Doubly Contrastive Learning for Source-Free Domain Adaptive Person Search

Yizhen Jia*, Rong Quan*, Yue Feng, Haiyan Chen, Jie Qin†

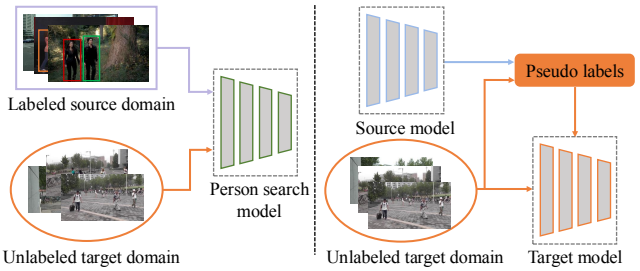
Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education
College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
yz.jia@nuaa.edu.cn

Abstract

Domain Adaptive Person Search (DAPS) aims to improve the generalization capability of person search models by training on both labeled source data and unlabeled target data, which is not that practical in real-world applications considering the storage/transmission costs and the privacy of source data. In this paper, we investigate a more practical and efficient person search setting, Source-Free Domain Adaptive Person Search (SFDA-PS), which seeks to generalize an existing source person search model to any unseen domain without requiring source data. Considering the absence of effective annotations in SFDA-PS, we propose a Doubly Contrastive Learning (DCL) method to adapt the target domain knowledge to the source model in a mutual learning and contrastive learning way. Specifically, we employ a mutual learning-based mean-teacher model as our baseline to incorporate target domain knowledge by pursuing prediction consistency between the teacher and student. Then, a Relation-embedded Contrastive (ReC) learning strategy is introduced to the detection head to ensure semantic consistency among proposals related to the same person while maintaining semantic distinction among proposals from different categories or persons. Furthermore, a Memory-aided Constrative (MaC) learning strategy is integrated into the re-identification (Re-ID) head to enhance its discriminative capability on target person embeddings. Extensive experiments on existing state-of-the-art person search models and two widely used benchmarks demonstrate the superiority of the proposed SFDA-PS task, as well as our proposed DCL.

Introduction

Person search (Xiao et al. 2017; Zheng et al. 2017) aims to identify a query person within realistic, uncropped scene images, which can be decomposed into two sub-tasks, pedestrian detection and person re-identification (Re-ID). Recently, person search has garnered significant interests in the computer vision community due to its extensive real-world applications. The primary driver of advancements in existing person search methods (Chen et al. 2020b; Li and Miao 2021; Zhang et al. 2021; Qin et al. 2023) is the availability of large-scale annotated benchmarks. Consequently, these models often exhibit poor generalization when applied to new domains not encountered during training due to the domain gap. However, collecting and annotating sufficient images to retrain or fine-tune models for each new domain is time-consuming and impractical.



(a) Previous DAPS task (b) The proposed SFDA-PS task

Figure 1: Comparison of the proposed Source-Free Domain Adaptive Person Search (SFDA-PS) setting with previous Domain Adaptive Person Search (DAPS).

Domain Adaptive Person Search (DAPS) (Li et al. 2022a) is proposed to enhance the generalization capability of person search models, enabling them to adapt to any target domain with no need for their labels. DAPS focuses on minimizing the domain gap between source and target domains by aligning their data distributions during training. Therefore, the training of DAPS typically requires the labeled source and unlabeled target data simultaneously, as shown in Fig. 1 (a). However, in practical applications, access to source data is often restricted due to concerns related to data privacy (particularly personal information), data transmission efficiency, and data proprietary. Furthermore, retraining a model from scratch for each new domain is inefficient.

To more efficiently enhance the person search model's generalization capability, we propose a novel Source-Free Domain Adaptive Person Search (SFDA-PS) setting in this paper, as illustrated in Fig. 1 (b). With no need for labeled source data, SFDA-PS directly generalizes the source-trained person search models to unseen target domains in an unsupervised way. Compared to DAPS, SFDA-PS protects the privacy of the source domain. Moreover, storing and transferring a source model is more efficient than handling large source datasets. As shown in Tab. 1, the storage

*These authors contributed equally.
†Corresponding author.

Storage size (MB)	CUHK-SYSU	PRW
Source Dataset	1211.5	2865.9
Source Model	398.6	560.5

Table 1: Storage size of source models and compressed datasets. The source models used here are SeqNet (Li and Miao 2021) and ROI-AlignPS (Yan et al. 2023), which are trained on CUHK-SYSU (Xiao et al. 2017) and PRW (Zheng et al. 2017), respectively.

size of a source model is considerably smaller than a compressed dataset. Furthermore, given the large number of existing well-trained person search models (Li and Miao 2021; Yan et al. 2021; Yu et al. 2022), SFDA-PS has more potential than DAPS for advancing the field of person search.

Nevertheless, SFDA-PS is more challenging than DAPS since no annotations are available when adapting the source model to the target domain. An intuitive solution to this problem is to train the target model using pseudo labels generated by the source model. However, due to the domain gap between the source and target data, these pseudo labels are often noisy, significantly hindering representation learning. To address this issue, we propose a Doubly Contrastive Learning (DCL) method for SFDA-PS in this paper to generalize existing source models to the target domain in a mutual learning and contrastive learning way, which can mitigate the lack of effective annotations in supervised learning. Specifically, we first employ a mean-teacher model (Tarvainen and Valpola 2017) as our baseline, which is widely utilized in domain adaptation (Deng et al. 2021; Zhang, Wang, and He 2023; Liu, Li, and Yuan 2023), to incorporate the target domain knowledge into the source model in a teacher-student mutual learning manner. The mean-teacher model generates a weak and a strong augmentation of the target image, inputs them into the teacher and student models, respectively, and then utilizes a consistency cost between the outputs of the teacher and student to gradually adapt the source model to the target domain.

Next, we develop two novel task-oriented contrastive learning strategies to further integrate the target domain knowledge into the source model. As the detection performance profoundly affects the final person search results, a Relation-embedded Contrastive (ReC) learning strategy is introduced after the detection head. ReC updates the student model by maximizing semantic consistency among proposals related to the same person and ensuring semantic distinction among proposals related to different categories (people and background) or persons. Leveraging the proposals generated by the Region Proposal Network (RPN) in the Faster RCNN-based person search models (*e.g.*, NAE (Chen et al. 2020b) and SeqNet (Li and Miao 2021)), multiple views for any person are inherently provided, satisfying the requirements for contrastive learning (Chen et al. 2020c). Furthermore, a Memory-aided Contrastive (MaC) learning strategy is proposed to enhance the student's discriminative capability regarding person embeddings. Concretely, we begin by clustering person embeddings generated by the source model and use the cluster centers, outliers, and hard proposal embeddings to initialize a hybrid embedding memory bank. The student model is then updated by maximizing the similarity between person embeddings of the same identity (ID) and the differences between person embeddings of different IDs. The hybrid embedding memory bank is updated along with the teacher model after each training epoch.

In summary, the main contributions of this paper include:

- We propose a new SFDA-PS task, which is more efficient and practical in real-world applications. Compared to DAPS, SFDA-PS effectively mitigates data concerns, as well as avoids retraining the model from scratch whenever a new domain is encountered.
- We present the first SFDA-PS model, named DCL, which directly generalizes existing person search models to new domains. DCL effectively addresses the challenge of lack of annotations in SFDA-PS. It employs a mean-teacher framework as the baseline to transfer the target domain knowledge to the source model in a mutual learning way. Moreover, two task-oriented contrastive learning strategies, ReC and MaC, are introduced to further import the knowledge from the detection and Re-ID heads.
- Extensive experimental results demonstrate the effectiveness of DCL in generalizing state-of-the-art person search models. Remarkably, our method surpasses existing DAPS methods without utilizing any source data, highlighting the superior generalization capability of our approach.

Related Work

Person Search

Person search aims to jointly tackle pedestrian detection and person Re-ID tasks. Supervised person search methods have achieved impressive results and can be broadly categorized into two types, *i.e.*, two-step models and one-step models. The two-step approach (Zheng et al. 2017; Han et al. 2019; Wang et al. 2020), which utilizes separate networks for each sub-task, yields high performance but at the cost of increased computational and storage demands. In contrast, one-step models (Xiao et al. 2017; Yan et al. 2023), which solve both sub-tasks in an end-to-end manner, have become the mainstream approach. Inspired by advancements in object detection (Ren et al. 2015; Lin et al. 2017; Tian et al. 2019), numerous one-step methods (Yan et al. 2019; Chen et al. 2020b; Li and Miao 2021; Han et al. 2021; Yan et al. 2021; Qin et al. 2023) have been developed to make the joint framework more efficient and effective. However, these models often struggle to generalize well to new test domains.

DAPS (Li et al. 2022a; Cui et al. 2024; Almansoori et al. 2024) focuses on generalizing the model from a source domain to a new target domain and has shown significant progress. Unlike DAPS, which requires access to both source and target data, our work investigates a novel and more practical domain adaptation setting that adapts the pretrained source person search model to an unlabeled target domain without requiring access to labeled source data.

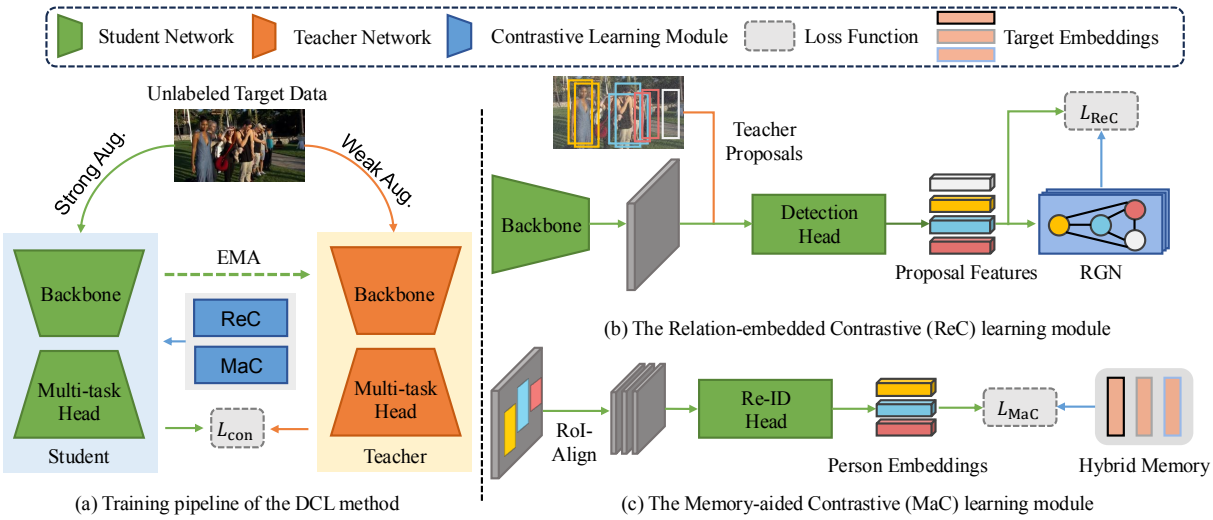


Figure 2: The overall architecture of the proposed DCL method. The mean-teacher framework is used as the baseline to transfer target knowledge through mutual learning. The teacher and student models use an identical person search structure, including a backbone for feature extraction and a multi-task head for detection and Re-ID. The teacher and student take a weak and a strong augmentation of the unlabeled target image as input, and the consistency loss (L_{con}) between their outputs is used to optimize the student model. The teacher is updated via the Exponential Moving Average (EMA) of student weights. Two task-oriented contrastive learning strategies, ReC and MaC, are used to further update the student.

Source-Free Domain Adaptation

In real-world applications, source data is often inaccessible during the adaptation process due to proprietary concerns, privacy regulations, or data transmission constraints. Many approaches have been proposed to address SFDA task for image classification (Liang, Hu, and Feng 2020; Zhang, Wang, and He 2023), object detection (Li et al. 2022b; VS, Oza, and Patel 2023), and segmentation (Lo et al. 2023). In classification, two main paradigms have emerged, *i.e.*, sample generation (Li et al. 2020; Hou and Zheng 2021) and pseudo label based self-training (Liang, Hu, and Feng 2020; Zhang, Wang, and He 2023). For source-free object detection, the pseudo label strategy is predominantly employed. SED (Li et al. 2021) introduces a self-training strategy by searching for an appropriate confidence threshold. UMA (Huang et al. 2021) and RCML (Lin et al. 2023) propose self-supervised feature learning with the mean-teacher framework (Tarvainen and Valpola 2017).

The SFDA task also has been explored for person Re-ID. Existing methods (Fu et al. 2019; Yang et al. 2020; Chen et al. 2023; Qu et al. 2024; Liu, Ye, and Du 2024) typically utilize reliable pseudo identity labels for self-supervised learning. However, these methods are specifically designed for Re-ID and cannot be directly applied to person search, which involves the additional task of person detection.

Methodology

Framework Overview

For a vanilla DAPS (Li et al. 2022a) task, we are given N_s labeled samples $\{x_s^n, y_s^n\}$ from the source domain $\mathcal{D}_{\mathcal{S}}$, where x_s^n denotes the n-th source image and y_s^n denotes the corresponding ground-truth labels, and N_t unlabeled samples $\{x_t^n\}$ from the target domain $\mathcal{D}_{\mathcal{T}}$, where x_t^n denotes the n-th target image without the ground-truth annotations. In contrast, the novel SFDA-PS considers a more realistic applicable scenario where only a source model $f(\theta)$ trained on the source domain $\mathcal{D}_{\mathcal{S}}$, and the unlabeled target domain $\mathcal{D}_{\mathcal{T}}$ are available during adaptation.

The overall architecture of the proposed DCL method is illustrated in Fig. 2. DCL employs a mean-teacher framework (Tarvainen and Valpola 2017) as the baseline to transfer target domain knowledge into the source model in a teacher-student mutual learning way. Furthermore, DCL develops two task-oriented contrastive learning strategies, ReC and MaC, to mitigate the bias towards the source data.

The Mean-Teacher Framework

Inspired by the recent success of the mean-teacher method in cross-domain adaptation object detection (Deng et al. 2021; VS, Oza, and Patel 2023), we formulate our DCL model within a mean-teacher framework. This framework consists of two models with identical architectures, a student model $f(\theta_{st})$ parameterized by Θ_{st} and a teacher model $f(\theta_{te})$ parameterized by Θ_{te}. The mean-teacher aims to evolve both models through mutual learning. As stated in (Cai et al. 2019) and (Deng et al. 2021), a key factor in the substantial improvement achieved by the mean-teacher method is the use of weak augmentations for the teacher and strong augmentations for the student, maintaining consistency between their predictions. Each sample prediction of the teacher can be seen as an ensemble of the student's current and earlier versions, indicating that the teacher is more robust and stable. Therefore, we directly use the teacher model in the inference stage.

At the beginning of DCL, both the teacher model $f(\theta_{te})$ and student model $f(\theta_{st})$ are initialized by the pre-trained source model $f(\theta)$. During the training stage, the student is optimized and the teacher is updated gradually by transferring the weights of the continuously learned student model. Given an unlabeled target image x_t, we first generate its strong augmentation $\hat{x}_t$ and weak augmentation $\bar{x}_t$. The teacher takes $\bar{x}_t$ as input to generate person bounding boxes, where the reliable ones with prediction confidence higher than a confidence threshold ε_h are preserved as pseudo bounding boxes. We generate each pseudo bounding box a pseudo identity label by clustering. The pseudo bounding boxes and identity labels constitute the pseudo labels $\hat{y}_t$ for the target training data. Then, these pseudo labels are used to update the student model, where the pseudo label supervision loss is formulated as follows:

$$L_{\text{base}} = L^{\text{rpn}}(\hat{x}_t, \hat{y}_t) + L^{\text{det}}(\hat{x}_t, \hat{y}_t) + L^{\text{re-id}}(\hat{x}_t, \hat{y}_t), \quad (1)$$

where L_{rpn} represents the loss of RPN, and L^{det} and $L^{\text{re-id}}$ denote the losses of detection and Re-ID heads, respectively.

In addition, we denote the student and the teacher detection class logits as p_{st} and p_{te}. To maintain consistency between the outputs of the teacher and student, we further minimize the discrepancy between p_{st} and p_{te}, which is typically computed using the KL-Divergence:

$$L_{\text{con}} = \text{KL}(\sigma(p_{st}), \sigma(p_{te})), \quad (2)$$

where σ denotes the softmax operator.

To obtain more stable pseudo labels, the teacher is gradually updated via the Exponential Moving Average (EMA) of student weights Θ_{st}:

$$\Theta_{te} \leftarrow \alpha\Theta_{te} + (1-\alpha)\Theta_{st}, \quad (3)$$

where α is the EMA rate that controls the update rate of teacher weights.

Although the mean-teacher framework enables the transfer of target knowledge into the source model, it is still not sufficient to eliminate the bias in feature space since it does not explicitly address the domain shift. Therefore, we further introduce ReC and MaC to facilitate model adaptation.

Relation-embedded Contrastive Learning

Noise in pseudo bounding boxes is unavoidable, even if they are selected with a high confidence threshold. Such noise can lead to error accumulation and performance damage. Therefore, we propose ReC, as illustrated in Fig. 2 (b), to eliminate the noise by modifying the bias of feature representations, which is done by ensuring feature consistency among proposals related to the same person, as well as feature distinction among proposals related to different categories or different persons.

We construct a relation graph to model the relationships among proposals. Specifically, as shown in Fig. 3, known the features of the proposals output by the detection head, we first learn the optimal relationships among them via two linear layers and a normalization layer. Then, we construct a relation graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where each vertex $v_i \in \mathcal{V}$ represents the feature of a proposal, each edge $e_{i,j} \in \mathcal{E}$ represents

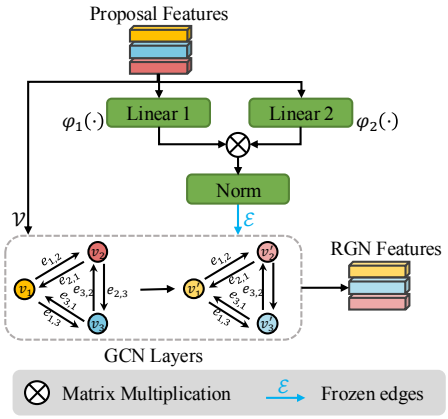


Figure 3: Details of the Relation Graph Network (RGN). RGN comprises learnable linear layers and Graph Convolution Network (GCN) layers. The inputs are the proposal features extracted from the detection head.

the optimal relationship between the i-th and j-th proposals. Note that we use the proposals of the teacher to extract proposal features and construct a relation graph for both teacher and student networks. The optimal relation $e_{i,j}$ can be represented as:

$$e_{i,j} = \frac{\exp(\phi(\varphi_1(v_i), \varphi_2(v_j)))}{\sum_k^{N_p} \exp(\phi(\varphi_1(v_i), \varphi_2(v_k)))}, \quad (4)$$

where $\phi(a,b) = a \cdot b^T$ is the inner product, φ_1 and φ_2 represent linear layers, and N_p is the number of RPN proposals.

Before ReC, a Graph Relation Network (GCN) is introduced to explore the optimal node relationships and preliminarily adjust the node features. We input $\mathcal{G}$ into the GCN, freeze the edges, and update the feature representations of the proposals according to the optimal relations. Subsequently, the output features of the RGN are fed into the classification layer of the detection sub-network to predict each proposal's class logits. We use $\tilde{p}_{st}$ and $\tilde{p}_{te}$ to represent the student and the teacher class logits predicted based on the RGN output features, respectively. Then, we minimize the discrepancy L_{RGN} between class logits p and $\tilde{p}$ to supervise the RGN parameters, which can be formulated as:

$$L_{\text{RGN}} = \text{KL}(\sigma(p_{st}), (\sigma(\tilde{p}_{st})) + \text{KL}(\sigma(p_{te}), (\sigma(\tilde{p}_{te})). \quad (5)$$

When the relationship matrix $(\mathcal{E})$ is obtained, we can generate the pair-wise relation labels and further construct positive/negative pairs. For any i-th and j-th proposal pair, the pair-wise relation label $M_{i,j}$ is generated by simply setting a threshold ϵ on normalized $\mathcal{E}$ which is typically formulated as:

$$M_{i,j} = \begin{cases} 1, & \text{if } e_{i,j} \geq \epsilon \\ 0, & \text{if } e_{i,j} < \epsilon. \end{cases} \quad (6)$$

$M_{i,j} = 1$ indicates the i-th and j-th proposals are highly related and form a positive pair, and vice versa for a negative pair.

Considering any i-th proposal as an anchor, we can define a positive set $T(i) = \{t \mid t \neq i, M_{i,t} = 1\}$. Each feature v_i

is projected as key $k_i = \psi_1(v_i)$ and query $q_i = \psi_2(v_i)$ in order to model better relations among the proposal features (Vaswani et al. 2017), where ψ_1 and ψ_2 are another linear learnable functions. The ReC loss L_{ReC} can be formulated as:

$$L_{\mathrm{ReC}} = \sum_{i}^{N_p} -\log \left\{ \frac{1}{\mid T \mid} \sum_{t}^{|T|} \frac{\exp\left(\phi\left(k_i, q_t\right)\right)}{\sum_{k,k\neq i}^{N_p} \exp\left(\phi\left(k_i, q_k\right)\right)} \right\}, \quad (7)$$

where $\mid T \mid$ is the number of $T(i)$. Note that ReC is used only to update the student parameters, whereas the teacher parameters are updated via EMA.

Memory-aided Contrastive Learning

The aim of the Re-ID sub-task is to improve the discrimination of person embeddings. Generally, previous supervised methods (Xiao et al. 2017; Yan et al. 2023) often achieve this goal by minimizing the feature discrepancy among the embeddings of the same person while maximizing the discrepancy among those of different people. However, in SFDA-PS, we have no ground-truth labels for the target images and can only use the pseudo labels generated from the teacher model, which inevitably contain noise. To this end, we propose MaC to make use of detection results and make person embeddings more discriminative, as illustrated in Fig. 2 (c).

MaC comprises two alternate processes, *i.e.*, constructing a memory bank and calculating the MaC loss to update the student. At the start of each epoch, we first generate pseudo bounding boxes and their corresponding embeddings from the teacher model. Then, the embeddings of the pseudo bounding boxes are clustered, and a hybrid embedding memory $\mathcal{B} = \{\mathcal{W}, \mathcal{O}, \mathcal{H}\}$ is constructed, which consists of three embedding types, cluster centroids $\mathcal{W}$, unclustered embeddings $\mathcal{O}$, and hard proposal embeddings $\mathcal{H}$. Specifically, we employ the clustering strategy (*e.g.*, DBSCAN) and the self-paced strategy (Ge et al. 2020) to obtain N_t^w clusters with centroids $\mathcal{W} = \left\{w_1, w_2, \dots w_{N_t^w}\right\}$, and N_t^o outlier embeddings $\mathcal{O} = \left\{o_1, o_2, \dots o_{N_t^o}\right\}$ not belonging to any cluster. The self-paced strategy is utilized to gradually create reliable clusters with the reliable criterion of measuring cluster independence and compactness. Hence, reliable pseudo identity labels can be obtained from the clustering results.

Inspired by (Li et al. 2022a), we also explore the potential of hard proposals to exploit target domain information sufficiently. Proposals with confidence in the range of $(\varepsilon_l, \varepsilon_h)$ are regarded as hard cases, where $\varepsilon_l < \varepsilon_h$, meaning the lower and higher bound thresholds. Hard proposals can be divided into three types: highly overlapped with high-confidence results, undetected persons, and background clusters. We exclude highly overlapped duplicates by further screening IoUs with high-confidence results, while the undetected persons and clusters belonging to the background are reserved for training to enhance the discrimination of the Re-ID branch. Then, the embeddings of these hard cases are added to $\mathcal{B}$.

Eventually, we build a hybrid memory $\mathcal{B} = \{\mathcal{W}, \mathcal{O}, \mathcal{H}\}$ to aid the Re-ID training. The MaC loss can be expressed as:

$$\begin{aligned} L_{\mathrm{MaC}} &= -\log \frac{\exp\left(\phi\left(x, b^+\right)\right)}{\sum_{b\in\mathcal{B}} \exp\left(\phi\left(x, b\right)\right)}, \\ \sum_{b\in\mathcal{B}} \exp\left(\phi\left(x, b\right)\right) &= \sum_{k=1}^{N_t^w} \exp\left(\phi\left(x, w_k\right)\right) + \\ &\sum_{k=1}^{N_t^o} \exp\left(\phi\left(x, o_k\right)\right) + \sum_{k=1}^{N_t^h} \exp\left(\phi\left(x, h_k\right)\right), \end{aligned} \quad (8)$$

where b_+ is the corresponding class prototype of the input embedding x. N_t^h is the number of the hard proposals in the memory, and h_k denotes the k-th embedding of the hard proposals.

Note that the hybrid memory is constructed before the start of each epoch, and the embeddings in the memory are not updated during the epoch.

Overall Loss Function

So far, we have introduced the mean-teacher framework, ReC, and MaC to tackle the source-free domain adaptation problem for person search effectively. The overall loss function of our proposed DCL method is formulated as follows:

$$L_{\mathrm{DCL}} = L_{\mathrm{base}} + L_{\mathrm{con}} + L_{\mathrm{RGN}} + L_{\mathrm{ReC}} + L_{\mathrm{MaC}}. \quad (9)$$

Experiments

Datasets and Settings

Datasets We conduct experiments on two general person search benchmarks, CUHK-SYSU (Xiao et al. 2017) and PRW (Zheng et al. 2017). CUHK-SYSU is a large-scale person search dataset that contains 18,184 scene images with 8,432 different identities and 96,143 annotated bounding boxes. The images come from two kinds of data sources (*i.e.*, street snaps and movies), covering diverse scenes under various viewpoints, lighting, resolutions, and occlusions. We utilize the standard training/test set, where the training set contains 5,532 identities and 11,206 images, and the test set contains 2,900 query persons and 6,978 images. PRW is extracted from video frames recorded by six static cameras on a university campus and contains 932 labeled persons with 43,110 bounding boxes. The dataset is split into a training set of 5,704 images with 482 different identities and a test set of 6112 images with 2,057 query persons.

Evaluation Protocols Following the settings in previous work (Chen et al. 2020b), the mean Average Precision (mAP) and Top-1 matching rate are adopted to evaluate the performance of person search. We also employ the Recall rate and Average Precision (AP) to measure the detection performance. For all the above evaluation metrics, the higher the value, the better the performance.

Generalization Setup To verify the generalization capability of DCL, we conduct experiments using three existing state-of-the-art person search models. Specifically, We take NAE (Chen et al. 2020b), SeqNet (Li and Miao 2021), and GLCNet (Qin et al. 2023) as the base person search networks for our DCL approach, respectively, and generalize these

Model	Type	CUHK-SYSU				PRW			
		Recall	AP	mAP	Top-1	Recall	AP	mAP	Top-1
NAE	S	41.3	37.8	39.4	41.0	92.9	86.8	31.3	79.4
	DCL	78.7	69.6	76.4	78.0	97.3	88.0	38.2	82.4
	O	90.4	86.3	93.2	93.6	93.5	89.5	40.5	79.6
SeqNet	S	58.5	52.8	54.8	56.8	92.6	87.6	29.8	77.5
	DCL	81.4	75.4	79.6	81.2	97.1	91.1	38.4	82.5
	O	91.4	88.5	93.9	94.5	96.7	94.2	46.3	82.9
GLCNet	S	63.0	56.7	61.4	64.0	94.5	89.5	26.0	80.3
	DCL	84.0	77.0	83.5	85.8	97.9	91.8	30.0	80.9
	O	91.4	88.2	95.5	96.5	97.0	94.2	49.5	87.5

Table 2: Performance of employing DCL method based on NAE (Chen et al. 2020b), SeqNet (Li and Miao 2021), and GLCNet (Qin et al. 2023) models on CUHK-SYSU and PRW. S: source only. O: oracle.

three models on the two tasks. The first task involves adapting a source model pre-trained on PRW to CUHK-SYSU, and the second is reversed. For brevity, only the target domain is specified when presenting the results.

Implementation Details

We implement our model with the PyTorch library and conduct all experiments on a single NVIDIA RTX A5000 GPU. In all of our experiments, the batch size is set to 1, and we adopt the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 0.0005. We optimize the model for 10 epochs, using an initial learning rate of 0.001, which is decreased by a factor of 10 at epoch 8. As for the data augmentation, we apply the random horizontal flip for weak augmentation and randomly add color jittering, grayscale, Gaussian blur, and cutout patches for strong augmentations, following UMT (Deng et al. 2021). N_p is set to 300 to construct the relation graph. ϵ is set to 0.5 to generate the pair-wise relation labels in ReC. We set default hyper-parameters $\varepsilon_h = 0.9$, $\varepsilon_l = 0.8$, and $\alpha = 0.8$.

Quantitative Results

Comparison on Different Person Search Models To assess the effectiveness of DCL, we conduct quantitative experiments on three Faster RCNN-based person search methods, respectively. The evaluation results are presented in Tab. 2. Source only refers to testing the pre-trained source model directly on the target domain. Oracle represents training a person search model on the target domain with ground-truth annotations with our own implementations, serving as a reference for the upper bound of adaptation performance.

Tab. 2 shows that the generalization ability of the source model is significantly improved with our DCL method. For instance, when employing DCL on CUHK-SYSU, the metrics of detection and Re-ID improve by over 30%, 20%, and 20% on NAE, SeqNet, and GLCNet, respectively, compared to the source only results. Since CUHK-SYSU contains more training data and diverse scenarios than PRW, the source model pre-trained on CUHK-SYSU exhibits better generalization capability than the one pre-trained on PRW, as evidenced by the comparison between the source only and oracle performance. Consequently, the improvements achieved by DCL on CUHK-SYSU are more pronounced than those on PRW. The significant improvements on CUHK-SYSU further demonstrate that our proposed DCL method can enhance generalization from simple to complex scenarios.

Method	CUHK-SYSU		PRW	
	mAP	Top-1	mAP	Top-1
OIM (Xiao et al. 2017)	75.5	78.7	21.3	49.9
CTXGraph (Yan et al. 2019)	84.1	86.5	33.4	73.6
HOIM (Chen et al. 2020a)	89.7	90.8	39.8	80.4
NAE (Chen et al. 2020b)	91.5	92.4	43.3	80.9
SeqNet (Li and Miao 2021)	93.8	94.6	46.7	83.4
AlignPS (Yan et al. 2021)	93.1	93.4	45.9	81.9
PSTR (Li et al. 2022a)	93.5	95.0	49.5	**87.8**
COAT (Yu et al. 2022)	94.2	95.5	51.0	86.8
GLCNet (Qin et al. 2023)	**95.7**	**96.3**	46.9	85.1
DPM (Zheng et al. 2017)	-	-	20.5	48.3
MGTS (Chen et al. 2018)	83.0	83.7	32.6	72.1
RDLR (Han et al. 2019)	93.0	94.2	42.9	70.2
OR (Yao and Xu 2020)	92.3	93.8	**52.3**	71.5
DAPS (Li et al. 2022a)	77.6	79.6	34.7	80.6
FOUS (Cui et al. 2024)	78.7	80.4	35.4	80.8
DDAM (Almansoori et al. 2024)	79.5	81.3	36.7	81.2
DCL (NAE)	*76.4*	*78.0*	*38.2*	*82.4*
DCL (SeqNet)	*79.6*	*81.2*	*38.4*	*82.5*
DCL (GLCNet)	*83.5*	*85.8*	*30.0*	*80.9*

Table 3: Comparison of mAP and Top-1 accuracy with the general state-of-the-art supervised and domain adaptive person search methods on CUHK-SYSU and PRW. The one-step and two-step supervised methods are grouped into the first and second categories, respectively. Our models are shown in *italics*. The best results are marked in **bold**.

Comparison to State-of-the-Art Methods As the proposed method of source-free domain adaptive person search is introduced for the first time in this paper, there is no method fairly comparable to our approach. Therefore, we compare the performance of DCL on NAE, SeqNet, and GLCNet with existing state-of-the-art person search methods under different settings, including supervised and domain adaptation methods, as shown in Tab. 3.

Despite the inevitable performance degradation due to the domain gap, our method still outperforms several supervised methods. For instance, DCL achieves higher mAP accuracy than OIM (Xiao et al. 2017), CTXGraph (Yan et al. 2019), DPM (Zheng et al. 2017), and MGTS (Chen et al. 2018) on PRW. Additionally, it is noteworthy that DCL achieves better performance than previous DAPS methods on both datasets, despite that DCL can only access the source model and unlabeled target data. This result suggests that DCL exhibits superior generalization capacity.

Ablation Study

Component Analysis We conduct an in-depth ablation analysis to investigate the impact of each component. Tab. 4

Mean-teacher	ReC	MaC	CUHK-SYSU			
			Recall	AP	mAP	Top-1
Source only			58.5	52.8	54.8	56.8
✓			73.4	67.4	65.3	66.8
✓	✓		78.0	71.8	65.8	67.5
✓		✓	<u>80.1</u>	<u>74.2</u>	<u>79.2</u>	<u>80.3</u>
✓	✓	✓	**81.4**	**75.4**	**79.6**	**81.2**

Table 4: Ablation study of the mean-teacher, ReC, and MaC on CUHK-SYSU. The best results in each metric are marked in **bold**, while the second results are marked with <u>underline</u>.

Strategy	CUHK-SYSU			
	Recall	AP	mAP	Top-1
w/o RGN	79.3	73.6	77.9	79.2
w/ RGN	81.4	75.4	79.6	81.2

Table 5: Comparative results when employing different strategies to model relationships among proposals.

presents the ablation results of the mean-teacher framework, ReC, and MaC on CUHK-SYSU. Additionally, we perform the second and third ablation experiments with the Re-ID sub-network frozen. The results show that using the mean-teacher framework alone enhances both detection and Re-ID performance, attributable to the mutual learning between the teacher and student models. As seen in Tab. 4, each contrastive module significantly improves detection and/or Re-ID performance. Notably, in the third row, we observe a substantial improvement in detection and a minor improvement in Re-ID, even though the Re-ID head is frozen at this moment. This suggests that ReC significantly boosts detection accuracy, thereby enhancing overall person search performance. Furthermore, we find that the two task-oriented contrastive learning strategies are complementary, yielding the best results when combining them with the mean-teacher framework.

Effectiveness of RGN To verify the effectiveness of RGN in learning the optimal relationships and enhancing representations, we conduct an analysis experiment by removing RGN and performing contrastive learning directly on the detection head output features. Without RGN, cosine similarity is used to measure relationships among the proposals and pair-wise relation labels are generated with the threshold ϵ. As shown in Tab. 5, employing RGN achieves better performance on CUHK-SYSU. This indicates that RGN effectively preserves more target-specific semantic information for modeling relationships and mitigates the impact of the domain gap.

Effectiveness of Hybrid Embedding Memory Tab. 6 reports the generalization performance using different types of embeddings within the hybrid embedding memory. "Cluster" refers to embedding clustered with DBSCAN and self-paced strategy (Ge et al. 2020). "Hard" represents embeddings obtained through hard proposal mining. The addition of hard case embeddings results in the best performance,

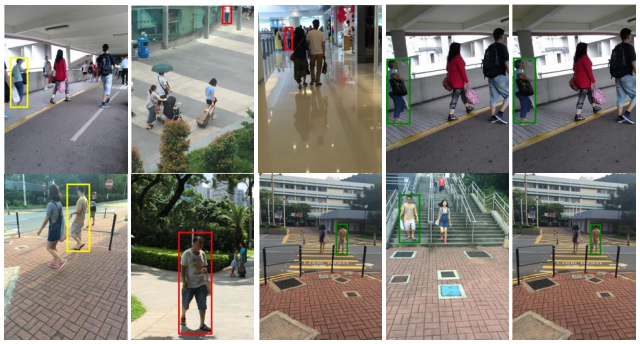


Figure 4: Visualization of DAPS (Li et al. 2022a) and three generalized models with DCL for challenging cases. The yellow bounding boxes denote the query, while the green and red boxes denote correct and incorrect Top-1 matches, respectively. * indicates the model adapted using DCL.

Cluster	Hard	CUHK-SYSU		PRW	
		mAP	Top-1	mAP	Top-1
		65.3	66.8	33.6	78.5
✓		77.9	78.8	37.1	81.3
✓	✓	79.2	80.3	38.2	82.0

Table 6: Ablation study of embedding clustering and hard proposal mining in hybrid memory.

indicating that these hard proposals effectively enhance the discriminative capability of person instances.

Qualitative Results

Fig. 4 presents several qualitative results on CUHK-SYSU, with query images captured by hand-held cameras. The Top-1 results of DAPS and the three models adapted using DCL are visualized for challenging cases, including obstacle occlusion (first row) and scale/viewpoint variation (second row). Our adapted models successfully localize and match the query person in difficult samples where DAPS fails, demonstrating the effectiveness of our DCL approach.

Conclusion

In this paper, we introduce a new person search task called SFDA-PS. SFDA-PS enables the generalization of an existing source model to unseen domains without requiring any annotated data, making it more efficient and less resource-intensive for real-world applications. To address SFDA-PS, we propose a novel DCL method, which employs a mean-teacher framework to integrate the target domain knowledge through mutual learning. Additionally, two task-oriented contrastive learning strategies are introduced, applied after the detection and Re-ID heads, to further incorporate the target domain knowledge into the source model. Extensive experiments on three advanced person search models and two large-scale benchmarks demonstrate the effectiveness of the new SFDA-PS task and the proposed DCL method.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 62276129 & No. 62206127), and the Fundamental Research Funds for the Central Universities (No. NS2024060).

References

Almansoori, M. K.; Fiaz, M.; Cholakkal, H.; and Yiz, J. 2024. DDAM-PS: Diligent Domain Adaptive Mixer for Person Search. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6688–6697.

Cai, Q.; Pan, Y.; Ngo, C.-W.; Tian, X.; Duan, L.; and Yao, T. 2019. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11457–11466.

Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; and Schiele, B. 2020a. Hierarchical online instance matching for person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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

Chen, D.; Zhang, S.; Yang, J.; and Schiele, B. 2020b. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chen, F.; Wang, N.; Tang, J.; Yan, P.; and Yu, J. 2023. Unsupervised person re-identification via multi-domain joint learning. *Pattern Recognition*, 138: 109369.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020c. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607. PMLR.

Cui, T.; Wang, H.; Peng, J.; Deng, R.; Fu, X.; and Wang, Y. 2024. Fast One-Stage Unsupervised Domain Adaptive Person Search. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 713–721.

Deng, J.; Li, W.; Chen, Y.; and Duan, L. 2021. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4091–4101.

Fu, Y.; Wei, Y.; Wang, G.; Zhou, Y.; Shi, H.; and Huang, T. S. 2019. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, 6112–6121.

Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; et al. 2020. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems*, 33: 11309–11321.

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

Han, C.; Zheng, Z.; Gao, C.; Sang, N.; and Yang, Y. 2021. Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Hou, Y.; and Zheng, L. 2021. Visualizing adapted knowledge in domain transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13824–13833.

Huang, J.; Guan, D.; Xiao, A.; and Lu, S. 2021. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34: 3635–3649.

Li, J.; Yan, Y.; Wang, G.; Yu, F.; Jia, Q.; and Ding, S. 2022a. Domain adaptive person search. In *European Conference on Computer Vision*, 302–318. Springer.

Li, R.; Jiao, Q.; Cao, W.; Wong, H.-S.; and Wu, S. 2020. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9641–9650.

Li, S.; Ye, M.; Zhu, X.; Zhou, L.; and Xiong, L. 2022b. Source-free object detection by learning to overlook domain style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8014–8023.

Li, X.; Chen, W.; Xie, D.; Yang, S.; Yuan, P.; Pu, S.; and Zhuang, Y. 2021. A free lunch for unsupervised domain adaptive object detection without source data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8474–8481.

Li, Z.; and Miao, D. 2021. Sequential end-to-end network for efficient person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liang, J.; Hu, D.; and Feng, J. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, 6028–6039. PMLR.

Lin, L.; Yang, Z.; Liu, Q.; Yu, Y.; and Lin, Q. 2023. Run and chase: Towards accurate source-free domain adaptive object detection. In *2023 IEEE International Conference on Multimedia and Expo*, 2453–2458. IEEE.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*.

Liu, F.; Ye, M.; and Du, B. 2024. Learning a generalizable re-identification model from unlabelled data with domain-agnostic expert. *Visual Intelligence*, 2(1): 28.

Liu, X.; Li, W.; and Yuan, Y. 2023. Decoupled Unbiased Teacher for Source-Free Domain Adaptive Medical Object Detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Lo, S.-Y.; Oza, P.; Chennupati, S.; Galindo, A.; and Patel, V. M. 2023. Spatio-temporal pixel-level contrastive learning-based source-free domain adaptation for video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10534–10543.

Qin, J.; Zheng, P.; Yan, Y.; Quan, R.; Cheng, X.; and Ni, B. 2023. Movienet-PS: A Large-Scale Person Search Dataset in the Wild. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Qu, X.; Liu, L.; Zhu, L.; Nie, L.; and Zhang, H. 2024. Instance-level Adversarial Source-free Domain Adaptive Person Re-identification. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, 27–28.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

VS, V.; Oza, P.; and Patel, V. M. 2023. Instance relation graph guided source-free domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3520–3530.

Wang, C.; Ma, B.; Chang, H.; Shan, S.; and Chen, X. 2020. Tcts: A task-consistent two-stage framework for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Yan, Y.; Li, J.; Qin, J.; Bai, S.; Liao, S.; Liu, L.; Zhu, F.; and Shao, L. 2021. Anchor-free person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yan, Y.; Li, J.; Qin, J.; Zheng, P.; Liao, S.; and Yang, X. 2023. Efficient person search: An anchor-free approach. *International Journal of Computer Vision*, 1–20.

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

Yang, F.; Li, K.; Zhong, Z.; Luo, Z.; Sun, X.; Cheng, H.; Guo, X.; Huang, F.; Ji, R.; and Li, S. 2020. Asymmetric co-teaching for unsupervised cross-domain person reidentification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12597–12604.

Yao, H.; and Xu, C. 2020. Joint person objectness and repulsion for person search. *IEEE Transactions on Image Processing*, 30: 685–696.

Yu, R.; Du, D.; LaLonde, R.; Davila, D.; Funk, C.; Hoogs, A.; and Clipp, B. 2022. Cascade transformers for end-to-end person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7267–7276.

Zhang, X.; Wang, X.; Bian, J.-W.; Shen, C.; and You, M. 2021. Diverse knowledge distillation for end-to-end person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3412–3420.

Zhang, Y.; Wang, Z.; and He, W. 2023. Class relationship embedded learning for source-free unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7619–7629.

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