Deep Intra-Image Contrastive Learning for Weakly Supervised One-Step Person Search

Jiabei Wang, Yanwei Pang, *Senior Member, IEEE*, Jiale Cao, Hanqing Sun, Zhuang Shao,
and Xuelong Li, *Fellow, IEEE*

arXiv:2302.04607v1 [cs.CV] 9 Feb 2023

Abstract—**Weakly supervised person search aims to perform joint pedestrian detection and re-identification (re-id) with only person bounding-box annotations. Recently, the idea of contrastive learning is initially applied to weakly supervised person search, where two common contrast strategies are memory-based contrast and intra-image contrast. We argue that current intra-image contrast is shallow, which suffers from spatial-level and occlusion-level variance. In this paper, we present a novel deep intra-image contrastive learning using a Siamese network. Two key modules are spatial-invariant contrast (SIC) and occlusion-invariant contrast (OIC). SIC performs many-to-one contrasts between two branches of Siamese network and dense prediction contrasts in one branch of Siamese network. With these many-to-one and dense contrasts, SIC tends to learn discriminative scale-invariant and location-invariant features to solve spatial-level variance. OIC enhances feature consistency with the masking strategy to learn occlusion-invariant features. Extensive experiments are performed on two person search datasets CUHK-SYSU and PRW, respectively. Our method achieves a state-of-the-art performance among weakly supervised one-step person search approaches. We hope that our simple intra-image contrastive learning can provide more paradigms on weakly supervised person search. The source code is available at https://github.com/jiabeiwangTJU/DICL.**

Index Terms—**Weakly supervised person search, contrastive learning, spatial-invariant, occlusion-invariant.**

Fig. 1. Intra-image contrast strategies of different weakly supervised one-step methods. CGPS (a) performs dense prediction contrasts using a single network with the input of an entire image. R-SiamNet (b) performs a one-to-one ground-truth contrast using Siamese network with the inputs of an entire image and cropped persons. Compared to these shallow intra-image contrast strategies, our method (c) exploits deep intra-image contrastive learning. We consider many-to-one Siamese contrasts between two branches of Siamese network and dense prediction contrasts in one branch of Siamese network. In addition, we perform occlusion-invariant contrast by randomly masking a portion of person in one branch of Siamese network.

I. INTRODUCTION

PERSON search is aimed at searching a given query person in a set of gallery images, which is regarded as a joint task of pedestrian detection and re-identification (re-id) [5], [6], [7]. Currently, most existing person search methods [2], [8], [9] are fully-supervised, which are developed based on person bounding-box annotations and identity annotations. Compared to the bounding-box annotations of pedestrian detection, the identity annotations of person re-id are substantially expensive and time consuming. On the one hand, it is difficult and expensive to obtain a large number of identity annotations across different scenes. There are nearly 72.7% and 17.4% person bounding boxes without identity annotations in CUHK-SYSU and PRW datasets [10], respectively. On the other hand, it is time consuming to accurately annotate the persons with same identity across different scenes, especially when the

person has a large variance in appearance, scale, occlusion, *etc*. Therefore, it is necessary to explore person search with only bounding-box annotations, termed as weakly supervised person search.

Recently, several contrastive learning based one-step methods have been proposed to perform weakly supervised person search. Two contrastive learning strategies are used in these methods, including memory-based contrast and intra-image contrast[1]. Memory-based contrast builds a cluster-level re-id memory bank over all training images, and employs it to supervise the re-id feature learning in current image. Intra-image contrast aims to exploit the useful contrast information within image for re-id feature learning. We aim to improve the intra-image part. Fig. 1 shows intra-image contrasts used in two typical weakly-supervised one-step methods CGPS [11] and R-SiamNet [10]. As shown in Fig. 1(a), CGPS explores dense prediction contrasts using a single network. It enforces the re-id features belonging to one person be close, and pushes the features belonging to different persons apart. In Fig. 1(b), R-SiamNet adopts a Siamese network for one-to-one Siamese contrast. It extracts two re-id features of a ground-truth in two branches and enforces these two re-id features with or without context be consistent. These methods achieve an initial success on weakly supervised one-step person search. However, it

J. Wang, Y. Pang, J. Cao, and H. Sun are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (E-mail: {jiabeiwang,pyw,connor,hqSun}@tju.edu.cn).

Z. Shao is with Warwick Manufacturing Group, University of Warwick, Coventry, CV4 7AL, UK (E-mail: Zhuang.Shao@warwick.ac.uk).

X. Li is with the School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, P.R. China (E-mail: li@nwpu.edu.cn).

[1]Intra-image contrast can also mean intra-batch contrast during training. For simplicity, we use intra-image contrast here.

still lags far behind fully supervised one-step person search approaches [12], [13].

We argue that one of key reasons is the insufficient information mining with shallow intra-image contrast, which cannot effectively deal with *spatial-level* and *occlusion-level* variance of person instance. The detected person in different images could have a large variance in scale or an inaccurate bounding-box localization (spatial-level variance). In addition, the detected person is sometimes occluded by surrounding objects (occlusion-level variance). As a result, it becomes difficult to match the re-id features of a same person across images due to these spatial-level and occlusion-level variance.

To address this issue caused by spatial-level and occlusion-level variance, we present a novel deep intra-image contrastive learning based method (DICL) that fully exploits instance-level information of each person in an image (Fig. 1(c)). Our DICL is built on a Siamese network with two novel contrast modules: spatial-invariant contrast (SIC) and occlusion-invariant contrast (OIC). In the Siamese network, a search branch extracts the re-id features from an entire image, while an instance branch extracts the re-id features from cropped ground-truths. Our SIC not only enforces many-to-one feature consistency between two branches of Siamese network, but also conducts dense prediction contrasts in the search branch. With many-to-one and dense contrasts, our SIC can learn scale-invariant and location-invariant features to deal with spatial-level variance. To deal with occlusion-level variance, our OIC performs the masking strategy in the search branch and enhances re-id feature consistency between two branches. Experimental results on two classical person search datasets CUHK-SYSU [5] and PRW [14] show the effectiveness and superiority of proposed DICL. Our contributions are summarized as follows:

- We propose a novel weakly supervised one-step person search framework via deep intra-image contrastive learning (DICL). Without the identity annotations, DICL fully mines the intra-image information to deal with spatial-level and occlusion-level variance.
- We design two intra-image contrast modules: spatial-invariant contrast (SIC) and occlusion-invariant contrast (OIC). SIC performs many-to-one Siamese contrast and dense prediction contrast to learn scale-invariant and location-invariant features. OIC learns occlusion-invariant features with masking strategy in search branch.
- Our method achieves favorable results on CUHK-SYSU and PRW, which outperforms previous weakly supervised one-step approaches. On CUHK-SYSU test set, it achieves the mAP of 87.4% and top-1 accuracy of 88.8%. On PRW test set, it achieves the mAP of 35.5% and top-1 accuracy of 80.9%.

II. RELATED WORK

In this section, we first introduce person search, including fully/weakly supervised person search. After that, we give a brief review on unsupervised person re-identification (re-id) and contrastive learning.

A. Person Search

Person search is a joint task of pedestrian detection and person re-id, which plays an important role in video surveillance [15], [16], [3], [4]. Existing person search methods can be divided into two main categories: two-step methods and one-step methods. Two-step methods [17], [18] detect pedestrians first and then identify them in the cropped person images, which perform two sub-tasks separately. Zheng *et al.* [14] explored different detectors and re-id methods for person search. Chen *et al.* [17], [2] argued that person foreground region plays more important role for re-id and proposed a mask-guided two-stream CNN model to emphasize the foreground information. Yao *et al.* [1] proposed to consider the objectness and repulsion information for similarity computation. Wang *et al.* [18], [19] employed an identity-guided query detector to generate query-like proposals and designed a detection results adapted re-ID model for improved re-id.

One-step methods [12], [13] perform pedestrian detection and person re-id in a unified end-to-end framework. Compared to two-step methods, one-step methods are simple but challenging, which needs to deal the diverse goals of pedestrian detection and person re-id. Most existing one-step methods are built on object detection framework Faster R-CNN [20], [21]. Xiao *et al.* [5] made an initial attempt by adding a fully-connected layer after R-CNN head network and generating the re-id features for online instance matching. Chen *et al.* [22] proposed to decompose detection and re-id features in the polar coordinate system. Li *et al.* [23] employed a cascaded head-networks for detection and re-id, where the detection is first performed and the re-id is second performed. Han *et al.* [24], [25] proposed to decouple detection and re-id by using RoI features only for re-id. Dong *et al.* [26] introduced a bi-directional interaction network taking both entire images and the cropped instance images as inputs during training, using accurate human appearance information from person patches to discriminate identities. Han *et al.* [27] perform detection, re-identification, and part classification together for improved person search. In addition, some other methods focus on improving online instance matching loss, such as HOIM [13] and OIM++ [28]. Instead of adopting anchor-based detector Faster R-CNN, Yan *et al.* [12] proposed to adopt anchor-free detector FCOS [29] for anchor-free person search, which avoids the hand-crafted anchor design. Two recent works PSTR [9] and COAT [30] employ the transformer-based network for person search.

Weakly supervised person search: Recently, several methods focus on weakly supervised person search. For example, two-step method CGUA [31] proposes a context-guided cluster algorithm and unpaired-assisted memory to take advantage of plenty of unpaired persons. One-step method CGPS [11] exploits detection context for dense prediction contrast using a single network. R-SiamNet [10] employs sparse ground-truths for intra-image contrast with a Siamese network. We argue that these intra-image contrasts in one-step methods are shallow, which does not fully exploit intra-image information.

Fig. 2. Architecture (a) of our deep intra-image contrastive learning (DICL) based method using a Siamese network, which has a search branch and an instance branch. DICL conducts intra-image contrast using two novel modules: spatial-invariant contrast (SIC) and occlusion-invariant contrast (OIC). SIC (b) performs many-to-one contrast in two branches of Siamese network and dense contrasts in all predictions of an image. OIC (c) enhances feature consistency using the masking strategy.

B. Unsupervised Person Re-ID

Unsupervised person re-id learns discriminative features to identify different persons in unlabeled cropped images. Most deep learning-based methods focus on the unsupervised domain adaption (UDA) task, which is with/without ID labels in source domain and without ID labels in target domain. The methods based on pseudo labels use clustering algorithm to classify instance features, such as k-means, DBSCAN [32], and Finch [33]. For example, BUC [34] treats each individual image as an identity and applies a bottom-up clustering to the feature embedding reducing the number of classes. SPCL [35] proposes a self-paced contrastive learning strategy with a novel clustering reliability criterion to generate more reliable clusters. Some methods utilize cross-domain transfer learning [36], [37]. DAPS [38] presents a unsupervised domain adaptive person search paradigm.

C. Contrastive Learning

Contrastive learning learns feature representation from positive or negative data pairs. SimCLR [39] introduces a learnable non-linear transformation head to train two augmentation paths with large batch sizes and long training steps. BYOL [40] proposes an online and target two interactive networks slowly updating with only positive pairs. SimSiam [41] represents a simple Siamese network without negative sample pairs, large batches or momentum encoders, using a stop-gradient operation to prevent collapsing. ContrastiveCrop [42] is a plug-and-play crop module for contrastive cropped pairs, which avoids false positives pairs and similar appearances pairs.

III. OUR METHOD

In this section, we first introduce the motivation, second present the overall architecture of proposed method, third describe two novel modules, and finally give the processes during training and inference.

A. Motivation

Contrastive learning has been initially used for weakly supervised one-step person search [11], [10], including memory-based contrast and intra-image contrast. However, the performance still lags far behind fully supervised one-step person search. We argue that one of the key reasons is that these methods do not fully mine intra-image information to deal with spatial-level and occlusion-level variance of person instance. For example, the detected person under different scenes may have different scales or inaccurate bounding-box localization, which is termed as spatial-level variance. The occlusion-level variance is that the detected person is occluded by other objects sometimes. As a result, it becomes challenging to accurately match the re-id features of same person across various images due to these spatial-level and occlusion-level variance.

To bridge the performance gap between weakly supervised one-step methods and fully supervised one-step methods, we present a novel deep intra-image contrastive learning based method (DICL) to fully exploit instance-level information of each person image considering spatial-level and occlusion-level variance of person instance.

B. Overall Architecture

Fig. 2 gives the overall architecture of our DICL. We adopt the Siamese network that contains a search branch and an instance branch. The search branch extracts the features of entire image for detection and re-id, while the instance branch

extracts the re-id features of fixed-size ground-truth pedestrians cropped and resized from the entire image. The feature extraction is performed based on Faster R-CNN [20]. Different to original Faster R-CNN, we make some modifications as follows. We use the fused feature map of last two layers of backbone for region proposal network (RPN) and RoI head network. The RoI head network consists of one 3×3 deformable layer [43], two 3×3 convolutional layers, and three parallelled fully-connected layers. In search branch, we first use an RoIAlign layer to extract the features of proposals and second employ RoI head network to perform classification, regression, and re-id feature generation. In instance branch, we remove the RoI layer and directly extract the re-id features using RoI head network.

Based on the re-id features generated in search branch and instance branch, we conduct intra-image contrastive learning and memory-based contrastive learning during training. For intra-image contrastive learning, we present a novel deep intra-image contrastive learning strategy that includes a spatial-invariant contrast module and an occlusion-invariant contrast module, described below. For memory-based contrastive learning, we build a cluster-level re-id memory bank using SPCL [35], and supervise the re-id feature learning like [11]. The memory bank is initialized with the averaged re-id features of two branches.

C. Spatial-Invariant Contrast

The same person under different cameras appears with the variance in scale. In addition, the detection cannot guarantee an accurate bounding-box localization like the ground-truth annotation. Therefore, the detected person in different images may have different scales or inaccurate bouding-box localization. We call it as spatial-level variance. The spatial-level variance will affect the accurate re-id matching across images. To deal with this spatial-level variance, we propose a spatial-invariant contrast (SIC) module that performs many-to-one Siamese contrasts between two branches of Siamese network and dense prediction contrasts in search branch.

The many-to-one Siamese contrast aims to enforce feature consistency between multiple predicted re-id features of a person in search branch and one re-id feature of a ground-truth in instance branch. We assign the predicted re-id features to the person when the corresponding predicted bounding-boxes have the IoU overlap higher than a given threshold with the person ground-truth bounding-box. Assuming that an image has n^p predictions and n^g ground-truths, the re-id features in search branch are represented as $f_i, i = 1, .., n^p$, and the re-id features in instance branch are represented as $f_j^g, j = 1, ..., n^g$. We enforce the re-id features of the same person in search branch to be consistent with re-id feature of corresponding ground-truth, where the loss of a prediction is written as

$$L_{mto}(i) = 1 - s(f_i, f_{j(i)}^g), i = 1, ..., n^p, \qquad (1)$$

where s represents the cosine distance similarity, and $j(i)$ represents the i-th predicted re-id feature corresponds to the j-th ground-truth. The dense prediction contrast aims to enforce the re-id features of the same person be closer and push the re-id features of different persons apart. The loss of dense contrast is formulated by triplet loss with the hard mining strategy [44], which is written as

$$L_{tri}(i) = m + \max_{f \in f_i^+} d(f_i, f) - \min_{f \in f_i^-} d(f_i, f), \qquad (2)$$

where d is a distance metric function, f_i^+ indicates a positive sample set that has the same identity with predicted f_i, f_i^- indicates a negative sample set that has different identities with predicted f_i, and m is a distance margin parameter. For a fair comparison, many-to-one Siamese contrast adopts the similar losses used in CGPS and R-SiamNet, and do not explore how to design a better contrastive loss, like NCE.

The many-to-one Siamese contrast is able to enforce the pedestrian re-id features of different scales and bounding-box locations to be consistent with each other, while dense prediction contrast makes the re-id features of the same person tighter with each other meanwhile keeps apart the features with different ids. Therefore, our SIC tends to learn discriminative scale-invariant and location-invariant, which can provide an accurate re-id matching even when the detected bounding-box is inaccurate or has a large variance in scale.

D. Occlusion-Invariant Contrast

Occlusion is very common in our daily life, which leads to a negative effect on person re-id. We call it as occlusion-level variance. To deal with this occlusion-level variance, we propose an occlusion-invariant contrast (OIC) module that performs the contrast between masked re-id feature and unmasked re-id feature. There are three different masking strategies, including masking on search branch, masking on instance branch, and masking on both search and instance branches. We observe that the masking on search branch performs better, as shown in Table V. The reason could be that instance branch without masking can provide consistent and occlusion-free feature, which can guide search branch to learn occlusion-invariant re-id feature.

Given a person in an image, we divide a person bounding-box into 14×6 grids, and randomly erase a portion of grids with pixel mean value. We perform the masking on each person with the probability of 50%, and mask at most 2 grids. We enforce the re-id feature of masked person in search branch to be consistent with unmasked re-id feature in instance search, where the loss of a prediction is written as

$$L_o(i) = 1 - s(f_i^o, f_{j(i)}^g), , i = 1, ..., n^o, \qquad (3)$$

where o represents the prediction of masked person, and n^o represents the number of predictions from occluded persons. The OIC module works based on SIC module. When they are performed simultaneously, OIC module also uses triplet loss between masked or unmasked instances.

Comparison with CGPS [11] and R-SiamNet [10]: Our proposed IDCL is related to CGPS and R-SiamNet, but has significant differences and contributions. (i) Our main contribution is a deep intra-image contrastive learning (DICL) framework. Our DICL provides two different contrast strategies, including spatial-invariant contrast (SIC) and occlusion-invariant contrast (OIC). Experimental results in Table VIII

Algorithm 1: Deep Into-image Contrastive learning Algorithm for weakly supervised person search

---

Input : Training images;
          Bounding-box annotations.
Output: Model of search branch.
for *each epoch* **do**
    a: Extract features of all training images with
    bounding-box annotations.
    b: Update memory bank and cluster.
    for *each mini-batch* **do**
        1: Randomly add mask patches on a portion of
        person regions of input image in search branch.
        2: Generate the re-id features of search branch and
        instance branch using Siamese network.
        3: Compute the contrast re-id loss by Eq. 4 and
        detection loss.
        4: Update the memory bank features.
    end
end

---

demonstrate that our DICL significantly surpasses CGPS and R-SiamNet. On PRW, the improvements are 19.3% and 14.3% in mAP, and the improvements are 12.9% and 7.5% in top-1 accuracy. We hope that our simple and effective DICL will serve as a strong baseline and help ease future research in this open-research area. (ii) The SIC in our DICL is a deep spatial-invariant contrast strategy. In contrast, CGPS and R-SiamNet are relatively shallow and do not fully exploit instance spatial information. For example, CGPS suffers from uncropped context interference when performing dense contrasts in search branch, while R-SiamNet ignores the impact of inaccurate bounding-box localization with one-to-one GT Siamese contrast. Experimental results in Table III demonstrate that even a simple combination of CGPS and R-SiamNet (*i.e.,* one-to-one GT Siamese contrast + dense contrasts in search branch) is inferior to our SIC (mAP: 31.4% vs 33.2%, top-1 accuracy: 77.3% vs 79.9%). (iii) Our SIC is a unified framework for spatial-invariant contrastive learning, where the contrast strategies in CGPS and R-SiamNet can be seen as a special case. In addition to SIC, we further introduce an occlusion-invariant contrast (OIC).

E. Training and Inference

During training, we build the initialized re-id memory bank by detecting all the training images and generating the re-id feature of each person with the averaged re-id feature of two branches. At the beginning of training, we perform the cluster on memory bank. At each iteration, the cluster-level memory bank is used for memory-based contrastive learning to supervise the re-id feature learning like [12]. In addition to memory-based contrast, we perform deep intra-image contrastive learning using the Siamese network. In search branch, we randomly add the mask on a portion of pedestrian region and perform the forward for detection and re-id feature generation. We keep the re-id features of all the n predicted bounding-boxes, where there are n^p re-id features without image masking and n^o re-id features with image masking. In instance branch, we crop and resize the ground-truth image regions into the fixed sizes and perform the

forward to generate the re-id features. With the re-id features in both branches, we perform spatial-invariant contrast and occlusion-invariant contrast. The contrast loss of our method can be written as

$$L_c = \frac{1}{n^p}\sum_i L_{mto}^i + \frac{1}{n^o}\sum_i L_o^i + \frac{1}{n}\sum_i L_{tri}^i + L_{oim}, \quad (4)$$

where L_{oim} is the OIM loss for memory-based contrast. Then, the overall loss can be written as $L_{all} = L_c + L_{det}$, where L_{det} is the detection loss. We summarize our DICL for weakly supervised person search in Algorithm 1.

During inference, we only keep the search branch in Siamese network for detection and re-id feature generation. We first employ search branch to perform detection and re-id feature generation on query image and a set of gallery images, respectively. After that, we set the re-id feature having the maximum overlap with the ground-truth bounding-box as re-id feature of each person. Finally, we calculate re-id feature similarities of query person and persons in gallery images, and select the matched persons according to the similarity scores.

IV. EXPERIMENTS

A. Datasets and Settings

CUHK-SYSU [5] is a large-scale person search dataset, including street and movie snapshots. It contains 18,184 images, 96,143 annotated pedestrian bounding-boxes, and 8,432 person identities. The training set contains 11,206 images, 55,272 pedestrian bounding-boxes, and 5,532 person identities, while the test set contains 6,978 images, 40,871 pedestrian bounding-boxes, and 2,900 person identities. During inference, the different queries have different gallery images, and the size of gallery images ranges from 50 to 4,000. If no special illustration, the gallery size is set to 100 as the default like most person search methods.

PRW [14] is captured in Tsinghua university using six cameras at different positions. Five cameras have a resolution of $1,080 \times 1,920$ pixels and another one is 576×720 pixels. There are 11,816 images, 43,110 pedestrian bounding-boxes, and 932 person identities. The training set contains 5,704 images, 18,048 pedestrian bounding-boxes and 482 person identities. The test set has 6,112 images, 25,062 pedestrian bounding-boxes, and 450 person identities.

Evaluation Protocol. We employ two standard evaluation metrics widely adopted in person search, including mean averaged precision (mAP) and top-1 accuracy.

B. Implementation Details

We implement our method on open-source library mmdetection [45]. We adopt ResNet50 pre-trained on the large-scale ImageNet-1K dataset [46], and use the fused feature map of last two layers for RPN and RoI head network. The fused feature map has the resolution of $H/16 \times W/16 \times 1024$ pixels. Search branch employs RoIAlign to generate the feature map of $14 \times 6 \times 1024$ pixels. Instance branch crops and resizes the pedestrian image into the fixed size of 224×96 pixels, so that the fused feature map in instance branch has the resolution of $14 \times 6 \times 1024$ pixels. The final fully-connected layer generates

TABLE I
IMPACT OF DIFFERENT MODULES IN OUR DICL, INCLUDING
SPATIAL-INVARIANT CONTRAST (SIC) AND OCCLUSION-INVARIANT
CONTRAST (OIC), FOR WEAKLY SUPERVISED PERSON SEARCH ON PRW
TEST SET.

Method (Weakly supervised)		re-id	
SIC	OIC	mAP	top-1
		25.6	73.2
✓		33.2	79.9
✓	✓	**35.5**	**80.9**

TABLE II
IMPACT OF DIFFERENT MODULES IN OUR DICL, INCLUDING
SPATIAL-INVARIANT CONTRAST (SIC) AND OCCLUSION-INVARIANT
CONTRAST (OIC), FOR FULLY SUPERVISED PERSON SEARCH ON PRW
TEST SET.

Method (Fully supervised)		re-id	
SIC	OIC	mAP	top-1
		43.6	78.8
✓		48.8	82.2
✓	✓	**49.6**	**82.4**

TABLE III
IMPACT OF INTEGRATING MANY-TO-ONE SIAMESE CONTRAST AND DENSE
CONTRAST IN OUR SPATIAL-INVARIANT CONTRAST (SIC) MODULE.

Num.	Many-to-one	Dense	mAP	top-1
(a)			25.6	73.2
(b)	✓		32.2	79.6
(c)		✓	31.4	77.3
(d)	✓	✓	**33.2**	**79.9**

TABLE IV
IMPACT OF DIFFERENT SIAMESE CONTRAST SETTINGS IN OUR
SPATIAL-INVARIANT CONTRAST (SIC) MODULE. GROUND-TRUTH
INDICATES GROUND-TRUTH BOUNDING-BOX OF A PERSON, WHILE
POSITIVES INDICATE ALL THE PREDICTIONS BELONGS TO A PERSON.

Num.	Search branch	Instance branch	mAP	top-1
(a)	Ground-truth	Ground-truth	25.6	73.2
(b)	Positives	Positives	31.2	79.7
(c)	Positives	Ground-truth	**33.2**	**79.9**

256-d vectors for re-id feature. We adopt the cluster hyper-parameters used in SPCL [35] to generate cluster-level re-id memory bank. In addition, we employ the prior that the persons in an image belong to different identities during cluster like [31], [11].

Our experiment is conducted on a single NVIDIA GeForce RTX 3090 GPU with SGD optimizer. The mini-batch is set as four images and the image scale is fixed to $1,500 \times 900$ pixels. There are 26 epochs in total, where the initial learning rate is set as 0.001 and decreased by a factor of 0.1 after epoch 16 and 22. During inference, the image is resized to $1,500 \times 900$ pixels.

C. Ablation Study

Here, we perform ablation study on PRW dataset to demonstrate the effectiveness of proposed method DICL.

Impact of different modules in DICL Table I shows the impact of our deep intra-image contrastive learning (DICL) for weakly supervised person search on PRW test set. The DICL contains a spatial-invariant contrast (SIC) module and an occlusion-invariant contrast (OIC) module. The baseline (a) only employs the ground-truth bound-box for intra-image contrast learning like [10]. When integrating our SIC module into the baseline, it provides 7.6% improvement on mAP and 6.7% improvement on top-1 accuracy. When integrating our OIC module into them, it further provides 2.3% improvement on mAP and 1.0% improvement on top-1 accuracy. Overall, our DICL provides 9.9% improvement on mAP and 7.7% improvement on top-1 accuracy. It demonstrates that our deep intra-image contrast learning is effective on weakly supervised person search.

In addition, we show the impact of our DICL for fully supervised person search in Table II. We build the re-id memory bank based on identity annotations, instead of unsupervised cluster. Our DICL achieves the mAP of 49.6% and the top-1 accuracy of 82.4%, which outperforms the baseline

by 6.0% and 3.6%. The related BINet [26] also adopts the Siamese framework for fully supervised person search, which achieves the mAP of 45.3% and the top-1 accuracy of 81.7%. Therefore, our DICL outperforms BINet by 4.3% on mAP and 0.7% on top-1 accuracy. It demonstrates the effectiveness of our deep intra-image contrast learning on fully supervised person search. Because this paper mainly focuses on weakly supervised settings, we do not add fully-supervised results in Table VIII. Meanwhile, our method has a larger improvement on weakly supervised person search, which demonstrates that intra-image contrast is more important for weakly supervised person search.

Impact of different contrasts in SIC We show the impact of integrating many-to-one Siamese contrast and dense prediction contrast in Table III. Many-to-one Siamese contrast denotes the contrasts between multiple predictions of a person in search branch and one ground-truth prediction of a person in instance branch. Dense prediction contrast denotes that the contrasts between all the predictions in search branch. When integrating single many-to-one Siamese contrast between two branches (b) or dense prediction contrast in search branch (c), it can significantly improve performance, respectively. When integrating these two contrasts together (d), it achieves the best performance. In addition, the method (c) can be treated as the combination of CGPS and R-SiamNet, which is inferior to our method (d). Therefore, our SIC methods outperform the simple combination of two weakly supervised methods CGPS and R-SiamNet.

Table IV shows the impact of different Siamese contrast settings. When we perform one-to-one Siamese contrast of ground-truth like [10], it (a) achieves 25.6% on mAP and 73.2% on top-1 accuracy. Further, we perform one-to-one Siamese contrasts of all positives, it (b) achieves 31.2% on mAP and 79.7% on top-1 accuracy. That provides a substantial improvement on the baseline with more diverse training samples. When we perform many-to-one Siamese contrasts between all positive samples and their corresponding ground-truth, it (c) achieves best performance. Compared to the strategy (b), the strategy (c) is more suitable to learn scale-

TABLE V
IMPACT OF MASKING STRATEGY ON DIFFERENT BRANCHES IN OUR
OCCLUSION-INVARIANT CONTRAST (OIC) MODULE.

Num.	Masking strategy	mAP	top-1
(a)	None	33.2	79.9
(b)	Search branch	**35.5**	**80.9**
(c)	Instance branch	33.1	79.5
(d)	Two branches	32.7	80.1

TABLE VI
IMPACT OF MASKING GRIDS IN OUR OCCLUSION-INVARIANT CONTRAST
(OIC) MODULE.

Num. of grids	0	2	4	6
mAP	33.2	35.5	35.4	33.0
Top-1 accuray	79.9	80.9	80.3	80.0

invariant and localization-invariant re-id feature.

Impact of masking strategy in OIC Table V shows the impact of the masking strategy on different branches of Siamese network. It achieves the best performance by performing the masking only on search branch. The masking strategies on instance branch or on both two branches do not improve the performance on mAP. We argue the reason is that instance branch aims to guide re-id feature to be consistent, while the masking strategy on instance branch will make it difficult to learn the consistent features. We also show the impact of masking grids in Table VI. We observe that masking 2 grids has best performance.

To show the ability of our OIC to deal with occlusion variance, we generate a masked test set by randomly masking 20% pixels of test images. Table VII compares the performance on original test test or masked test set. When using masked test set, all the methods have the performance drop. We observe that it has less performance drop when using our OIC module. It demonstrates that our occlusion-invariant contrast can better deal with occlusion variance.

D. Comparison With State-Of-The-Art Methods

We compare our proposed method DICL with the state-of-the-art methods, including fully supervised and weakly supervised approaches. They can be divided into two-step approaches and one-step approaches.

Comparison on CUHK-SYSU In Table VIII, we compare our DICL with some state-of-the-art approaches with the gallery size of 100, including fully-supervised and weakly-supervised methods. Our method DICL achieves the mAP of 87.4% and top-1 accuracy of 88.8%, which is superior to current weakly supervised one-step person search methods. CGPS [11] has the mAP of 80.0% and top-1 accuracy of 82.3%, while R-SiamNet [10] has the mAP of 86.0% and top-1 accuracy of 87.1%. Our proposed method outperforms CGPS by 7.4% on mAP and 6.5% on top-1 accuracy, and outperforms R-SiamNet by 1.4% on mAP and 1.7% on top-1 accuracy. In addition, our method outperforms some fully supervised one-step methods such as RCAA [52] and CTXG [53]. The weakly-supervised two-step CGUA [31] has the mAP of 91.0% and top-1 accuracy of 92.2%, which uses

TABLE VII
COMPARISON OF W/WO OCCLUSION-INVARIANT CONTRAST (OIC)
MODULE UNDER ORIGINAL AND MASKED PRW TEST SET.

Test set	without OIC		with OIC	
	mAP	top-1	mAP	top-1
Original	33.2	79.9	35.5	80.9
Masked	15.0	60.2	20.9	67.6
△ (drop)	18.2	19.7	14.6	13.3

TABLE VIII
COMPARISON WITH STATE-OF-THE-ART METHODS, INCLUDING FULLY
SUPERVISED AND WEAKLY SUPERVISED APPROACHES ON CUHK-SYSU
AND PRW TEST SETS.

	Methods	CUHK-SYSU		PRW	
		mAP	top-1	mAP	top-1
Two-step	*Fully supervised:*				
	IDE [14]	-	-	20.5	48.3
	MGTS [17]	83.0	83.7	32.6	72.1
	CLSA [47]	87.2	88.5	38.7	65.0
	RDLR [48]	93.0	94.2	42.9	70.2
	IGPN [49]	90.3	91.4	47.2	87.0
	TCTS [18]	93.9	95.1	46.8	87.5
	Weakly supervised:				
	CGUA [31]	91.0	92.2	42.7	86.9
One-step	*Fully supervised:*				
	OIM [5]	75.5	78.7	21.3	49.4
	IAN [50]	76.3	80.1	23.0	61.9
	NPSM [51]	77.9	81.2	24.2	53.1
	RCAA [52]	79.3	81.3	-	-
	CTXG [53]	84.1	86.5	33.4	73.6
	QEEPS [54]	88.9	89.1	37.1	76.7
	HOIM [13]	89.7	90.8	39.8	80.4
	BINet [26]	90.0	90.7	45.3	81.7
	NAE [22]	91.5	92.4	43.3	80.9
	PGA [55]	90.2	91.8	42.5	83.5
	AlignPS [12]	93.1	93.4	45.9	81.9
	DMRNet [24]	93.2	94.2	46.9	83.3
	AGWF [27]	93.3	94.2	53.3	87.7
	SeqNet [23]	93.8	94.6	46.7	83.4
	BUFF [56]	91.6	92.2	44.9	86.3
	OIMNet++[28]	93.1	94.1	47.7	84.8
	PSTR [9]	93.5	95.0	49.5	87.8
	COAT [30]	94.2	94.7	53.3	87.4
	Weakly supervised:				
	CGPS [11]	80.0	82.3	16.2	68.0
	R-SiamNet [10]	86.0	87.1	21.2	73.4
	DICL (Ours)	**87.4**	**88.8**	**35.5**	**80.9**

two independent networks for detection and re-id, and focuses on memory-based contrast. Compared to weakly-supervised two-step CGUA, our one-step DICL only utilizes a single network for detection and re-id, and focuses on deep intra-image contrast. Therefore, our DICL is a simple framework for end-to-end person search.

Fig 5 further presents the performance under different gallery sizes of [50, 100, 500, 1000, 2000, 4000]. The result shows that our method works better than weakly supervised one-step methods, especially under larger gallery sizes, which indicates that our method is capable of searching targets in

| (a) Query | (b) Gallery 1 | (c) Gallery 2 | (d) Query | (e) Gallery 1 | (f) Gallery 2 |

Fig. 3. Qualitative results of our method on PRW test set, where the red box represents the query and the green box represent search result in the gallery image. Our method finds the query persons with various views and scales.

Fig. 4. Qualitative results of our method on CUHK-SYSU test set, where the red box represents the query and the green box represent search result in the gallery image. Our method matches the query persons in different scenes.

large and challenging scenes.

Comparison on PRW We also compare our method with state-of-the-art approaches on PRW test set in Table VIII. Compared to CUHK-SYSU, PRW test set has a larger gallery size (6,112) and is challenging. Our method achieves a mAP of 35.5% and a top-1 accuracy of 80.9%, which outperforms CGPS by 19.3% on mAP and 12.9% on top-1 accuracy, and outperforms R-SiamNet by 14.3% on mAP and 7.5% on top-1 accuracy. In addition, our method outperforms fully supervised methods MGTS [52] and CTXG [53] by 2.9% and 2.1% on mAP respectively.

Qualitative results Fig. 3 and Fig. 4 give some qualitative results of our DICL on PRW and CUHK-SYSU test sets. Our method successfully detects and recognizes the query persons in different gallery images, where the persons have different poses, scales, background and visibility. For example, as shown in the last row of Fig. 3, the girl changes greatly in position, scale and visible part. We also observe that our method occasionally struggles at similar appearance, heavy occlusion, and nearby pedestrian interference in Fig. 6. Therefore, it is necessary to address these issues in future.

V. CONCLUSION

In this paper, we have proposed a novel deep intra-image contrastive learning (DICL) framework for weakly supervised one-step person search. We employed a Siamese network

Fig. 5. Comparison with different weakly supervised one-step methods under different gallery sizes on CUHK-SYSU test set.

with a search branch and an instance branch to learn discriminative features to deal with spatial-level and occlusion-level variances. Specifically, we introduced two novel intra-image contrast modules: spatial-invariant contrast (SIC) and occlusion-invariant contrast (OIC). SIC performs many-to-one Siamese contrasts between two branches of Siamese network and dense contrasts in all the predictions in search branch, which can learn discriminative scale-invariant and location-invariant re-id features. OIC preforms the masking strategy on

(a) Similar appearance   (b) Occlusion scenes   (c) Nearby pedestrians

Fig. 6. Some failure cases of our method. We observe that it is still challenging to detect and recognize the persons at heavy occlusion, extreme low-light conditions, etc.

search branch and improves feature consistency to occlusion variance. Extensive experiments on two datasets demonstrate the effectiveness of our method.

Currently, our method mainly focuses on deep intra-image contrast, which does not consider to improve memory-based contrast. In future, we will exploit how to improve the unsupervised cluster for memory-based contrast.

REFERENCES

[1] H. Yao and C. Xu, "Joint person objectness and repulsion for person search," *IEEE Transactions on Image Processing*, vol. 30, pp. 685–696, 2021.

[2] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search by separated modeling and a mask-guided two-stream cnn model," *IEEE Transactions on Image Processing*, vol. 29, pp. 5669–4682, 2020.

[3] C. Gao, G. Cai, X. Jiang, F. Zheng, J. Zhang, Y. Gong, F. Lin, X. Sun, and X. Bai, "Conditional feature learning based transformer for text-based person search," *IEEE Transactions on Image Processing*, vol. 31, pp. 6097–6108, 2022.

[4] Y. Chen, R. Huang, H. Chang, C. Tan, T. Xue, and B. Ma, "Cross-modal knowledge adaptation for language-based person search," *IEEE Transactions on Image Processing*, vol. 30, pp. 4057–4069, 2021.

[5] 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, pp. 3376–3385.

[6] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao, "From handcrafted to deep features for pedestrian detection: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4913–4934, 2022.

[7] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, 2022.

[8] C. Zhao, Z. Chen, S. Dou, Z. Qu, J. Yao, J. Wu, , and D. Miao, "Context-aware feature learning for noise robust person search," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 7047–7060, 2022.

[9] J. Cao, Y. Pang, R. M. Anwer, H. Cholakkal, J. Xie, M. Shah, and F. S. Khan, "Pstr: End-to-end one-step person search with transformers," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9458–9467.

[10] C. Han, K. Su, D. Yu, Z. Yuan, C. Gao, N. Sang, Y. Yang, and C. Wang, "Weakly supervised person search with region siamese networks," in *IEEE International Conference on Computer Vision*, 2021, pp. 11986–11995.

[11] Y. Yan, J. Li, S. Liao, J. Qin, B. Ni, X. Yang, and L. Shao, "Exploring visual context for weakly supervised person search," in *AAAI Conference on Artificial Intelligence*, 2022, pp. 3027–3035.

[12] Y. Yan, J. Li, J. Qin, S. Bai, S. Liao, L. Liu, F. Zhu, and L. Shao, "Anchor-free person search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7690–7699.

[13] D. Chen, S. Zhang, W. Ouyang, J. Yang, and B. Schiele, "Hierarchical online instance matching for person search," in *AAAI Conference on Artificial Intelligence*, 2020, pp. 10518–10525.

[14] 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, pp. 3346–3355.

[15] X. Ke, H. Liu, W. Guo, B. Chen, Y. Cai, and W. Chen, "Joint sample enhancement and instance-sensitive feature learning for efficient person search," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7924–7937, 2022.

[16] W. Yang, H. Huang, X. Chen, and K. Huang, "Bottom-up foreground-aware feature fusion for practical person search," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 262–274, 2022.

[17] 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, pp. 764–781.

[18] 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, pp. 11949–11958.

[19] C. Wang, B. Ma, H. Chang, S. Shan, and X. Chen, "Person search by a bi-directional task-consistent learning model," *IEEE Transactions on Multimedia*, 2021.

[20] S. Ren, K. He, R. B. 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, 2017.

[21] 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, pp. 2117–2125.

[22] 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, pp. 12612–12621.

[23] Z. Li and D. Miao, "Sequential end-to-end network for efficient person search," in *AAAI Conference on Artificial Intelligence*, 2021, pp. 2011–2019.

[24] 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 *AAAI Conference on Artificial Intelligence*, 2021, pp. 1505–1512.

[25] C. Han, Z. Zheng, K. Su, D. Yu, Z. Yuan, C. Gao, N. Sang, and Y. Yang, "Dmrnet++: Learning discriminative features with decoupled networks and enriched pairs for one-step person search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[26] 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, pp. 2836–2845.

[27] B.-J. Han, K. Ko, and J.-Y. Sim, "End-to-end trainable trident person search network using adaptive gradient propagation," in *IEEE International Conference on Computer Vision*, 2021, pp. 925–933.

[28] S. Lee, Y. Oh, D. Baek, J. Lee, and B. Ham, "Oimnet++: Prototypical normalization and localization-aware learning for person search," in *European Conference on Computer Vision*, 2022, pp. 621–637.

[29] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *IEEE International Conference on Computer Vision*, 2019, pp. 9626–9635.

[30] R. Yu, D. Du, R. LaLonde, D. Davila, C. Funk, A. Hoogs, and B. Clipp, "Cascade transformers for end-to-end person search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7267–7276.

[31] C. Jia, M. Luo, C. Yan, X. Chang, and Q. Zheng, "Cgua: Context-guided and unpaired-assisted weakly supervised person search," *arXiv:2203.14307*, 2022.

[32] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.

[33] M. S. Sarfraz, V. Sharma, and R. Stiefelhagen, "Efficient parameter-free clustering using first neighbor relations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8934–8943.

[34] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8738–8745.

[35] Y. Ge, F. Zhu, D. Chen, R. Zhao, and H. Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-id," in *The International Conference on Neural Information Processing Systems*, 2020, pp. 11309–11321.

[36] 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, pp. 994–1003.

[37] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7202–7211.

[38] J. Li, Y. Yan, G. Wang, F. Yu, Q. Jia, and S. Ding, "Domain adaptive person search," in *European Conference on Computer Vision*, 2022.

[39] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, 2020, pp. 1575–1585.

[40] J.-B. Grill, F. A. Florian Strub, C. Tallec, P. Richemond, E. Buchatskaya, B. A. P. Carl Doersch, Z. Guo, M. G. Azar, K. K. Bilal Piot, R. Munos, and M. Valko, "Bootstrap your own latent - a new approach to self-supervised learning," in *The International Conference on Neural Information Processing Systems*, 2020, pp. 21 271–21 284.

[41] X. Chen and K. He, "Exploring simple Siamese representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 745–53.

[42] X. Peng, K. Wang, Z. Zhu, M. Wang, and Y. You, "Crafting better contrastive views for siamese representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 031–16 040.

[43] 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, pp. 764–773.

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

[45] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, C. Z. Dazhi Cheng, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, YueWu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "Mmdetection: Open mmlab detection toolbox and benchmark," 2019.

[46] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[47] X. Lan, X. Zhu, and S. Gong, "Person search by multi-scale matching," in *European Conference on Computer Vision*, vol. 11205, 2018, pp. 553–569.

[48] 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, pp. 9813–9822.

[49] 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, pp. 2582–2591.

[50] 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.

[51] H. Liu, J. Feng, Z. Jie, J. Karlekar, B. Zhao, M. Qi, J. Jiang, and S. Yan, "Neural person search machines," in *IEEE International Conference on Computer Vision*, 2017, pp. 493–501.

[52] X. Chang, P. Huang, Y. Shen, X. Liang, Y. Yang, and A. G. Hauptmann, "RCAA: Relational context-aware agents for person search," in *European Conference on Computer Vision*, 2018, pp. 86–102.

[53] 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, pp. 2158–2167.

[54] 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, pp. 811–820.

[55] H. Kim, S. Joung, I.-J. Kim, and K. Sohn, "Prototype-guided saliency feature learning for person search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4865–4874.

[56] W. Yang, H. Huang, X. Chen, and K. Huang, "Bottom-up foreground-aware feature fusion for practical person search," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 1, pp. 262–274, 2022.