysis, video retrieval, and human-computer interaction. Despite tremendous progress achieved by recent works [2], [4], this task still suffers from the issues inherited from both detection and person re-ID, *i.e.*, viewpoint and pose variance, occlusion, complex background, false alarms in detection, misalignment, etc.

According to the structure designation of the two sub-tasks, *i.e.*, pedestrian detection and person re-ID, the existing works can be divided into two-step and one-step manners. Two-step methods [4], [10], [11], [12], [13], [14] sequentially process the sub-tasks with two separate networks. A detector is applied on raw images to predict the bounding boxes and a followed re-ID network extracts identification features from the detected person images. In contrast, one-step methods [1], [2], [3], [5], [15], [16], [17], [18] learn person localization and identification in parallel with the underlying network shared, exhibiting higher efficiency. Given an uncropped input image, the model predicts the bounding boxes and the corresponding identification features of all the detected persons within a single network.

Despite significant progress that has been made in the one-step person search [1], [2], [3], [5], three crucial issues still have not been fully solved by previous works. 1) 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 [7] supervise the shared Region-of-Interest (RoI) features with multi-task losses, *i.e.*, regression loss, foreground-background classification loss, and identification loss. As Fig. 1(a) shows, pedestrian detection focuses on learning the commonness of all persons while recognition aims to distinguish the differences among multiple identities [12]. The competing objectives of these sub-tasks make the RoI features difficult to optimize. 2) Limited by the GPU memory, the small batch size induces the inconsistent memory bank under the end-to-end fashion, as shown in Fig. 1(b). Previous works [15] maintain an exponential moving average 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. Since metric learning requires vast informative similarity pairs, the features with less consistency lead to sub-optimal identification feature learning. 3) The inherent relationships among unlabeled persons are underutilized as they are only taken as negative samples. As Fig. 1(c) shows, in the person search dataset, it is intractable to recognize and annotate all the person identities. Thus, there are some instances with only bounding box labels, termed as unlabeled identities. In the CUHK-SYSU dataset [15], 72.7% of pedestrians have no identity

- *C. Han, C. Gao and N. Sang are with the Key Laboratory of Ministry of Education for Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, 430074, China. E-mail: (hcc, cgao, nsang)@hust.edu.cn (Corresponding author: Nong Sang).*
- *Z. Zheng is with Sea-NExT Joint Lab, School of Computing, National University of Singapore, Singapore 118404. E-mail: zdzheng@nus.edu.sg*
- *Y. Yang is with School of Computer Science, Zhejiang University, Zhejiang 310058, China. E-mail: yangyics@zju.edu.cn*
- *K. Su, D. Yu, Z. Yuan are with the ByteDance, Shanghai, 201103, China. E-mail: (sukai, yudongdong, yuanzehuan)@bytedance.com*

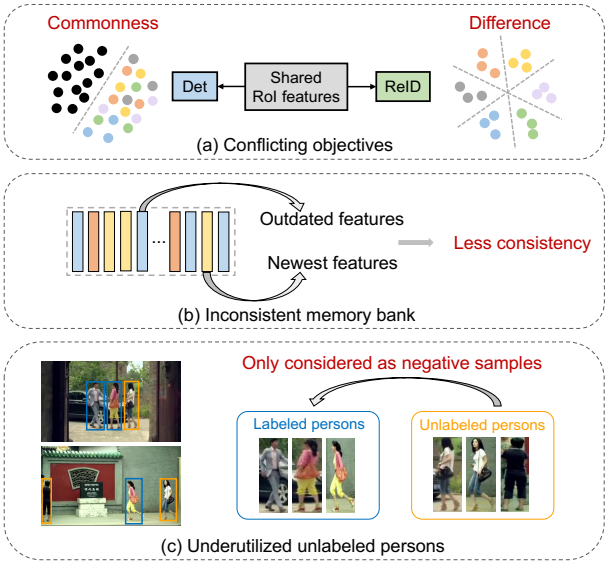


Fig. 1. Typical issues for one-step person search task. (a) Conflicting objectives: the detection task aims at finding commonness of all persons while the recognition task focuses on the difference of multiple instances. (b) Inconsistent memory bank: the outdated features are less consistent with newest ones caused by the limited batch size. (c) Underutilized unlabeled persons: the inherent relationships among unlabeled persons are unexplored since they are only taken viewed as the negative sample pairs.

annotations. Recent works [1], [2], [15] memorize these unlabeled samples in a circular queue which are considered as the negative samples of the labeled ones. However, this manner discards the great potential of unlabeled persons. To tackle these challenges, we propose the DMRNet++ containing three aspects as follows:

First, we rethink the decoupling and integration of pedestrian detection and identification in the one-step person search framework, establishing a task-decoupled framework (TDF). Different from previous works that apply conflicting objectives on RoI features, our motivation is that RoI features should be specific to the re-ID task as they contain the detailed recognition patterns of detected persons. Besides, pedestrian detection can be learned well based on the coarse convolutional features according to the success of one-stage detectors. Therefore, we take the one-stage detector as our base network. Detection and re-ID modules are branched from the layers of the feature pyramid network (FPN) [19], which contain rich visual information and could burden multiple types of task-specific feature encoding. The fine-grained RoI features sampled from FPN are only fed into the re-ID module for recognition. Based on the RoI align, interference may be involved due to the large receptive fields of network [1]. Here we introduce a point-based spatial sampling strategy to extract the RoI features, providing a more flexible approach to draw the details of pedestrians. We demonstrate that this new design makes the two sub-tasks substantially decoupled and facilitates the learning for both tasks. Specifically, the decoupled network with RetinaNet backbone [20] achieves 7.0% improvements on mAP compared to the popular baseline with Faster R-CNN.

Second, to further boost the identification feature learning, we build a memory-reinforced mechanism (MRM). Inspired by the recent unsupervised contrastive learning method [21], 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. The dual networks reinforce each other and converge to robust solution states. Experimental evidence proves that the developed mechanism is more effective than the look-up table.

Third, we introduce an unlabeled-aided contrastive loss (UCL) to exploit the potentials of the unlabeled identities. Different from previous works that only take the unlabeled persons as negative pairs, we model the recognition process as semi-supervised learning. Apart from the labeled persons, contrastive learning is also applied to unlabeled identities, enhancing the generalization of models. Without the identity labels, it is natural to construct the positive pairs through augmentations for unlabeled persons. To further exploit the aggregations, more informative positive pairs can be established by selecting the k-reciprocal nearest neighbors [22] of the unlabeled samples. With the inevitable noisy positives, different forms of contrastive loss are applied according to the similarities. Experimental evidence proves the necessity and effectiveness. Further, with this semi-supervised modeling, our proposed UCL can be generalized to other unlabeled scenarios, alleviating the burden of costly labeling.

Our main contributions can be summarized as follows:

- We propose a simplified one-step framework that decouples the optimization of detection and identification. In particular, a point-based spatial sampling is employed to generate RoI features, which 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.
- We model the recognition process as semi-supervised learning, and introduce a unlabeled-aided contrastive loss to further explore the potentials of unlabeled identities.
- Our model is easy to train and efficient to use. Adequate experiments show the competence of our DMRNet++, and the performance surpasses all the one-step methods.

A preliminary version of this work was published in [23]. We have extended our conference version as follows. (1) Owing to the inadequate exploration of the unlabeled identities, we model the re-ID module learning as a semi-supervised task. By developing an unlabeled-aided contrastive loss, more positive pairs of the unlabeled samples are constructed. To ease the effect of noisy positives, a loose constraint is applied on the positive samples with lower similarities. Enriched pairs ensure contrastive learning on unlabeled identities, thus improving the generalization of the learned model. (2) We enhance the original decoupled framework to generate more discriminative RoI features. With the proposed point-based spatial sampling strategy, a more flexible manner is provided to draw the details of pedestrians. This reduces the involved interference of RoI features caused by the large receptive fields of networks. (3) We validate the competence of DMRNet++ by exploring various data augmentations, incorporating with different detectors, and evaluating it under cross-dataset scenarios. Both quantitative and qualitative analyses are presented. (4) The experiments exhibit the effectiveness and efficiency of our DMRNet++, which reaches comparable performance with the state-of-the-art two-step method.

2 Related Work

2.1 Person search

Person search has raised a lot of interest in the computer vision community recently [10], [11], [12], [14], [15]. In the existing literature, existing approaches can be generally categorized into two families of work according to the training steps.

Two-step methods [4], [10], [11], [12], [13], [14] separate the person search task into two sub-tasks, *i.e.*, the pedestrian detection and person re-ID, trained with two independent models. Zheng *et al.* [10] first make a thorough evaluation on various combinations of different detectors and re-ID networks. They also propose a Confidence Weighted Similarity (CWS), incorporating the detection confidence into similarity matching. Lan *et al.* [11] observe the resolution diversity problem and propose a Cross-Level Semantic Alignment (CLSA) network to solve the multi-scale matching problem. Chen *et al.* [12] consider the contradictory objective problem and extract more representative features by a two-steam model. Han *et al.* [13] develop an RoI transform layer that enables gradient backpropagated from the re-ID network to the detector, obtaining more reliable bounding boxes with the localization refinement. Chan *et al.* [14] introduce reinforcement learning to the detection network by constantly trying to adjust the bounding box in various ways to find the perfect match. Wang *et al.* [4] point out the consistency problem that the re-ID model trained with hand-drawn images is not available. They alleviate this issue by producing query-like bounding boxes as well as training with detected bounding boxes.

One-step methods [1], [2], [3], [5], [15], [16], [17], [18] 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. Meanwhile, it meets more challenges, such as the contradictory objective problem and the redundant context information of instances. Xiao *et al.* [15] 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. With the usage of more unlabeled instances, a better feature space is learned. Xiao *et al.* [16] apply the center loss to this task, enhancing the discrimination of feature embeddings. Several methods consider leveraging the query image extensively. Munjal *et al.* [17] first introduce a query-guided end-to-end person search network. With the global context from both query and gallery images, the well-designed framework generates query-relevant proposals and learns query-guided re-ID scores. Yan *et al.* [18] explore the contextual information and build a graph learning framework to employ context pairs to update target similarity. To incorporate the query information into the detection network, Dong *et al.* [1] propose a Siamese network that takes both scene images and cropped person patches as input. With the guidance of the cropped patches, the learned model can focus more on persons. Yan *et al.* [24] first employ the anchor-free framework to tackle this task and address three misalignment issues in scale, region, and task levels.

As pointed out by [12], pedestrian detection focuses on learning the commonness of all persons while person re-ID aims to distinguish the differences among multiple identities. Chen [2] solves this problem by disintegrating the embeddings into norm and angle, which are used to measure the detection confidence and identity similarity, respectively. However, this method ignores the effect of the regression loss, and excessive contexts still hamper feature learning. Different from [2], we identify that the inherently defective module design is the core cause of the conflict and hinders effective feature learning.

2.2 Pedestrian detection

Pedestrian detection plays a crucial role in the person search framework. There are several commonly used detectors in traditional object detection, including Deformable Part Model (DPM) [25], Aggregated Channel Features (ACF) [26], Locally Decorrelated Channel Features (LDCF) [27] and Integrate Channel Features (ICF) [28]. In recent years, with the advent of Convolutional Neural Network (CNN), the object detection task is soon dominated by CNN-based detectors. According to whether there is a regional proposal network to generate proposals, the methods can be broadly divided into two categories: one-stage manner [20], [29], [30] and two-stage manner [7], [31], [32], [33].

The two-stage manner is composed of a proposal generator and a region-wise prediction subnetwork ordinarily. As a representative two-stage detector, Faster R-CNN [7] has been extended into numerous variants [33], [34], [35], [36], [37]. Faster R-CNN proposes a region proposal network (RPN) to generate proposals in the first stage, and then refine the object localization in the second stage. It greatly reduces the amount of computation while shares the characteristics of the backbone network. Lin *et al.* [19] design a top-down architecture with lateral connections for building multi-level semantic feature maps at multiple scales, called the Feature Pyramid Network (FPN). Using FPN in a basic detection network can assist in detecting objects at different scales.

Due to the high efficiency, the one-stage manner has attracted much more attention recently. YOLO [29], [38] directly detects objects through a single feed-forward network with fast detection speed. SSD [30] spreads out default boxes on multi-scale layers within a ConvNet, predicting the object category and box offsets. RetinaNet [20] solves the problem of class imbalance by focal loss, which attends to the learning on hard examples and down-weight the contribution of numerous easy negatives.

Recently, anchor-free detectors have raised more interest due to their simple structures and efficient implementations. FCOS [39] employs the center point of the objects to define positives, then predicts the four distances from positives to object boundary. RepPoints [40] first locates several self-learned keypoints and then predicts the bound of the spatial extent of objects. Without excessive hyper-parameters caused by anchors, these methods are more potential in terms of generalization ability.

2.3 Person Re-Identification

Person Re-Identification [8], [41], [42], [43], [44] aims at searching a query person from the cropped gallery images containing the same person in a cross-camera mode. Recently, deep learning dominates the re-ID research community with significant advantages in retrieval accuracy. Most methods focus on producing identity-discriminative representations, including representative feature mining [8], [45], [46], [47], [48], [49], [50] and deep metric learning [51], [51], [52], [53]. Sun *et al.* [8] propose a generalized Part-based Convolutional Baseline (PCB) to extract several body parts features, allowing various partition strategies for part extraction, *e.g.*, pose estimation, human parsing, and uniform partitions. Wang *et al.* [54] introduce a multi-branch deep network for learning discriminative representations, termed

as the Multiple Granularity Network (MGN). MGN learns global and local representation with a certain granularity of body partition. Meanwhile, some deep metric learning methods [10], [51], [52] are also widely used in person re-ID. Cross-entropy loss is widely used in the existing methods [10], [46], [55], which treats the training process as an image classification problem. The contrastive loss [56], [57] optimize the pairwise relationship by pulling positive sample pairs closer while pushing negative pairs farther than distance threshold. Zheng *et al.* [53] show that the contrastive loss can work well with cross-entropy loss to further boost the performance. Hermans *et al.* [51] develop a triplet hard loss, which applies an online triplet hard negative mining method in a mini-batch, promoting the result increasingly. Moreover, Chen *et al.* [52] propose a quadruplet loss, which aims to further reduce the intra-class and enlarge the inter-class variations.

3 Background

Xiao *et al.* [15] firstly propose the one-step person search work. As the most representative framework, it is widely adopted in the following methods [1], [2], [3], [5], [16], [17], [18]. Specifically, the pipeline is based on the Faster R-CNN [7], as illustrated in Fig. 2(a). With the shared backbone, detection and re-ID head are branched from the RoI features. For the re-ID head, the features are supervised by the developed Online Instance Matching (OIM) loss. Together with the detection losses in the RPN head and RoI head, the whole network is trained in an end-to-end fashion.

With a memory bank mechanism, the OIM loss is designed to enable a better convergence with large but sparse identities in the classification task. Specifically, suppose there are C labeled identities in the training set, a look-up table $W \in \mathbb{R}^{C\times d}$ is constructed to memorize the class centroid embeddings, where d denotes the feature dimension. For unlabeled persons, a circular queue $V \in \mathbb{R}^{M\times d}$ is built to store the diverse embeddings, which are used as the negative samples of labeled ones. Different from the parameters of classifiers, the look-up table and circular queue are considered as external buffers.

In a mini-batch, we denote the feature of a labeled person $x \in \mathbb{R}^d$. We then compute the cosine similarities with all the labeled and unlabeled identities by Wx and Vx, respectively. The probability of x being recognized as the identity with class-id i is defined by a softmax function:

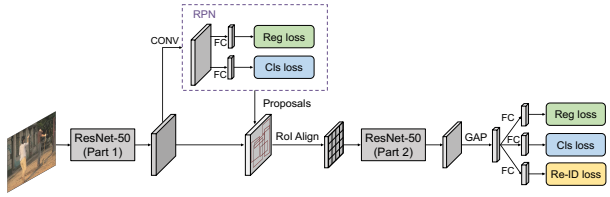
$$p_i = \frac{\exp\left(w_i x/\tau\right)}{\sum_{j=1}^{C}\exp\left(w_j x/\tau\right)+\sum_{k=1}^{M}\exp\left(v_k x/\tau\right)}, \quad (1)$$

where w_i is i-th class centroid embedding in the look-up table, and $w_i x$ measures how well x matches the i-th class centroid. τ is a temperature parameter that controls the concentration level of the distribution. The final objective is to minimize the negative log-likelihood, which is formulated as:

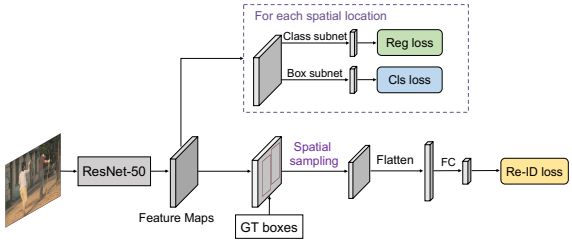
$$\mathcal{L}_{oim} = -E_x\left[\log p_{gt}\right], \quad (2)$$

where p_{gt} is the predicted probability of the ground-truth class. During backward, the out-of-date features in V are dequeued and the new unlabeled features in the current batch are enqueued. The component in the look-up table W is updated with momentum η:

$$w_{gt} \leftarrow \eta w_{gt} + (1-\eta)x. \quad (3)$$



(a) General one-step person search pipeline



(b) Our decoupled one-step person search pipeline

Fig. 2. Comparisons of different training pipelines. (a) General one-step person search pipeline. Multi-task losses are applied on the shared RoI features. (b) Our decoupled one-step person search pipeline. The RoI features are specific to the re-ID task.

4 Proposed Method

Although the one-step person search exhibits higher efficiency, several challenges still exist in this line of works. To address the three issues mentioned in Fig. 1, we develop an enhanced decoupled and memory-reinforced network (DMRNet++) outlined in Fig. 3. It contains a task-decoupled framework (TDF) to disentangle the tightly coupled pipelines (Sec. 4.1). A memory-reinforced mechanism (MRM) is introduced to ensure a more consistent memory bank (Sec. 4.2). To further exploit the underutilized unlabeled persons, an unlabeled-aided contrastive loss (UCL) is proposed for recognition (Sec. 4.3).

4.1 Task-Decoupled Framework

As exhibited in Fig. 2(a), the general person search pipeline is widely adopted in the existing one-step algorithms [1], [2], [3], [5], [16], [17], [18]. To disentangle this tightly coupled pipeline, we propose a task-decoupled framework shown in Fig. 2(b). We take the following aspects into account when designing our framework:

Spatial-decoupled: There exist contradictory objectives when supervising the shared RoI features with multi-task losses. 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. Thus, we target to spatially decouple the imposed feature spaces of two sub-tasks.

Representative: It is crucial to generate representative features for pedestrians during retrieval. Relying on the RoI align layer with the proposals is suboptimal, especially in the beginning there are several incorrect results for detection. Even if the proposals are accurate enough, the features extracted by RoI align may involve interferential cues outside the boxes due to the large receptive field [1]. To alleviate the effect of excess context, a point-based strategy is proposed to provide a more flexible approach to drawing the details of pedestrians.

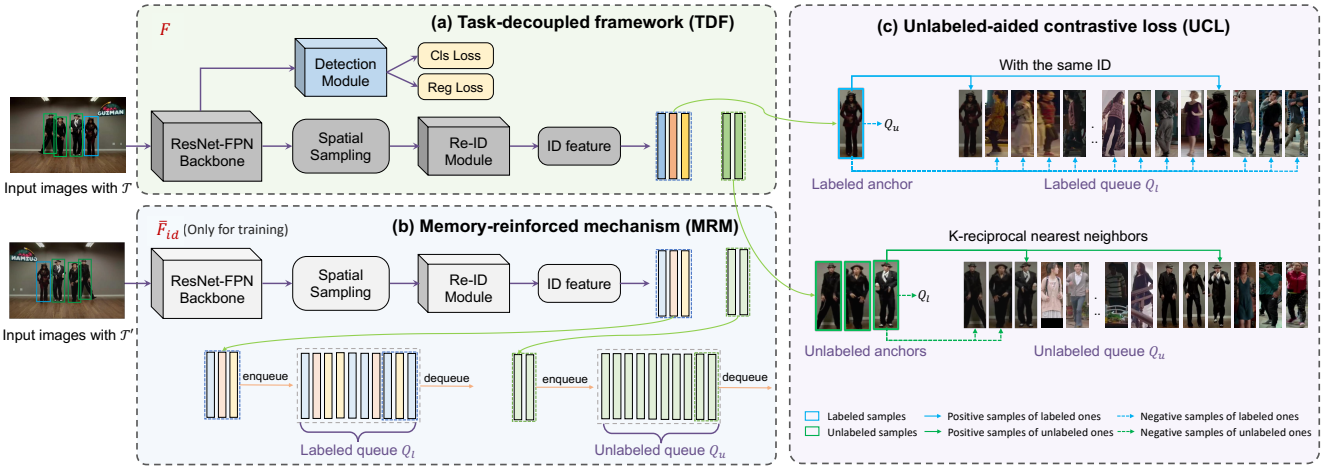


Fig. 3. An overview of the enhanced Decoupled and Memory-Reinforced Networks (DMRNet++) for one-step person search. (a)F is our task-decoupled framework. The images are input to a shared backbone, then detection and re-ID modules are branched from the layers of FPN. With a point-based spatial sampling, the RoI features are specific to the re-ID task. The generated ID features are termed labeled or unlabeled anchors. (b)$\bar{F}_{id}$ is a slowly-updating network counterpart, which is utilized to consistently encode the re-ID features, maintaining consistent memory banks, *i.e.*, Q_l and Q_u. Thus, multiple positive and negative similarity pairs can be built between anchors and queued embeddings. (c) Diagram of the proposed UCL. For the labeled anchor (blue box), its positive pairs (blue solid line) are selected from Q_l that with the same identities. The rest ones in Q_l and Q_u are considered as negative pairs (blue dotted line). For the unlabeled anchor (green box), the assigned positive pairs (green solid line) are obtained by k-reciprocal nearest neighbors. Other samples within the same image and those in Q_l are negative pairs (green dotted line).

Simplicity: For the person search task, the detector only requires to distinguish person or background, rather than the multi-classification task in object detection [7]. It is unnecessary to apply foreground-background classification loss in both the RPN head and the RoI head. Besides the redundant loss function, the split of the backbone also makes the re-ID module awkward. To tackle this, we take simplicity into account when designing the framework.

The above criteria motivate a sample framework to generate representative RoI features for identification, which is spatial-decoupled for the sub-tasks. As illustrated in Fig. 2(b), instead of the multi-task losses under a shared feature space, our design contains three core aspects:

First, since the RoI features contain the detailed recognition patterns of detected persons, they can be specific to the re-ID task. Meanwhile, 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 [20], FCOS [39] and RepPoints [40]. Specifically, we adopt the ResNet-50 [58] with FPN [19] as the shared backbone. For the detection head, a class subnet and a box subnet based on FPN are employed to perform foreground-background classification and bounding box regression on each location. To extract the RoI features for identification, we conduct spatial sampling on the FPN layers to produce fine-grained embeddings. Since FPN include rich semantic information while RoI features contain specific content, this design makes the two sub-tasks substantially decoupled.

Second, to produce more discriminative RoI features, we employ the ground truth bounding boxes for the correctness and introduce a spatial sampling strategy to ensure representativeness. On one hand, we extract RoI features only based on the ground truth bounding boxes during 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. On the other hand, we improve the box-based spatial sampling methods with a point-based strategy to extract RoI features. As mentioned above, the RoI align relies on rectangular bounding boxes, *i.e.*, the ground truth boxes in training and the final predictions in testing. The bounding box is convenient to use but provides coarse localization and results in the corresponding rough extraction of persons. In this paper, to focus on the discriminative part of pedestrians, we aim at learning a set of points instead of the boxes to extract RoI features. Specifically, given the input feature map f and a ground truth box of size $b_w \times b_h$, we first sample $g \times g$ initial points inside the box area uniformly, denoted as $\{a_{ij}\}(0 <= i, j < g)$. Inspired by [35], additional point-wise offsets $\Delta a \in \mathbb{R}^{g \times g \times 2}$ are learned based on the initial points to augment the spatial sampling locations, which is formulated as:

$$\Delta a = \alpha \cdot MLP\left(f', \theta_a\right) \circ (b_w, b_h), \tag{4}$$

where f' and θ_a denote the extracted RoI features based on the initial points and the learnable parameters, respectively. MLP represents a three-layer fully connected subnet, which learns the normalized point-wise offsets. $\circ$ denotes element-wise product. Following [35], the offsets are transformed by element-wise product with the width and height of the box, with a pre-defined scalar α. For (i, j)-th point, the new augmented point is obtained by adding the learned offset Δa_{ij}. To generate the RoI features $\hat{f}$ by these irregular points, Eq. (5) is implemented by the bilinear interpolation operation [59] on f.

$$\hat{f}(i, j) = f\left(a_{ij} + \Delta a_{ij}\right). \tag{5}$$

Third, by removing the detection loss from the RoI head, the re-ID module is solely preserved with high simplicity. Without the separated ResNet-50 and global average pooling, the extracted RoI feature is flattened and transformed by a fully connected layer. This new design is both efficient and simple compared with the general one-step pipeline.

Discussion: Previous works [2], [12] also study the contradictory goals of the two sub-tasks. [12] adopts the two-step approach to avoid conflict. For the one-step manner, [2] reconciles the conflict by factorizing embeddings into magnitude and direction for foreground scoring and re-ID, respectively. From another perspective, we identify that the inherently defective module design is the main cause of this issue and hinders the effective feature learning of the one-step models. By decoupling the two sub-tasks spatially, the contradictory goals are conducted on different feature spaces. In the experiments, we make comparisons on different network designs. It shows that the negative effect of the inferior module can be alleviated clearly with a simple head disentanglement. Further, almost 7.0% improvements can be achieved with our decoupled design, more details are exhibited in Sec. 5.4.

4.2 Memory-Reinforced Mechanism

Effective feature learning is challenging for the one-step person search. Although the OIM loss solves the gradient issue by introducing the memory bank mechanism, the memorized embeddings are not consistently encoded. Specifically, limited by the GPU memory constraints in the end-to-end fashion, the batch size is relatively small in the one-step person search. In the OIM loss, samples in each mini-batch select their proxy embeddings as positive pairs from the memory bank. However, the proxy is updated when it meets the sample with the same category. This induces that the encoders of the comparing features come from different iterations with a long time apart. The memorized embeddings could be outdated as the weights of the model evolve and are less consistent. 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 [21], [60], 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 embeddings within the look-up table, we maintain a queue-style memory bank for the labeled instances. 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.

In this paper, we maintain a labeled queue $Q_l \in \mathbb{R}^{L\times d}$ containing the features of L labeled persons, and an unlabeled queue $Q_u \in \mathbb{R}^{U\times d}$ containing the features of U unlabeled persons, where d is the feature dimension.

A slow-moving average of the network. To make the stored features encoded consistently, we introduce a slow-moving average of the network for generating features in the memory bank. We denote our decoupled network as F, where its parameters θ are updated by the back-propagation. The slow-moving average of the network is denoted by $\bar{F}_{id}$. Its parameters $\bar{\theta}$ are updated by exponential moving average (EMA) [60] at each iteration:

$$\bar{\theta} \leftarrow m\bar{\theta} + (1-m)\theta, \tag{6}$$

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 the detection modules. $\bar{F}_{id}$ requires no gradient and brings little overhead at each iteration.

4.3 Unlabeled-aided Contrastive Loss

In previous OIM-based methods [1], [2], [5], [15], the unlabeled persons are only considered as negative classes for all the labeled identities, enlarging the underlying high-dimensional visual space. However, the inherent relationships among unlabeled persons are lacking exploration. To exploit the potential of the unlabeled samples, we develop an unlabeled-aided contrastive loss with both labeled and unlabeled samples as anchors.

Inspired by recent contrastive learning frameworks [21], [61], [62], two correlated views of the same example are processed by F and $\bar{F}_{id}$ to maximize the agreement via the proposed contrastive loss. Different augmented inputs are beneficial to the samples in yielding effective embeddings, especially for the unlabeled ones. Besides, by exploring the k-reciprocal neighbors [22], our method can better leverage the unlabeled identities and harness the inherent semantics to enrich the original supervised representations. The whole recognition process can be modeled as semi-supervised learning. Further, our framework can be generalized to more scenarios with under-labeled data. The loss calculation can be divided into four steps as follows:

First, given a batch of input images, two separate data augmentation operators are sampled from the same set of augmentations with stochasticity, denoted as $(t \sim \mathcal{T})$ and $(t' \sim \mathcal{T})$. $\bar{F}_{id}$ and F are trained to promote the consistency of the sample with different data augmentations by contrastive learning. For one augmented image set, the features insistently encoded by $\bar{F}_{id}$ are employed to update the Q_l and Q_u. As Fig. 3 shows, these newest embeddings are enqueued while the outdated ones are dequeued, preserving consistent queues with fixed lengths. The other augmented image set is processed by F, producing the embeddings considered as the anchors. Suppose there is a **labeled anchor** x_l with the identity of i and an **unlabeled anchor** x_u.

Second, for the **labeled anchor** x_l, assuming that there are K positive samples in Q_l sharing the same identity with x_l, and the rest J ones in Q_l and Q_u are viewed as negative samples. The positive and negative cosine similarities are denoted as $\{s_p^i\}(i = 1, 2, ..., K)$ and $\{s_n^j\}(j = 1, 2, ..., J)$, respectively. Different from the look-up table [15] that provides a single positive sample, *i.e.*, the centroid embedding, our method fits well the multi-positive scene. Since the memory bank is decoupled from the number of identities, we obtain multiple positive and negative samples for the labeled anchor x_l. To achieve balanced gradient learning, we use a simplified loss function from [63] as multi-positive contrastive learning:

$$\mathcal{L}_l = log[1 + \sum_{i=1}^{K}\sum_{j=1}^{J}\exp(\gamma(s_n^j - s_p^i))], \tag{7}$$

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. With this supervision, F and $\bar{F}_{id}$ reinforce each other and their parameter spaces converge to robust solution states.

Third, the labeled queue Q_l with a large size has ensured adequate negative samples, substituting the effect of unlabeled identities. To exploit the inherent relationship among these unlabeled identities, we apply contrastive learning to construct positive and negative pairs. Discriminative embeddings can be learned by

enriching sample pairs. Without the identity annotations, it is natural to construct the positive pairs through different augmentations for unlabeled persons. For further aggregating the similar samples, we select the k-reciprocal nearest neighbors of unlabeled features to establish more informative positive pairs. Specifically, for a specific **unlabeled anchor** x_u, we calculate the cosine similarities between x_u and the features in unlabeled queue Q_u. The similarity list is arranged in descending order, where we select the top-k similarities, and their corresponding samples are defined as the k-nearest neighbors of x_u, *i.e.*, $N(x_u, k)$. It means the top-k samples in Q_u with the highest similarities to x_u.

$$N(x_u, k) = \{q_1^u, q_2^u, ..., q_k^u\}. \tag{8}$$

Our hypothesis is that if two persons belong to the k-nearest neighbors of each other, they are more likely to be the same identity [22]. Under this criterion, the satisfied samples are called the k-reciprocal nearest neighbors. Specifically, for the unlabeled anchor x_u, its k-nearest neighbors are $N(x_u, k) = \{q_i^u\}$. Among $\{q_i^u\}$, whose k-nearest neighbors containing the unlabeled anchor in reverse are called the k-reciprocal nearest neighbors of x_u, defined as $R(x_u, k)$. The formulation is as follows:

$$R(x_u, k) = \{q_i^u | (q_i^u \in N(x_u, k)) \cap (x_u \in N(q_i^u, k))\}. \tag{9}$$

The obtained samples in $R(x_u, k)$ are taken as the candidate positive pairs of x_u. Considering that the pedestrians in a scene image belong to different categories, we remove the samples within the same image of x_u in $R(x_u, k)$, and the rest are considered as positive samples. With this strategy, more potential positive pairs can be exploited, as illustrated in Fig. 3. To construct the negative pairs of x_u, we can also employ the prior, thus taking the instance within the same image of x_u as the negative samples. Together with the memorized features in Q_l, plenty of negative pairs can be built.

Fourth, supposing there are K' positive pairs and J0 negative ones for x_u totally. The corresponding cosine similarities are represented as $\{\hat{s}_p^i\}(i = 1, 2, ..., K')$ and $\{\hat{s}_n^j\}(j = 1, 2, ..., J')$, respectively. Considering some noisy positive samples may be involved, we adopt two forms of contrastive losses for the unlabeled samples. With a defined threshold μ, the positive samples with similarities greater than μ are considered as reliable positives. They are supervised by the upper formula of Eq. (10), which is the same as Eq. (7). For the positive pairs with lower similarities, we employ a loose variation of contrastive loss:

$$\mathcal{L}_u = \begin{cases} \log[1 + \sum_{i=1}^{K'} \sum_{j=1}^{J'} \exp(\gamma(s_n^j - s_p^i))] & \text{if } s_p^i > \mu \\ -\log \frac{\sum_{i=1}^{K'} \exp(\gamma \hat{s}_p^i)}{\sum_{i=1}^{K'} \exp(\gamma \hat{s}_p^i) + \sum_{j=1}^{J'} \exp(\gamma \hat{s}_n^j)} & \text{if } s_p^i < \mu \end{cases} \tag{10}$$

Consequently, the unlabeled-aided contrastive loss consists of the ones on both labeled and unlabeled anchors.

$$\mathcal{L}_{UCL} = \mathcal{L}_l + \mathcal{L}_u. \tag{11}$$

In UCL, we employ two kinds of contrastive losses for different samples based on the following considerations. For the labeled anchors, since we have the identity information, it is intuitive to make every s_p greater than every s_n. The Eq. (7) ensures this property and keeps the balance of gradient. Similarly, for the unlabeled anchors, their k-reciprocal nearest neighbors with higher similarity are considered more reliable ones. Thus, the Eq. (7) is also suitable, which can be rewritten as:

$$\begin{aligned} \mathcal{L}_l &= \log[1 + \sum_{i=1}^{K} \sum_{j=1}^{J} \exp(\gamma(s_n^j - s_p^i))] \\ &= \log[1 + \sum_{j=1}^{J} \exp(\gamma s_n^j) \sum_{i=1}^{K} \exp(-\gamma s_p^i)] \\ &= -\log \frac{(\sum_{i=1}^{K} \exp(-\gamma s_p^i))^{-1}}{(\sum_{i=1}^{K} \exp(-\gamma s_p^i))^{-1} + \sum_{j=1}^{J} \exp(\gamma s_n^j)}. \end{aligned} \tag{12}$$

The target of this equation is to make the $(\sum_i \exp(-\gamma s_p^i))^{-1}$ greater, which means decreasing the $\sum_i \exp(-\gamma s_p^i)$. It requires each s_p^i to be greater enough, and can be regarded as the hard mining of positive sample pairs.

However, for the positive pairs with lower similarities of unlabeled anchors, it is more likely to involve some noise, *e.g.*, the wrong identities. The hard mining on positive pairs may overemphasize the false positives, and damage the learned embeddings. To make it robust to noise, we use a loose variation of contrastive loss for the unlabeled anchor and its positive pairs with lower similarities. Different from the upper formula in Eq. (10) that each s_p^i requires to be greater enough, the lower one exhibits loose constraints which tries to increase $\sum_i \exp(\gamma \hat{s}_p^i)$. It means that if a relatively large similarity is reached by other positive pairs, the false positive can be ignored under this easy mining.

The training process is summarized in Algorithm 1. We optimize the joint model with both the UCL loss in the re-ID module, and the classification loss $\mathcal{L}_{cls}$ and regression loss $\mathcal{L}_{reg}$ in the detection module. The overall objective function is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{UCL} + \mathcal{L}_{cls} + \mathcal{L}_{reg}. \tag{13}$$

Algorithm 1 Training process of DMRNet++

Require: Training datasets with bounding box and identity annotation;
Require: Initialize the backbone F and $\bar{F}_{id}$ with pretrained ResNet-50;
Require: Initialize the Q_l and Q_u with zeros;
Require: Scale factor γ for Eq. (7)(10), momentum m for Eq. (6), k for Eq. (9), similarity threshold μ for Eq. (10)
for each epoch **do**
  for each mini-batch **do**
    1: Apply two data augmentations on the input images.
    2: Encode labeled and unlabeled anchors $\{x_l^i\}, \{x_u^j\}$ by F.
    3: Encode labeled and unlabeled features $\{\bar{x}_l^i\}, \{\bar{x}_u^j\}$ by $\bar{F}_{id}$.
    4: Update Q_l and Q_u with $\{\bar{x}_l^i\}, \{\bar{x}_u^j\}$, respectively.
    5: For the labeled anchors $\{x_l^i\}$, the samples with the same identity in Q_l are positive pairs, while the rest in Q_l and those in Q_u are considered as negative pairs. Compute the pairwise loss by Eq. (7).
    6: For the unlabeled anchors $\{x_u^j\}$, the positive pairs are selected by Eq. (9). The samples within the same image and those in Q_l are taken as negative pairs. Compute the pairwise loss by Eq. (10)
    7: Compute the overall loss by Eq. (13).
    8: Update the encoder F by back-propagation.
    9: Update the encoder $\bar{F}_{id}$ with momentum m by Eq. (6).
  end for
end for

5 Experiments

In this section, we first describe the datasets and evaluation protocols in Sec. 5.1, after which the implementation details are elaborated in Sec. 5.2. Then we compare with state-of-the-art methods in Sec. 5.3. Next, we conduct comprehensive ablation studies to explore the effects of different components for DMRNet++ in Sec. 5.4. More parameter analysis are shown in Sec. 5.5 and adequate visualization results are exhibited in Sec. 5.6.

5.1 Datasets and Settings

CUHK-SYSU. CUHK-SYSU [15] 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. PRW [10] is extracted from the video frames which are captured by six spatially disjoint cameras. There are totally $11,816$ frames with the $43,110$ annotated bounding boxes. PRW also contains unlabeled identities and labeled identities which is ranged from 1 to 932. In the 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 [15], [17]. 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.

5.2 Implementation Details

The backbone in our network adopts the ResNet-50 [58] pretrained on the ImageNet dataset, together with the FPN [19]. For the detection network, we use the latest PyTorch implementation of RetinaNet [20], FCOS [39], ATSS [64], Foveabox [65] and RepPoints [40] released by OpenMMLab [66]. Actually, our framework is compatible with most detectors. The input images are resized to 1333×800 by default. We also evaluate with larger resolutions and multi-scale training strategy, detailed in Sec. 5.4. The batch size is set to 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}. We employ a step decay learning rate schedule with a warm-up strategy, where the learning rate is gradually increased to 1×10^{-3} in the first epoch, then decreased to 1×10^{-4} and 1×10^{-5} at epoch 8 and epoch 11. Our model is trained for 12 epochs totally. The queue sizes L and U are set to 8196 and 8196 for CUHK-SYSU while 1024 and 1024 for PRW. The momentum factor m is set to 0.999, and the scale factor γ is set to 16. The k is set to 5 in Eq. (8) and 2 in Eq. (9). The similarity threshold in Eq. (10) is set to 0.7. These hyper-parameters are tuned on a validation set of the CUHK-SYSU dataset. 100 identities are randomly split from the original training set as the query for validation, and the rest are considered as the new training set. For the gallery set, it is the same as the original training set. All experiments are implemented on the PyTorch framework, and the network is trained on an NVIDIA GeForce GTX 1080 Ti.

5.3 Comparisons with the State-of-the-Art Methods

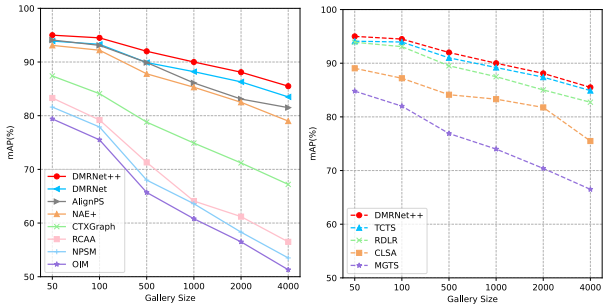
5.3.1 Evaluation On the CUHK-SYSU dataset

Compared methods. In this section, we compare our proposed methods with the state-of-the-arts, including both one-step and two-step manners. For the two-step approach, there are 12 baselines by combining different pedestrian detectors and person re-ID works. Specifically, three baseline detection networks are used to detect persons, including ACF [26], CCF [70], and Faster R-CNN based on ResNet-50 (CNN) [75]. For the re-ID networks, representative person descriptors are used to extract features, such as DenseSIFT-ColorHist (DSIFT) [67], Local Maximal Occurrence (LOMO) [69], Bag of Words (BoW) [71], and IDNet (the re-ID part of OIM [15]). The distance metric methods (*i.e.*, KISSME [68], XQDA [69]) are combined with person descriptors for re-ID. Besides, some two-step methods exhibit excellent performance are also compared, including RCAA [14], MGTS [12], CLSA [11], RDLR [13], TCTS [4]. We also compare our methods with one-step methods, including OIM [15], NPSM [72], RCAA [14], MGTS [12], CLSA [11], NAE [2], BINet [1], PGA [74] and AlignPS [24].

TABLE 1
Experimental comparisons with state-of-the-art methods on the CUHK-SYSU dataset. 'MS' denotes the multi-scale training strategy.'DCN' means the deformable conv layers are used in ResNet.

Method	mAP(%)	Rank-1(%)
Two-Step Methods		
ACF [26]+DSIFT [67]+Euclidean	21.7	25.9
ACF [26]+DSIFT [67]+KISSME [68]	32.3	38.1
ACF [26]+LOMO+XQDA [69]	55.5	63.1
CCF [70]+DSIFT [67]+Euclidean	11.3	11.7
CCF [70]+DSIFT [67]+KISSME [68]	13.4	13.9
CCF [70]+LOMO+XQDA [69]	41.2	46.4
CCF [70]+IDNet	50.9	57.1
CNN [7]+DSIFT [67]+Euclidean	34.5	39.4
CNN [7]+DSIFT [67]+KISSME [68]	47.8	53.6
CNN [7]+Bow [71]+Cosine	56.9	62.3
CNN [7]+LOMO+XQDA [69]	68.9	74.1
CNN [7]+IDNet	68.6	74.8
RCAA [14]	79.3	81.3
MGTS [12]	83.0	83.7
CLSA [11]	87.2	88.5
RDLR [13]	93.0	94.2
TCTS [4]	**93.9**	**95.1**
One-Step Methods		
OIM [15]	75.5	78.7
IAN [16]	76.3	80.1
NPSM [72]	77.9	81.2
CTXGraph [18]	84.1	86.5
DC-I-Net [73]	86.2	86.5
QEEPS [17]	88.9	89.1
BINet [1]	90.0	90.7
PGA [74]	90.2	91.8
NAE [2]	91.5	92.4
NAE+ [2]	92.1	92.9
AlignPS [24] (MS)	93.1	93.4
AlignPS+ [24] (MS+DCN)	94.0	94.5
DMRNet	93.2	94.2
DMRNet++	**94.4**	**95.5**
DMRNet++ (MS)	**94.5**	**95.7**

Experimental results. Tab. 1 shows the performance comparisons on the CUHK-SYSU dataset. Both the mAP and rank-1 accuracy are reported for evaluation. The results of two-step methods are shown in the upper block while the one-step methods are exhibited in the lower block. When the gallery size is set to 100, our proposed DMRNet++ reaches $\mathbf{94.4\%}$ on mAP and $\mathbf{95.5\%}$ on



(a) One-step methods (b) Two-step methods

Fig. 4. Comparisons with different methods at varying gallery sizes on the CUHK-SYSU dataset. (a) and (b) shows the comparisons with one-step methods and two-step methods, respectively.

rank-1. It can be seen that our method outperforms all other one-step methods. In the existing literature, although one-step methods exhibit high efficiency, the performance is always inferior to two-step ones. It is significant that our DMRNet++ exceeds the state-of-the-art two-step method TCTS [4]. Compared with TCTS, our method obtains the performance gain of 0.5% and 0.4% in terms of mAP and rank-1 metric. We also exhibit the results when applying the multi-scale training strategy, referring to Sec. 5.4 for detailed settings, and the performance is slightly promoted. The results show the effectiveness of our decoupled network and memory-reinforced mechanism. The decoupled design alleviates the conflict of optimization. The consistently encoded memory bank ensures sufficient positive and negative pairs, providing effective identification embeddings. The result also demonstrates the potential of the one-step method on both efficiency and accuracy.

To evaluate the performance consistency, we also compare with other competitive methods under varying gallery sizes of $[50, 100, 500, 1000, 2000, 4000]$. Fig. 4(a) shows the comparisons with one-step methods while Fig. 4(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.

5.3.2 Evaluation On the PRW dataset

Compared methods. Similarly, we compare with the state-of-the-art methods on the PRW dataset [10]. There are 9 traditional two-step person search methods, which combine the separated detection network (*i.e.*, ACF [26], DPM-Alex [25], LDCF [27]) and re-ID methods (*i.e.*, LOMO [69]+XQDA [69], IDE$_{det}$ [10]). Confidence Weighted Similarity (CWS) [10] is proposed to incorporate the detection confidence when similarity matching. Other two-step methods include MGTS [12], CLSA [11], RDLR [13], TCTS [4]. For the one-step methods, we compare our methods with OIM [15], NPSM [72], RCAA [14], MGTS [12], CLSA [11], NAE [2], BINet [1], PGA [74] and AlignPS [24].

Experimental results. The comparisons on the PRW dataset [10] are shown in Tab. 2. Following the benchmarking setting [10], the gallery contains all 6112 testing images. It can be seen that our proposed DMRNet++ reaches **51.0**% on mAP and **86.8**% on rank-1. For the combinations of detection methods and re-ID models, DPM-Alex [25]+IDE$_{det}$ [10] +CWS [10] achieves the best performance, while we surpass it by a large margin.

TABLE 2
Experimental comparisons with state-of-the-art methods on the PRW dataset. 'MS' denotes the multi-scale training strategy.'DCN' means the deformable conv layers are used in ResNet.

Method	mAP(%)	Rank-1(%)
Two-Step Methods		
ACF-Alex [26]+LOMO+XQDA [69]	10.3	30.6
ACF-Alex [26]+IDE$_{det}$ [10]	17.5	43.6
ACF-Alex [26]+IDE$_{det}$ [10]+CWS [10]	17.8	45.2
DPM-Alex [25]+LOMO+XQDA [69]	13.0	34.1
DPM-Alex [25]+IDE$_{det}$ [10]	20.3	47.4
DPM-Alex [25]+IDE$_{det}$ [10]+CWS [10]	20.5	48.3
LDCF [27]+LOMO+XQDA [69]	11.0	31.1
LDCF [27]+IDE$_{det}$ [10]	18.3	44.6
LDCF [27]+IDE$_{det}$ [10]+CWS [10]	18.3	45.5
MGTS [12]	32.6	72.1
CLSA [11]	38.7	65.0
RDLR [13]	42.9	70.2
TCTS [4]	**46.8**	**87.5**
One-Step Methods		
OIM [15]	21.3	49.9
IAN [16]	23.0	61.9
NPSM [72]	24.2	53.1
CTXGraph [18]	33.4	73.6
DC-I-Net [73]	31.8	55.1
QEEPS [17]	37.1	76.7
PGA [74]	42.5	83.5
NAE [2]	43.3	80.9
NAE+ [2]	44.0	81.1
BINet [1]	45.3	81.7
AlignPS [24] (MS)	45.9	81.9
AlignPS+ [24] (MS+DCN)	46.1	82.1
DMRNet	46.9	83.3
DMRNet++	**51.0**	**86.8**
DMRNet++ (MS)	**52.1**	**87.0**

Compare with the state-of-the-art one-step work AlignPS+ [24], our method outperforms it by 4.9% on mAP even without the multi-scale training strategy and deformable convolution layers. It is observed that the multi-scale training strategy further improves the performance on mAP clearly. More discussions on the image resolutions are shown in Sec. 5.4. Compared with the CUHK-SYSU dataset, the PRW dataset lacks the diversity of clothing. Many identities have similar appearances, making it challenging to distinguish these persons. This causes poor performance of mAP in existing methods. Surprisingly, our DMRNet++ surpasses the best two-step work [4] by 4.2% on mAP, showing the effectiveness of DMRNet++ under various scenarios.

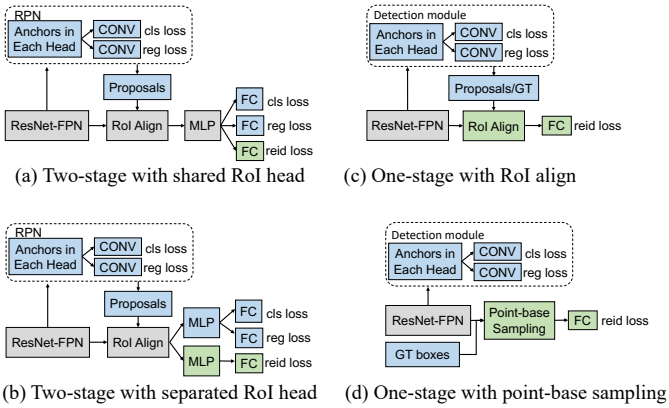
5.4 Ablation Study on DMRNet++

In this section, we conduct detailed ablation studies to evaluate the effectiveness of the DMRNet++. We first explore the effect of different network designs, and analyze the effectiveness of the proposed components in DMRNet++. Then we investigate the impact of various image augmentations, and compare two memory bank mechanisms. After that, the performance of DMRNet and DMRNet++ under different settings, *e.g.*, detectors, resolutions are exhibited. Next, we validate the generalization ability of the proposed methods under the cross-dataset scenario. Finally, we show the efficiency of the proposed method by runtime comparisons.

Comparisons on different network designs. To investigate what causes the poor performance of one-step person search, we first

TABLE 3
Comparisons of different network designs on the CUHK-SYSU and PRW datasets. The performance of re-ID and detector trained in a single network is represented, denoted as R and D. D-S denotes the result of the separated trained detector.

Methods	CUHK-SYSU				PRW			
	(R)mAP	(R)Rank-1	(D)mAP	(D-S)mAP	(R)mAP	(R)Rank-1	(D)mAP	(D-S)mAP
Faster R-CNN + OIM [15]	75.5	78.7	-	-	21.3	49.9	-	-
Faster R-CNN (FPN) w/ (a) + OIM	84.3	84.6	86.9	92.2	29.0	51.5	93.1	95.1
Faster R-CNN (FPN) w/ (b) + OIM	87.5	87.7	89.8		34.8	58.1	93.9	
RetinaNet (FPN) w/ (c) + OIM (Proposals)	90.0	90.8	91.2	92.3	36.0	73.3	94.7	95.3
RetinaNet (FPN) w/ (c) + OIM (GT)	90.3	91.0	91.4		36.1	73.6	94.8	
RetinaNet (FPN) w/ (d) + OIM (Proposals)	90.8	91.7	91.3		37.0	74.8	94.7	
RetinaNet (FPN) w/ (d) + OIM (GT)	90.9	91.7	91.4		37.2	74.9	94.9	
RepPoints (FPN) w/ (c) + OIM (GT)	92.4	93.2	91.7	93.1	39.1	73.6	94.7	95.4
RepPoints (FPN) w/ (d) + OIM	93.0	93.9	91.6		40.3	75.3	94.7	



(a) Two-stage with shared RoI head (c) One-stage with RoI align

(b) Two-stage with separated RoI head (d) One-stage with point-base sampling

Fig. 5. Comparisons on different network designs. (a) The general person search pipeline is based on a two-stage detector, where the RoI head is shared by detection and re-ID. (b) Separated RoI heads are employed to ease the coupling of sub-tasks. (c) With the discarded detection losses in RoI head, the RoI features extracted by RoI align are specific to identification. (d) A point-based spatial sampling is introduced to generate the RoI features.

conduct several experiments to illustrate the comparisons among different network options, as shown in Fig. 5. Detailed results with various settings are shown in Tab. 3.

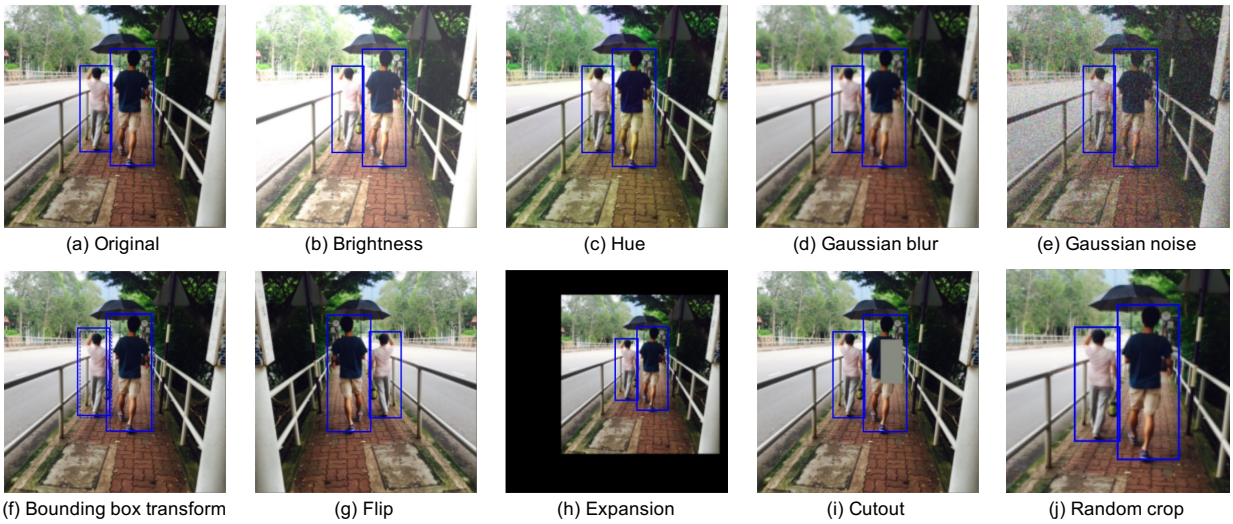
For fair comparisons, we incorporate FPN into the general one-step framework [15] as our strong baseline (a), and this improves the performance by a large margin, *e.g.*, about 8% promotion on mAP. When it comes to tangled sub-tasks (detection and re-ID) conflict in the one-step person search, it is natural to think about decoupling different tasks from the shared backbone. For this purpose, (b) employs separated RoI heads for detection and re-ID training. In Tab. 3, the results perform better than a shared RoI head manner on both re-ID and detection tasks. This indicates the inherently defective module design makes the network severely coupled. It harms the optimization on both sub-tasks when sharing feature space and 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 shown in Fig. 5(c), one-stage detectors can be well incorporated and the RoI features are specific for identification. This manner achieves the rank-1 of 90.8% and 73.3% on two datasets, surpassing the baseline on both re-ID and detection by a large margin. It shows the spatial-aware decoupling benefits the optimization on two sub-tasks. Note that the performance of separated trained detectors for one-stage (RetinaNet) or two-stage (Faster R-CNN) is almost the same. This denotes the improvements originate from the decoupled design, other than a superior detector. In Fig. 5(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 boxes. The comparisons are shown in Tab. 3. The two approaches show similar performance under box-based and point-based sampling manners. We only use the ground truth boxes since it saves much computational cost in training.

As shown in Fig. 5(d), our proposed task-decoupled network additionally introduces a point-based spatial sampling strategy to extract the RoI features. Compared with the box-based RoI align, this manner focuses more on the pedestrians, and reduces the contextual interference caused by the large receptive field. On the PRW dataset, our proposed point-based strategy can boost the rank-1 by 1.3% and 1.7% with RetinaNet and RepPoints, respectively. This shows the superiority of extracting features with irregular points other than rectangular boxes. Some visualizations further support the results, which are detailed in Sec. 5.6. Finally, based on our proposed task-decoupled framework, the performance achieves the mAP of 93.0% and 40.3% on the CUHK-SYSU and PRW datasets.

TABLE 4
Component analysis of the proposed DMRNet++ on the CUHK-SYSU and PRW datasets. In TDF, there are two manners to generate RoI features, *i.e.*, box-based RoI align and point-based spatial sampling. In UCL, $\mathcal{L}_l$ and $\mathcal{L}_u$ denote the Eq. (7) and Eq. (10). $\mathcal{L}'_u$ means only the upper loss in Eq. (10) is applied on the unlabeled samples.

TDF		MRM	UCL			More	CUHK-SYSU		PRW	
box	point		$\mathcal{L}_l$	$\mathcal{L}_u$	$\mathcal{L}'_u$	Augs	mAP	Rank-1	mAP	Rank-1
							84.3	84.6	29.0	51.5
✓							92.4	93.2	39.1	73.6
✓		✓	✓				92.9	93.7	46.0	83.2
	✓	✓	✓				93.5	94.4	47.2	83.8
	✓	✓	✓	✓			93.9	94.9	49.5	85.7
	✓	✓	✓	✓		✓	94.3	95.3	50.2	86.1
	✓	✓	✓		✓	✓	94.1	95.1	49.0	85.3



(a) Original (b) Brightness (c) Hue (d) Gaussian blur (e) Gaussian noise

(f) Bounding box transform (g) Flip (h) Expansion (i) Cutout (j) Random crop

Fig. 6. Illustrations of the data augmentation. Given the original image (a), (b)-(e) show the appearance transformation while (f)-(j) exhibit the spatial/geometric transformations. Each augmentation can transform data stochastically with some internal parameters.

TABLE 5
Comparisons of different image augmentations in DMRNet++ on the CUHK-SYSU and PRW datasets.

Brightness	Hue	Blur	Noise	Bbox-Trans	Flip	Expand	Cutout	Crop	CUHK-SYSU		PRW	
									mAP	Rank-1	mAP	Rank-1
									93.3	94.1	48.8	84.6
✓									92.8	93.7	47.9	83.9
	✓								92.0	93.2	45.8	83.4
		✓							93.4	94.1	49.0	85.5
			✓						92.9	93.9	48.8	84.6
				✓					93.6	94.7	49.1	85.4
					✓				93.9	94.9	49.5	85.7
						✓			93.5	94.6	49.6	85.7
							✓		93.2	94.3	49.4	85.5
								✓	92.9	93.9	48.5	85.2
				✓	✓	✓			**94.3**	**95.3**	49.9	85.9
		✓		✓	✓	✓	✓		94.0	95.1	**50.2**	**86.1**

Effectiveness of the components. We investigate the effectiveness of each component in DMRNet++ on the CUHK-SYSU and PRW datasets. As shown in Tab. 4, the experiments are divided into four groups, and the first one shows the strong baseline. The effectiveness of our proposed TDF has been discussed above. The promotion is obvious by comparing the first two rows.

Then, we explore the effect of our proposed MRM, as shown in the second group of Tab. 4. With a slow-moving average network counterpart, the queue-style memory banks are consistently encoded. Without the look-up table preserving the centroids, the original OIM loss is extended to a multi-positive pairwise loss $\mathcal{L}_l$ in Eq. (7). It shows that when applying the MRM with $\mathcal{L}_l$, the performance is improved clearly, especially for the PRW datasets. This is caused by the hundreds of samples under some categories in PRW. In the baseline, the samples within a batch are pulled to their memorized centroids. With the proposed MRM, the labeled queue maintains sufficient positive samples with high consistency, providing stronger intra-class compactness.

In our proposed UCL, apart from the labeled anchors, we also take the unlabeled samples as anchors. Different from previous works that only view the unlabeled samples as negative pairs for the labeled samples, we firstly take the recognition process as semi-supervised learning. The results are shown in the third and fourth groups of Tab. 4. $\mathcal{L}_u$ and $\mathcal{L}'_u$ mean the different contrastive loss are applied on unlabeled samples, *i.e.*, Eq. (10) and the upper loss in Eq. (10), respectively. From the results, we can draw two observations. First, both $\mathcal{L}_u$ and $\mathcal{L}'_u$ can promote the performance clearly. It shows the underlying potentials of unlabeled samples. Second, $\mathcal{L}_u$ performs better than $\mathcal{L}'_u$. It exhibits the benefit of applying loose form contrastive loss on positive pairs with lower similarities, which is robustness to some false positive samples. The improvement is more obvious on the PRW dataset in which vast similar dresses may lead to more noise in positive pairs, and $\mathcal{L}_u$ can well suppress this issue. Third, besides the basic horizontal flipping, more augmentations enrich the sample pairs and promote the feature learning.

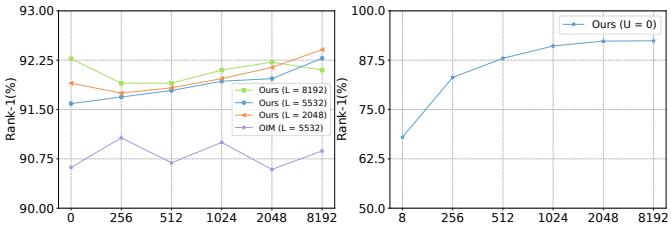
Comparisons with different augmentations. In our framework, the input images are applied with different augmentations, then processed by F and $\bar{F}_{id}$ .To analyze the impact of data augmentations, we apply several augmentations on DMRNet++. As Fig. 6 shows, (a) is the original image and the augmentations in the first row involve appearance transformation, such as color

distortion (color brightness (b), color hue(c)) [76], Gaussian blur (d) and Gaussian noise (e). The second row shows the spatial and geometric transformations, including our proposed bounding box transformation (f), image horizontal flipping (g), expansion [30] (h), cutout [77] (i), and random cropping (j).

The performance with different augmentations is shown in Tab. 5, the first row shows the baseline result that without augmentations. We conduct the experiments with individual image transformation firstly. It is observed that few appearance transformations suffice to learn good representations. We speculate that person re-ID relies heavily on color, so changing the appearance may lead to inferior performance. For the geometric transformations, image horizontal flipping is widely used in the person search task [5], [17], which also shows a positive effect in our experiments. The random crop can be seen as a zoom-in operation that produces larger training instances. In contrast, the image expansion is implemented as a zoom-out operation that creates more small training examples. In Tab. 5, the result is inferior to the baseline when applying random crop. One possibility is that some persons may be removed by this operation since there are several instances in a scene image. The image expansion is beneficial to small objects, thus it improves the performance clearly in PRW dataset that contains plentiful small persons. To preserve the robustness of the identification features with variable bounding boxes, we develop an augmentation called the bounding box transformation. It shifts the cropping area in a fixed range randomly, ensuring the completeness of the image and the randomness of the instance. Bounding box transformation is a kind of random crop for the person search task. Further, cutout could enhance the robustness of the model against occlusion, thus promoting the accuracy slightly. When composing augmentations, the quality of representation gains more improvements. For the CUHK-SYSU dataset, the augmentation set contains bounding box transform, image horizontal flipping and expansion. For the PRW dataset, the augmentations consist of bounding box transform, image horizontal flipping, gaussian blur, expansion and cutout.

Comparisons of memory bank mechanisms with varying sizes. We analyze the effect of different memory bank mechanisms with varying sizes, *i.e.*, the look-up table with OIM loss, and our memory-reinforced mechanism with pairwise loss. They are implemented on the same network, as described in Fig. 5(c). Here we remove the point-based strategy and $\mathcal{L}_u$ for fair comparisons. The results are shown in Fig. 7, L is the length of the look-up table or the queue with labeled samples, and U is the length of the queue with unlabeled ones. The comparisons provide the following observations.

- To explore the effect of taking unlabeled samples as negative pairs, we compare OIM (L = 5532) with our method (L = 2048/5532/8192) under different sizes of U. As shown in Fig. 7 (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 shown in Fig. 7(b), when U is set to zero, our method benefits from a larger L without the unlabeled queue. This is intuitive since more positive/negative sample pairs can be exploited.
- As shown in Fig. 7(a)(b), when two methods reach similar results, our method only uses a small size of labeled queue ($L = 1024, U = 0$), which is more efficient than OIM.



(a) Unlabeled queue size U (b) Labeled queue size L

Fig. 7. Comparison between two memory bank mechanisms with varying sizes. The numbers of labeled and unlabeled samples are denoted as L and U, respectively.

TABLE 6
The results of DMRNet and DMRNet++ with different detectors on the CUHK-SYSU and PRW datasets. D denotes the result of detection.

Methods	Detector	CUHK-SYSU			PRW		
		mAP	Rank-1	mAP(D)	mAP	Rank-1	mAP(D)
DMRNet	FCOS [39]	87.6	90.4	91.0	40.1	80.8	95.0
DMRNet++		90.4	92.9	90.7	45.3	83.4	95.1
DMRNet	ATSS [64]	91.1	92.7	91.9	43.6	82.4	95.4
DMRNet++		92.5	94.0	91.6	46.2	84.6	95.3
DMRNet	RetinaNet [20]	91.2	92.5	91.3	44.6	82.0	94.7
DMRNet++		93.2	94.3	91.3	50.5	86.2	94.8
DMRNet	Foveabox [65]	90.7	92.3	91.5	45.8	82.8	95.4
DMRNet++		92.9	94.2	91.6	50.1	86.5	95.1
DMRNet	RepPoints [40]	92.9	93.7	91.7	46.0	83.2	94.6
DMRNet++		94.3	95.3	91.8	50.2	86.1	94.8

Effectiveness on different detectors. As illustrated in Tab. 6, to verify the expandability of DMRNet++, different detection networks are incorporated into our framework, including RetinaNet [20], RepPoints [40], FCOS [39], ATSS [64] and Foveabox [65]. We not only consider the anchor-based detection networks (*e.g.*, RetinaNet), but also evaluate the anchor-free detectors (*e.g.*, RepPoints, FCOS). Tab. 6 exhibits the mAP and rank-1 of person search, and the mAP of jointly trained detector denoted as 'mAP (D)'. DMRNet++ consistently improves the DMRNet by significant margins on various detectors. Especially, when incorporated with RetinaNet, the mAP is promoted by 5.9% on the PRW dataset. This confirms the effectiveness and robustness of our method when extended to different detectors.

Effectiveness on different image resolutions. As shown in Tab. 7, we compare the proposed DMRNet++ with DMRNet under different image resolutions. The first two rows show the results under a single-scale training setting with 1333×800 and 1500×900 input sizes. A large image size generally benefits the learned feature embeddings, and the mAPs are improved on both datasets. The last two rows exhibit the performance with a multi-scale training strategy. Similar to [24], the longer side of the input image is randomly resized from 667 to 2000 during training, while the test image is re-scaled to a fixed size of 1500×900. From Tab. 7, it is observed that the results are significantly improved on the PRW dataset, while the performance

TABLE 7
The results of DMRNet and DMRNet++ with different input resolutions on the CUHK-SYSU and PRW datasets. The first two rows show the single-scale training while the last two rows are multi-scale training.

Methods	Resolution		CUHK-SYSU		PRW	
	train	test	mAP	Rank-1	mAP	Rank-1
DMRNet	1333 × 800	1333×800	92.9	93.7	46.0	83.2
DMRNet++			94.3	95.3	50.2	86.1
DMRNet	1500 × 900	1500×900	93.2	94.2	46.9	83.4
DMRNet++			94.4	95.5	51.0	86.8
DMRNet	[667, 2000]	1500×900	92.5	93.7	48.1	84.6
DMRNet++			93.7	94.6	50.5	86.3
DMRNet	[1333, 2666]	2000×1200	93.1	94.2	50.5	84.5
DMRNet++			94.5	95.7	52.1	87.0

TABLE 8
Evaluations of the proposed methods under the cross-dataset setting. PRW→CUHK-SYSU: the model is trained on PRW dataset while tested on CUHK-SYSU, and vice versa.

Methods	PRW→CUHK-SYSU			CUHK-SYSU→PRW		
	mAP	Rank-1	mAP(D)	mAP	Rank-1	mAP(D)
OIM-base	49.4	54.9	65.1	20.5	42.5	87.6
DMRNet	50.8	55.7	64.8	25.3	71.9	88.2
DMRNet++	52.1	57.5	64.4	27.5	76.6	87.2

on the CUHK-SYSU dataset is decreased slightly. This implies the fragileness of CUHK-SYSU with small input sizes. We further increase the resolution under the multi-scale training, as shown in the last row. The longer side of the input image is randomly resized from 1333 to 2666 during training, and the test image is fixed to 2000×1200. The results are increased on both datasets. Especially on the PRW dataset, we achieve 52.1% on mAP and 87.0% on rank-1 accuracy. Moreover, It is seen that DMRNet++ consistently improves the DMRNet by significant margins under different resolution settings. The results further show the effectiveness of our proposed DMRNet++.

Effectiveness on cross-dataset scenario. To validate the generalization ability of our framework, we conduct cross datasets comparison between datasets. Specifically, we directly utilize the model trained on a source dataset (*e.g.*, CUHK-SYSU) to evaluate on a different target dataset (*e.g.*, PRW). We compare our proposed DMRNet and DMRNet++ with the re-implemented OIM baseline. The results are presented in Tab. 8, from which we draw two observations.

First, three methods exhibit similar detection performance while different search results. This implies the DMRNet++ manifests a strong discriminative ability under the cross-dataset setting. When trained on the CUHK-SYSU and tested on the PRW dataset, DMRNet++ outperforms OIM [15] by a large margin on mAP and rank-1 accuracy.

Second, although the accuracy drops for both cross-dataset scenarios, the model trained with CUHK-SYSU is slightly better than PRW. Since the CUHK-SYSU dataset contains diverse scenes, it shows better capability when transferring, especially on the mAP of detection.

Runtime comparisons. Efficiency is one advantage of our framework. To show the efficiency of our method, we compare the runtime with other methods in the inference stage. The average runtime of the detection and re-ID for a panorama image is reported. For a fair comparison, we test the models with an input image size of 900 × 1500, which is the same as other works [2], [12], [17]. Since the methods are implemented with different GPUs, we also report the TFLOPs. As shown in Tab. 9, upon normalization with TFLOPs, our framework is much faster than the two-step method MGTS [12]. Moreover, our method is more efficient than NAE+ [2], which is the current state-of-the-art one-step method. Compared with DMRNet, DMRNet++ exhibits higher performance with similar efficiency.

TABLE 9
Runtime comparisons of different methods.

Methods	GPU	TFLOPs	Time
MGTS [12]	K80	8.7	1296
QEEPS [17]	P6000	12.0	300
NAE [2]	V100	14.1	83
NAE+ [2]	V100	14.1	98
DMRNet	V100	14.1	66
DMRNet++	V100	14.1	67

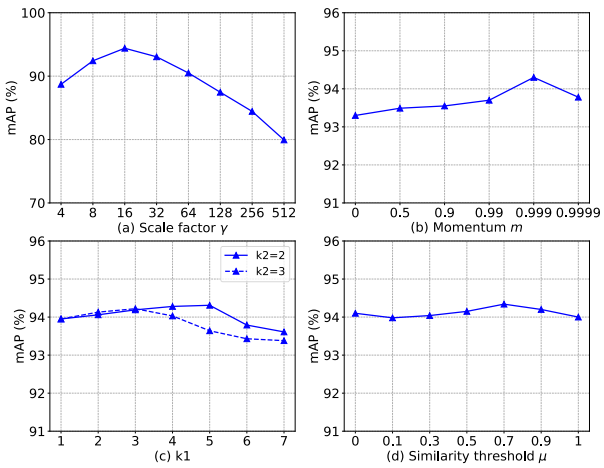


Fig. 8. Performance of our framework with different values of scale factor γ, momentum m, k-reciprocal neighbors and similarity threshold μ. The mAP (%) on the CUHK-SYSU dataset is shown.

5.5 Parameter Analysis

We analyze the important parameters in our proposed DMRNet++. All the parameters are tuned on the CUHK-SYSU dataset, and the same values are directly employed on other datasets. Experimental results are presented in Fig. 8.

Scale factor γ. Scale factor determines the largest scale of each similarity score. As shown in Fig. 8(a), we study the effects of the scale factor on DMRNet++ by vary γ from 4 to 512. We observe that the performance is robust in the interval of $[8, 32]$ and our framework achieves the optimal result when setting the scale factor γ as 16.

Momentum factor m. The performance of our method with different momentum factors is shown in Fig. 8(b). We obtain the optimal result when m is set to 0.999. This indicates a relatively

Fig. 9. Illustration of the point-based sampling positions (7×7) on the backbone network. Different colors denote different identities, and the images are from the PRW dataset.

large momentum facilitates learning discriminative identification features. When m is zero, it means the parameters of f and $\bar{F}_{id}$ are identical. Surprisingly, with the least consistent encoding, our mechanism still slightly outperforms the look-up table, showing the effectiveness of the queues.

K-reciprocal neighbors. Fig. 8(c) shows the effect on different k-reciprocal neighbors. Here we utilize two sizes in Eq. (9), where $N(x_u, k_1)$ and $N(q_i^u, k_2)$. When k_1 is equal to 1, it means only the two augmented views of the same instance are considered as a positive pair. The solid curve and dashed curve show the results with different values of k_1 when $k_2 = 2$ and $k_2 = 3$, respectively. Obviously, the performance grows as k_1 increases in a reasonable range, and achieves the optimal result when $k_1 = 5$ and $k_2 = 2$. As the value of k_1 continues growing, more false positive samples are included in the k-reciprocal set. Taking these samples as positive pairs causes a decline of performance.

Similarity threshold μ. In Eq. (10), the similarity threshold μ is adopted to assign different forms of contrastive loss. For the reliable positive pairs with similarities larger than μ, the upper loss in Eq. (10) with hard positive mining is applied. The positive samples with similarities less than μ are applied with a loose constraint, the lower one in Eq. (10). Fig. 8(d) shows the effect of different thresholds. When μ approaches 0, all the positive pairs of unlabeled anchors are supervised by the upper loss in Eq. (10). The inevitable noise makes sub-optimal results. When μ is set to 1, the supervision is only the lower loss in Eq. (10). The loose form cannot provide strong constraints. The optimal result is achieved when setting μ as 0.7.

5.6 More Qualitative Analysis.

Illustration of the point-based sampling positions. To analyze the effect of the point-based spatial sampling, we visualize the sampling positions on the backbone network on the PRW dataset. As illustrated in Fig. 9, different from the initial points that sampled uniformly in the bounding boxes, the augmented points focus more on the human body with the additional offsets. Note that different colors represent different identities. Compared to the box-based sampling approach with fixed rectangles, our point-based spatial sampling provides a more flexible way with the irregular points.

Illustration of the k-reciprocal set. To evaluate the quality of assigned positive pairs for unlabeled anchors, we visualize the k-reciprocal set in the training process. As illustrated in Fig. 10, given an unlabeled anchor x_u (green bounding box), its 5-nearest

Fig. 10. Visualizations of k-reciprocal nearest neighbors of unlabeled anchors x_u. Given x_u (green bounding box), its k-nearest neighbors $N(x^u, k)$ (blue bounding boxes) are shown on the right. Below each neighbor q_i^u, its k-nearest neighbors $N(q_i^u, k)$ are exhibited. If x^u and q_i^u are the k-nearest neighbors reciprocally, q_i^u is selected to the k-reciprocal set (red dotted rectangle) as the positive pairs.

neighbors $N(x^u, k)$ are shown in the right side, denoted as q_i^u. Below each neighbor q_i^u , its 2-nearest neighbors $N(q_i^u, k)$ are exhibited. In Fig. 10 (a), it can be seen that the unlabeled anchor x_u is included in the 2-nearest neighbors of each q_i^u, thus the five samples are selected to the k-reciprocal set (red dotted rectangle) as the positive pairs of x_u. As Fig. 10 (b) exhibits, the 2-nearest neighbors of q_5^u do not contain the unlabeled anchor, while x_u is only included in the 2-nearest neighbors of $q_1^u, q_2^u, q_3^u, q_4^u$. Therefore, the k-reciprocal neighbors of x_u are composed of these four samples. As shown in Fig. 10 (c), the positive sample contains only the one that with different augments of x_u. This is the extreme case and at least one positive sample pair is guaranteed.

Since this positive pair assignment is conducted on the unlabeled queue, we have no identity annotation for each person. Therefore, except for the unlabeled anchors are denoted as green bounding boxes, other selected neighbors are all shown with the blue boxes, including the right or wrong matches. Combined with the context, we can judge that most selected k-reciprocal neighbors have the same identity as the unlabeled anchor. This verifies the importance of exploring the positive pairs for unlabeled anchors and shows the effectiveness quantificationally.

Visualization results on the CUHK-SYSU dataset. As shown in Fig. 11, we present the visualization results of both DMRNet and DMRNet++ for comparison, in which the rank-3 search results on the CUHK-SYSU dataset are exhibited. Given each query person in the green bounding box, the search results of DMRNet and DMRNet++ are shown on the left and right parts, respectively. Red and blue bounding boxes represent the wrong and correct results, respectively. As Fig. 11 shows, the task of person search is challenging due to the severe occlusion, low resolution, and

Fig. 11. Qualitative search results on the CUHK-SYSU dataset. Given the query persons (green boxes), we show the rank-3 search results of DMRNet and DMRNet++. Red/blue boxes represent the wrong/correct results, respectively.

changing illumination. From the search result of DMRNet, we observe that the most wrong matches are caused by similar appearances, especially under the same scene with the ground truth person. This indicates that excessive contextual information may be involved with the box-based spatial sampling method. Based on the point-based spatial sampling in DMRNet++, irregular points are learned to locate the positions containing discriminative features. More informative sample pairs are explored by applying contrastive learning on unlabeled identities. This further promotes the generalization of our model. When evaluating DMRNet++, the results show that our enhanced network can locate and match all the target persons correctly.

Visualization results on the PRW dataset. The visualization of search results on the PRW dataset is shown in Fig. 12. We present the rank-5 results of three methods for comparison, including re-implemented OIM [15], DMRNet [23] and DMRNet++. It is obvious that the main difficulties lie in resolution variations and similar appearance when evaluating on PRW. The query person in Fig. 12(a) exhibits a small size, leading to the failure of rank-4 for the OIM. DMRNet can match most targets and the DMRNet++ shows the optimal results. This shows the robustness of DMRNet++ on small targets. In Fig. 12(b), the query person wears a white shirt and black trousers. For the OIM method, it is easy to confuse the persons having similar appearances. Although the search results of OIM have the same clothing, its rank-5 all failed to discover the target person. DMRNet shows a better result, and the wrong match only occurs at the small targets. DMRNet++ can locate and match each target, exhibiting higher performance. It indicates more discriminative embeddings are ensured based on our point-based sampling and UCL loss. However, our DMRNet++ also encounters incorrect matches in some conditions. As shown in Fig. 12(c), the small target with similar clothes causes the wrong match. In real-world applications, due to the occlusion, pose variation, illumination, and viewpoint, the task of person search is still challenging.

6 Conclusion

In this work, we propose an enhanced decoupled and memory-reinforced network (DMRNet++) for one-step person search. Our main purpose is to address the challenges in one-step pipelines, *i.e.*, conflicting objectives, inconsistent memory bank and underutilized unlabeled identities. To tackle these issues, we improve the network design, memory bank mechanism and loss function respectively. All these designs aim at producing more discriminative features for retrieval. First, we propose a task-decoupled framework (TDF) that substantially decouples the two sub-tasks. This design alleviates optimization conflicts on the RoI head, facilitating the features learning for recognition. Second, we introduce a memory-reinforced mechanism (MRM) to ensure the consistency of memory bank. By incorporating a slow-moving average of the network, the memorized features can be consistently encoded, promoting effective identification learning. Third, we develop an unlabeled-aided contrastive loss (UCL) to exploit the potentials of the unlabeled identities. By applying contrastive learning on the unlabeled identities, more informative positive and negative sample pairs are explored, promoting highly discriminative identification feature embeddings.

Due to the massive simplification of the pipeline design, our model is easy to train and efficient to deployment. It sets a new state-of-the-art among one-step methods and outperforms a lot of existing two-step methods. We hope that our findings can encourage a shift in the framework of the one-step person search and drive more research in this field. In the future, we will continue to explore the person search task with the help of 3D prior knowledge [78], [79].

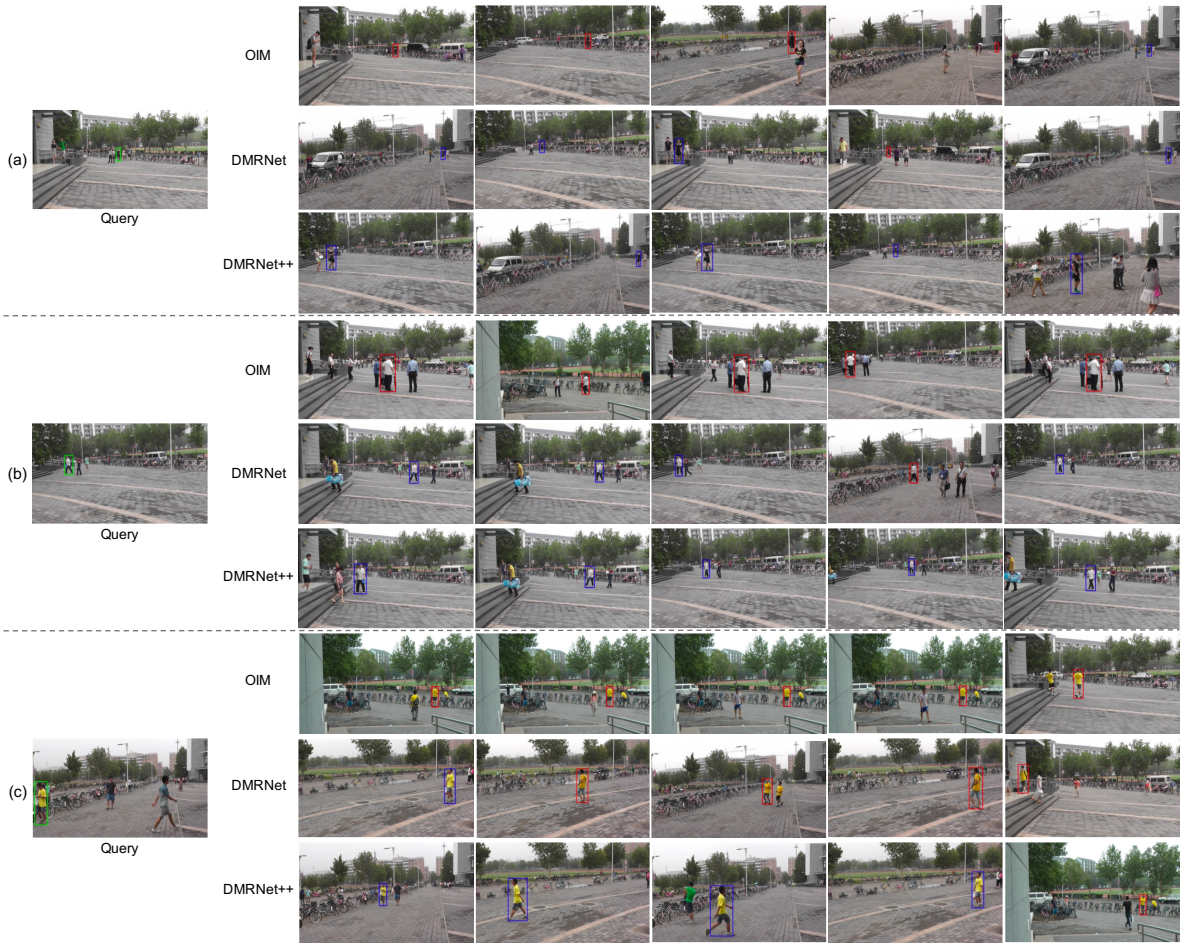


Fig. 12. Qualitative search results on the PRW dataset. Given the query persons (green boxes), we show the top rank-5 search results of three methods, including OIM [15], our proposed DMRNet and DMRNet++. Red/blue boxes represent the wrong/correct results, respectively.

ACKNOWLEDGMENTS

This work was supported by 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

[1] W. Dong, Z. Zhang, C. Song, and T. Tan, "Bi-directional interaction network for person search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[2] D. Chen, S. Zhang, J. Yang, and B. Schiele, "Norm-aware embedding for efficient person search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[3] Y. Zhong, X. Wang, and S. Zhang, "Robust partial matching for person search in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[4] C. Wang, B. Ma, H. Chang, S. Shan, and X. Chen, "Tcts: A task-consistent two-stage framework for person search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[5] W. Dong, Z. Zhang, C. Song, and T. Tan, "Instance guided proposal network for person search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[6] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2011.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[8] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, and S. Wang, "Learning part-based convolutional features for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[9] X. Qian, Y. Fu, T. Xiang, Y.-G. Jiang, and X. Xue, "Leader-based multi-scale attention deep architecture for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 371–385, 2019.

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

[11] X. Lan, X. Zhu, and S. Gong, "Person search by multi-scale matching," in *European Conference on Computer Vision*, 2018.

[12] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a mask-guided two-stream cnn model," in *European Conference on Computer Vision*, 2018.

[13] C. Han, J. Ye, Y. Zhong, X. Tan, C. Zhang, C. Gao, and N. Sang, "Re-id driven localization refinement for person search," in *IEEE International Conference on Computer Vision*, 2019.

[14] X. Chang, P.-Y. Huang, Y.-D. Shen, X. Liang, Y. Yang, and A. G. Hauptmann, "Rcaa: Relational context-aware agents for person search," in *European Conference on Computer Vision*, 2018.

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

[16] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei, and J. Feng, "Ian: the individual aggregation network for person search," *Pattern Recognition*, vol. 87, pp. 332–340, 2019.

[17] B. Munjal, S. Amin, F. Tombari, and F. Galasso, "Query-guided end-to-

end person search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

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

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

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision*, 2017.

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

[22] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[23] C. Han, Z. Zheng, C. Gao, N. Sang, and Y. Yang, "Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1505–1512.

[24] Y. Yan, J. Li, J. Qin, S. Bai, S. Liao, L. Liu, F. Zhu, and L. Shao, "Anchor-free person search," *arXiv:2103.11617*, 2021.

[25] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[26] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[27] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Advances in Neural Information Processing Systems*, 2014.

[28] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *British Machine Vision Conference*, 2009.

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

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

[31] R. Girshick, "Fast r-cnn," in *IEEE International Conference on Computer Vision*, 2015.

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

[33] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2018.

[34] G. Song, Y. Liu, and X. Wang, "Revisiting the sibling head in object detector," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[35] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *IEEE International Conference on Computer Vision*, 2017.

[36] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[37] Z. Cai and N. Vasconcelos, "Cascade r-cnn: high quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[38] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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

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

[41] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017.

[42] A. Barman and S. K. Shah, "A graph-based approach for making consensus-based decisions in image search and person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[43] Y. Shen, T. Xiao, S. Yi, D. Chen, X. Wang, and H. Li, "Person re-identification with deep kronecker-product matching and group-shuffling random walk," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[44] J. Li, S. Zhang, Q. Tian, M. Wang, and W. Gao, "Pose-guided representation learning for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[45] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 3037–3045, 2018.

[46] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *European Conference on Computer Vision*, 2018.

[47] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep crf for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[48] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *IEEE International Conference on Computer Vision*, 2017.

[49] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *IEEE International Conference on Computer Vision*, 2017.

[50] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

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

[52] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[53] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM transactions on multimedia computing, communications, and applications (TOMM)*, vol. 14, no. 1, pp. 1–20, 2017.

[54] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *ACM International Conference on Multimedia*, 2018.

[55] C. Han, R. Zheng, C. Gao, and N. Sang, "Complementation-reinforced attention network for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3433–3445, 2019.

[56] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *European Conference on Computer Vision*, 2016.

[57] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[59] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015.

[60] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems*, 2017.

[61] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[62] C. Doersch, A. Gupta, and A. Zisserman, "Crosstransformers: spatially-aware few-shot transfer," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 981–21 993, 2020.

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

[64] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

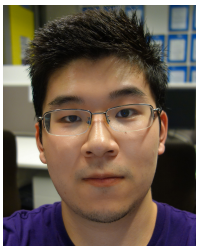
[65] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "Foveabox: Beyound anchor-based object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 7389–7398, 2020.

[66] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

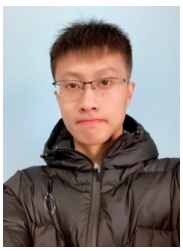
[67] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[68] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
[69] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
[70] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *IEEE International Conference on Computer Vision*, 2015.
[71] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision*, 2015.
[72] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan, "Neural person search machines," in *IEEE International Conference on Computer Vision*, 2017.
[73] L. Zhang, Z. He, Y. Yang, L. Wang, and X.-B. Gao, "Tasks integrated networks: Joint detection and retrieval for image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
[74] H. Kim, S. Joung, I.-J. Kim, and K. Sohn, "Prototype-guided saliency feature learning for person search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
[75] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
[76] A. G. Howard, "Some improvements on deep convolutional neural network based image classification," *arXiv preprint arXiv:1312.5402*, 2013.
[77] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.
[78] J. Rajasegaran, G. Pavlakos, A. Kanazawa, and J. Malik, "Tracking people by predicting 3d appearance, location and pose," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2740–2749.
[79] Z. Zheng, N. Zheng, and Y. Yang, "Parameter-efficient person re-identification in the 3d space," *arXiv:2006.04569*, 2020.

Chuchu Han received the B.S. degree in School of Automation from ChongQing University, China, in 2017. She is currently pursuing the Ph.D. degree in pattern recognition and intelligent systems from Huazhong University of Science and Technology, China. Her research interests include surveillance video analysis, and image retrieval.



Zhedong Zheng received the Ph.D. degree from the University of Technology Sydney, Australia, in 2021 and the B.S. degree from Fudan University, China, in 2016. He is currently a postdoctoral research fellow at Sea-NExT Joint Lab, School of Computing, National University of Singapore. He was an intern at Nvidia Research (2018) and Baidu Research (2020). His research interests include robust learning for image retrieval, generative learning for data augmentation, and unsupervised domain adaptation.



Kai Su received the B.S. degree from Soochow University, China, in 2016, and the M.S. degree from the PALM lab, Southeast University, China, in 2019. Currently, he is an R&D engineer at Bytedance AI Lab. His main research interests include computer vision and machine learning.

Dongdong Yu recieved the Ph.D. degree in Pattern Recognition and Intelligent Systems from the Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is a Researcher at ByteDance AI Lab. His research interests focus on image classification, human keypoint, and scene parsing.

Zehuan Yuan received the B.S. degree and Ph.D. degree both from Department of computer science and technology, Nanjing University,China. He is currently a researcher in Bytedance AI Lab. He has published more than fifteen academic papers on the top-tier international journals and conferences, such as ICML, CVPR, IJCAI,AAAI,ICLR,ECCV, etc. His research interests lie in computer vision and machine learning.

Changxin Gao received the Ph.D. degree in pattern recognition and intelligent systems from Huazhong University of Science and Technology in 2010. He is currently an associate professor at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China. His research interests are pattern recognition and surveillance video analysis.

Nong Sang received the Ph.D. degree in pattern recognition and intelligent systems from Huazhong University of Science and Technology in 2000. He is currently a professor at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China. His research interests include object detection and recognition, object tracking, image/video semantic segmentation, intelligent processing and analysis of surveillance videos.

Yi Yang received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He is currently a professor with Zhejiang University, China. He was a Post-Doctoral Research with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. His current research interest includes machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval, video analysis and video semantics understanding.