Unsupervised Domain Adaptive Person Search via Dual Self-Calibration

Linfeng Qi[1], Huibing Wang[1]*, Jiqing Zhang[1], Jinjia Peng[2], Yang Wang[3]

[1]School of Information Science and Technology, Dalian Maritime University, Dalian, China
[2]School of Cyber Security and Computer, Hebei University, Baoding, China
[3]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
{qilinfeng, huibing.wang, jqz}@dlmu.edu.cn, pengjinjia@hbu.edu.cn, yangwang@hfut.edu.cn

Abstract

Unsupervised Domain Adaptive (UDA) person search focuses on employing the model trained on a labeled source domain dataset to a target domain dataset without any additional annotations. Most effective UDA person search methods typically utilize the ground truth of the source domain and pseudo-labels derived from clustering during the training process for domain adaptation. However, the performance of these approaches will be significantly restricted by the disrupting pseudo-labels resulting from inter-domain disparities. In this paper, we propose a Dual Self-Calibration (DSCA) framework for UDA person search that effectively eliminates the interference of noisy pseudo-labels by considering both the image-level and instance-level features perspectives. Specifically, we first present a simple yet effective Perception-Driven Adaptive Filter (PDAF) to adaptively predict a dynamic filter threshold based on input features. This threshold assists in eliminating noisy pseudo-boxes and other background interference, allowing our approach to focus on foreground targets and avoid indiscriminate domain adaptation. Besides, we further propose a Cluster Proxy Representation (CPR) module to enhance the update strategy of cluster representation, which mitigates the pollution of clusters from misidentified instances and effectively streamlines the training process for unlabeled target domains. With the above design, our method can achieve state-of-the-art (SOTA) performance on two benchmark datasets, with 80.2% mAP and 81.7% top-1 on the CUHK-SYSU dataset, with 39.9% mAP and 81.6% top-1 on the PRW dataset, which is comparable to or even exceeds the performance of some fully supervised methods.

Code — https://github.com/whbdmu/DSCA

Introduction

Person search is a task that combines pedestrian detection (Girshick et al. 2014; Ren et al. 2015) and re-identification (ReID) (Ye et al. 2020; Peng et al. 2020; Wang et al. 2022; Ye et al. 2023) to achieve localization and identification of the target pedestrians in real-life scenes. Supervised learning in person search tasks has made significant advancements (Jaffe and Zakhor 2023; Jiang et al. 2024).

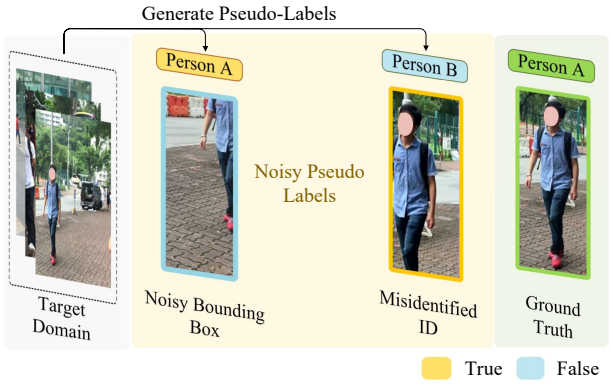

Figure 1: Noisy pseudo-labels consist of low-quality pseudo bounding boxes and misidentified pseudo identities. Where the quality of bounding boxes is the result of unsupervised detection, the reliability of the identities is affected by the clustering features.

However, models trained in one domain often struggle to generalize effectively to other domains due to factors such as person viewpoints and camera configuration (Liu et al. 2024). Moreover, annotating sufficient training data for a specific scenario is laborious and expensive. Therefore, unsupervised domain adaptation (UDA) offers significant potential in real-world scenarios for person search.

UDA (Ganin and Lempitsky 2015; Kang et al. 2019) for person search aims to narrow the gap between ideal training data and the realities of practical applications. However, this field is still in its infancy. Current approaches involve combining the pre-trained model on the source domain with clustering to produce pseudo labels on the target domain, which are then employed to facilitate the feature representation learning (Huang et al. 2023; Wang et al. 2024a). Specifically, DAPS (Li et al. 2022) is the pioneering work that proposes an implicit alignment module to transfer the knowledge of well-annotated source domain data to the unlabeled target domain for UDA person search. Based on this, DDAM (Almansoori, Fiaz, and Cholakkal 2024) further enhances domain adaptation by generating a hybrid domain representation, which improves knowledge transfer between the source and target domains. Although effective, these methods fail

*Corresponding Author.

to acknowledge that pseudo-labels frequently include incorrect labels due to domain gaps. As shown in Figure 1, the incorrect pseudo-labels are comprised of two parts: the noisy pseudo bounding boxes obtained from unsupervised detection, and the misidentified pseudo-IDs obtained from the clustering of the extracted instance-level features. These noisy labels potentially lead to misleading feature learning and impair the effectiveness of domain adaptation.

To tackle the above challenges, we propose a novel framework dubbed Dual Self-Calibration (DSCA) that eliminates the interference caused by noisy pseudo-labels on the UDA person search from both the image and instance perspectives. Specifically, we first present a Perceptually Driven Adaptive Filter (PDAF) to effectively identify and eliminate invalid regions of interest from image-level features, alleviating distractions from incorrect pseudo bounding boxes. Our PDAF consists of two key components: Perception-Driven Threshold (PDT) and Self-Calibrating Filter (SCF). The SCT has the ability to adaptively filter misjudged and meaningless features according to the threshold predicted by PDT, effectively enhancing the robustness of domain adaptation. Besides, to address the issue of clusters being contaminated by misidentified pseudo-ID instances, we further develop a novel Cluster Proxy Representation (CPR) to enhance the cluster update strategy. The CPR designs a memory structure that moves the utilization and update of the memory dictionary from the instance level to the cluster level. Misidentified instances are treated as part of a cluster proxy rather than as separate instances. As cluster proxies are continuously updated, the effect of misidentified instances will be eliminated. Note that our DSCA only requires clustering once per-training epoch, unlike other methods that generate pseudo-labels through multiple clustering (Ge et al. 2020). Meanwhile, the memory dictionary merely utilizes and updates the unique cluster proxy for instances in the same cluster. These measures effectively simplified the training process and thus improved efficiency. Extensive experiments on different target domain datasets validate the effectiveness of the proposed DSCA, which outperforms existing SOTA unsupervised domain adaptive methods by significant margins. Ablation experiments also evidence the importance of each key component of DSCA.

In summary, this paper makes the following contributions:

- This paper proposes a novel dual self-calibration framework for UDA person search. The framework ensures the effectiveness of domain adaptation by eliminating the negative effects of potentially erroneous pseudo-labels. Its performance is superior to the current SOTA methods.
- We present the Perception-Driven Adaptive Filter to suppress invalid pseudo bounding boxes at different scales in the backbone.
- We propose a simplified and efficient cluster update strategy, Cluster Proxy Representation, to address the contamination of clusters by misidentified instances.

Method

In this section, we introduce the Dual Self-Calibration (DSAC) framework. First, we outline the network architecture. Second, the details of the Perception Driven Adaptive Filter (PDAF) proposed in this paper are elaborated. Finally, we illustrate the working principle of Clustered Proxy Representation (CPR).

Framework Overview

The person search model is based on the end-to-end architecture of the DAPS network, which includes an implicit domain alignment module (DAM) (Chen et al. 2018b; Deng et al. 2018) designed to reduce the discrepancy between two domains. By this approach, we adhere to the principle of "sample quality first" and utilize our novel method to obtain more refined feature representation and pseudo-labels for the UDA tasks.

As illustrated in Figure 2, the DSCA framework comprises perception-calibrated person search and unlabeled target domain preprocessing. 1) In perception-calibrated person search, DSCA uses the first four convolutional layers of ResNet-50 (He et al. 2016) as a backbone network to extract image-level features from the input images in both the source and target domains. Implement PDAF for each layer of the backbone network to ensure the concentration of extracted features on the foreground targets, thereby facilitating multi-scale feature filtering. Subsequently, reliable image-level features are fed into the detection network via RPN(Ren et al. 2015) generate region proposals and regressed through the box head to output the detected bounding boxes. Finally, based on the instance-level features extracted from the bounding boxes, the more sophisticated detection and re-identification tasks are continued at the NAE head (Chen et al. 2020b). The DAM is used to eliminate inter-domain disparities at both the image and instance scales. 2) Before each epoch's training phase, DSCA applies the CPR pipeline for unlabeled target domain preprocessing. The network directly extracts training proposals from the target domain using RPN. The boxes obtained from training proposals are directly utilized as pseudo bounding boxes. To mitigate the impact of misidentified instances, DSCA presents CPR for training proposals, employing a single clustering strategy that directly assigns and utilizes pseudo-IDs. Furthermore, the average features within clusters are applied to initialize the memory dictionary for Contrastive Learning (He et al. 2020; Dai et al. 2022).

Perception-Driven Adaptive Filter

This paper proposes a Perception-Driven Adaptive Filter (PDAF). Through multi-scale filtering in the backbone, it eliminates misfocusing on scene information and ensures task flow reliability.

Perception-Driven Threshold. Backbone network is a powerful feature extractor designed to provide effective feature representations for downstream tasks. However, during the learning process of UDA, some invalid feature representations are likely to be erroneously recognized as targets and used for generating pseudo-labels due to their highlighting. Therefore, to obtain accurate filtering thresholds to eliminate these invalid representations, this paper proposes a Perception-Driven Threshold (PDT). The PDT distinguishes

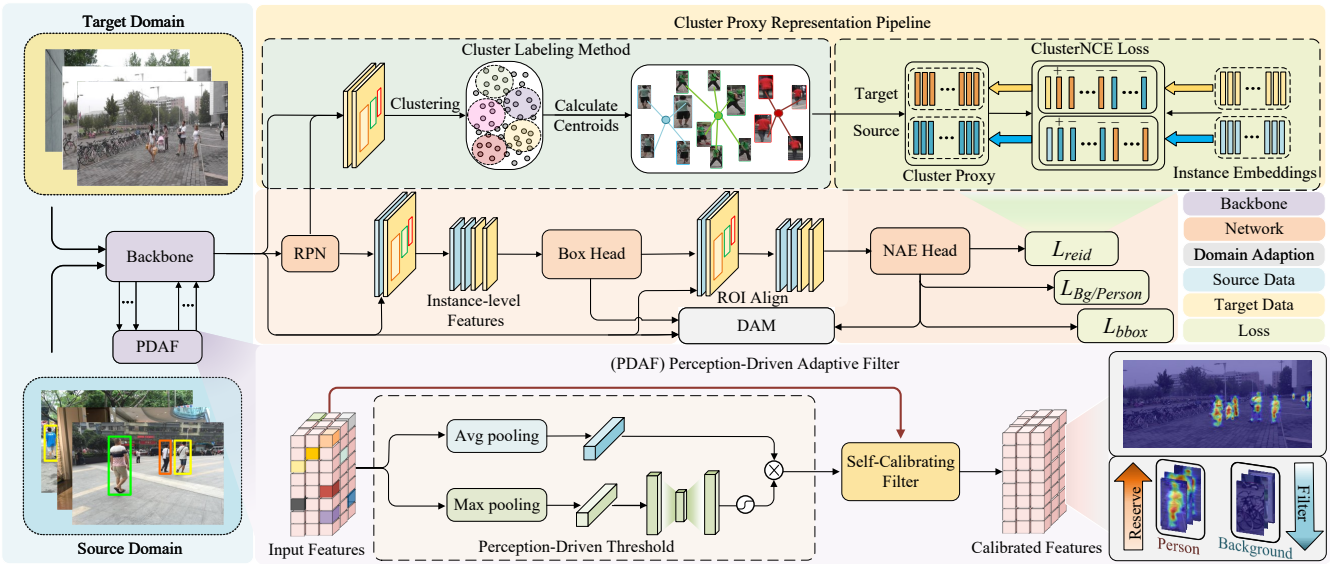

Figure 2: The design architecture of the DSCA framework. For each training period, DSCA alternates between two phases:(1) Cluster Proxy Representation Pipeline. Using RPN generated proposal boxes to annotate unlabeled samples in the target domain, the annotated samples are clustered to assign pseudo-labels and initialize the cluster proxy dictionary. (2) Perception-Driven Adaptive Filter. Image-level features with valid foreground information after PDAF purification are employed for downstream domain alignment and person search tasks.

scene information based on the foreground target perception capability obtained by training the model on the source domain.

Specifically, PDT generates a channel attention map using the inter-channel relationships of the features and uses this as an adaptive threshold. To achieve this, we used pooling to compress the input image-level features' dimensions for spatial information aggregation. With the average pooling and max pooling operations, we aggregate the spatial information of image-level features $F \in R^{H\times W\times C}$ from different perspectives to obtain two spatial contextual representations: $F_{avg} \in R^{1\times 1\times C}$ and $F_{max} \in R^{1\times 1\times C}$.

In the context of person search, average pooling aggregates global spatial messages and evaluates the overall effectiveness of the channel. Therefore, in this paper, we consider $F_{avg} \in R^{1\times 1\times C}$ as a kind of basic threshold $\tau' \in R^{1\times 1\times C}$ and use it as a benchmark for distinguishing between foreground and background information. It can be described as:

$$\tau' = AvgPool\left(F\right), \tag{1}$$

while τ' achieves a basic ability to distinguish and shows good performance in fixed scenarios, it is not flexible enough. For complex real-world scenarios, its effectiveness will be greatly reduced. Therefore, PDT uses the max pooling features F_{max} to guide τ' to ensure that the obtained thresholds can be adapted to scene changes. This is because max pooling features F_{max} can be employed as a key indicator to ascertain the presence of foreground information in channels, and they are relatively less sensitive to scene variations (Wei et al. 2021). Specifically, F_{max} be passed into a multi-layer perceptron (MLP) and use the sigmoid function to scale the output to the range of 0 to 1, so we can obtain the adaptive scaling factor. Finally, we use element-wise multiplication to obtain the PDT. In short, it can be calculated as:

$$\begin{aligned} \alpha &= \sigma\left(MLP\left(MaxPool\left(F\right)\right)\right), \\ \tau &= \alpha\left(AvgPool\left(F\right)\right), \end{aligned} \tag{2}$$

where σ denotes the sigmoid function, $\alpha \in R^{1\times 1\times C}$ represents the adaptive scaling factor, and $\tau \in R^{1\times 1\times C}$ represents our Perception-Driven Threshold.

Self-Calibrating Filter. The design of the threshold function is of paramount importance in the context of filter design. Traditional threshold functions (Donoho and Johnstone 1994; Donoho 1995) encompass both hard and soft thresholds. However, neither of these functions is optimal for person search.

The hard threshold function can directly filter out the part of the input features that are smaller than the threshold value without interfering with the foreground features. It can be expressed as follows:

$$f_{hard}(x,\tau) = \begin{cases} x & |x| \geq \tau \\ 0 & |x| < \tau, \end{cases} \tag{3}$$

where x denotes the value of the input and τ means the threshold. However, this simply truncated approach impairs the structural information and affects the ability to model fitting. The soft threshold function performs a panning operation on the input features according to the threshold value and then eliminates the portion below the threshold. It can be described as:

$$f_{soft}(x,\tau) = \begin{cases} x-\tau & x \geq \tau \\ 0 & |x| < \tau \\ x+\tau & x \leq -\tau. \end{cases} \tag{4}$$

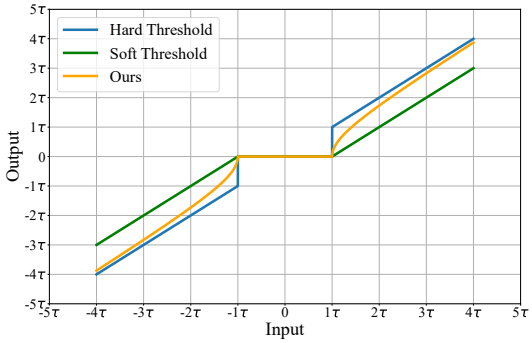

Figure 3: Comparison of our High-Order Soft Threshold with two classical threshold functions, take $n = 2$ as an example.

The application of a soft threshold ensures the smooth filtering of features. Nonetheless, the filtering process results in a constant bias between the input and output features, which in turn affects the quality of the foreground features.

Based on the above analyses, we rethink the design of the threshold function and propose a Self-Calibrating Filter (SCF). It fully combines the advantages of the traditional threshold functions while overcoming the shortcomings of both. This filter consists of a High-Order Soft Threshold (HST) and a Self-Calibrating Transition (SCT) in two parts. Specifically, HST further improves the accuracy of the filter by using a high-order function instead of the traditional linear method. As shown in Figure 3, HST fully adapts to the needs of the person search task and grasps the differences between the foreground features and scene features. The formula of HST is as follows:

$$f_{hst}(x,\tau) = \begin{cases} sgn(x)(|x|^n - \tau^n)^{\frac{1}{n}} & |x| \geq \tau \\ 0 & |x| < \tau, \end{cases} \tag{5}$$

where $sgn()$ returns the sign of the input value. Although theoretically, the larger n is, the better the function fitted by HST. In actual applications, due to the use of higher-order functions instead of linear computation, neural networks may suffer from potential gradient explosion and gradient vanishing problems. This makes it difficult for gradient descent-based optimization algorithms to update model parameters effectively. To solve this problem, we propose the Self-Calibrating Transition (SCT) strategy. By introducing learnable parameters, we can self-calibrate and adjust the form of the threshold function. This allows the function to be closer to the soft threshold at the beginning of training, and then gradually transition to HST as the training proceeds. The formula of SCF is as follows:

$$f_{scf}(x,\tau) = \begin{cases} \lambda\left(x^2 - \tau^2\right)^{\frac{1}{2}} + (1-\lambda)(x-\tau) & x \geq \tau \\ 0 & |x| < \tau \\ -\lambda\left(x^2 - \tau^2\right)^{\frac{1}{2}} + (1-\lambda)(x+\tau) & x \leq -\tau, \end{cases} \tag{6}$$

where $\lambda \in R^{1\times1\times C}$ is a set of learnable parameters initialized as 0. This gradual transition strategy can not only effectively alleviate the gradient problem but also give full play to the advantages of HST in terms of expressive ability.

Cluster Proxy Representation

To reduce the difficulty of pseudo-labels optimization in UDA person search, some previous studies have designed an instance-level contrastive learning framework inspired by SPCL (Ge et al. 2020). Specifically, a memory dictionary containing all instances is built and updated, but the utility is at the cluster level. This approach can lead to meaningless or even misidentified samples being included as target instances in the training process.

In this paper, we design the Cluster Proxy Representation (CPR). CPR is different from existing memory structures. It uses a unique proxy to represent each category during training, and it updates the memory by cluster, thus effectively reducing the effect of instance pollution.

Cluster Labeling Method. This paper uses the person search framework pre-trained in the source domain to generate training proposals and pseudo bounding boxes, then uses the DBSCAN (Ester et al. 1996) algorithm to assign the pseudo-labels. After the above operation, the unlabeled dataset D_T becomes $D'_T = \left\{(x_i, l_i^1, ..., l_i^n)\right\}_{i=1}^N$, where x_i denotes the i-th image in this dataset, l_i is the pseudo-label of the instances in this image, N denotes the number of images, and n denotes the number of instances in the image.

For each cluster, whether it belongs to the source or target domain, we maintain a unique cluster proxy representation, which is initialized with the average instance features of it:

$$c_k = \frac{1}{N_k} \sum_{f_i \in F_k} f_i, \tag{7}$$

where c_k denotes the cluster proxy representation for the k-th cluster, F_k represents a set of features with the same cluster k, and f_i is the feature of each instance. Then use the memory dictionary $M = \{(c_1, c_2, ..., c_K)\}$ to store these representation uniformly.

Cluster Memory Updating. For CPR, memory updates consist of two parts: online and offline. For the online phase, this work recognizes the M as a look-up table (LUT) (Xiao et al. 2017) that stores all marked pedestrian proxy representations and calculates the loss using the following equation:

$$L = -\log \frac{\exp\left(f \cdot c_+ / \tau\right)}{\sum_{k=0}^{K} \exp\left(f \cdot c_k / \tau\right)}, \tag{8}$$

where f is an instance query feature extracted for re-identification and c_+ shares the same cluster with f. τ is a temperature factor that adjusts the softness of the probability distribution. During backward, if the target label is k, then we will update the k-th column of the M by

$$c_k \longleftarrow (1-\gamma)\, c_k + \gamma f, \tag{9}$$

where $\gamma \in [0, 1]$ is the momentum factor used to control the update speed of the proxy. Through this cluster-level update strategy, the pollution caused by misidentified instances will be overwritten by subsequent updates.

In the offline stage, to confirm the effectiveness of the cluster proxy in the entire UDA learning process, this paper uses an asynchronous learning strategy. In each training

Method	Backbone	PRW		CUHK-SYSU	
		mAP	top-1	mAP	top-1
Fully-Supervised					
OIM (Xiao et al. 2017)	ResNet-50	21.3	49.4	75.5	78.7
MGTS(Chen et al. 2018a)	VGG-16	32.6	72.1	83.0	83.7
HOIM(Chen et al. 2020a)	ResNet-50	39.8	80.4	89.7	90.8
NAE+ (Chen et al. 2020b)	ResNet-50	44.0	81.1	92.1	92.9
RDLR (Han et al. 2019)	ResNet-50	42.9	70.2	93.0	94.2
TCTS (Wang et al. 2020)	ResNet-50	46.8	87.5	93.9	95.1
AlignPS+ (Yan et al. 2021)	ResNet-50	46.1	82.1	94.0	94.5
SeqNet (Li and Miao 2021)	ResNet-50	46.7	83.4	93.8	94.6
PSTR (Cao et al. 2022)	PVTv2-B2	56.5	89.7	95.2	96.2
SeqNeXt (Jaffe and Zakhor 2023)	ConvNeXt-B	57.6	89.5	96.1	96.5
SEAS(Jiang et al. 2024)	ConvNeXt-B	60.5	89.5	97.1	97.8
Weakly-Supervised					
CGPS(Yan et al. 2022)	ResNet-50	16.2	68.0	80.0	82.3
R-SiamNet(Han et al. 2021)	ResNet-50	21.2	73.4	86.0	87.1
SSL(Wang et al. 2023)	ResNet-50	30.7	80.6	87.4	88.5
DICL(Wang et al. 2024b)	ResNet-50	35.5	80.9	87.4	88.8
Unsupervised					
DAPS(Li et al. 2022)	ResNet-50	34.7	80.6	77.6	79.6
DDAM(Almansoori, Fiaz, and Cholakkal 2024)	ResNet-50	36.7	81.2	79.5	81.3
DSCA(Ours)	ResNet-50	**39.9**	**81.6**	**80.2**	**81.7**

Table 1: Comparison of mAP(%) and top-1 accuracy(%) with the fully supervised, weakly supervised, and unsupervised methods on CUHK-SYSU and PRW datasets.

epoch, the clustered proxy and pseudo-labels are reinitialized together. At the same time, to ensure consistency of the training objectives, we introduce instance features extracted from the previous epoch and smooth the instances using the Exponential Moving Average (EMA). This process can be described as follows:

$$c_k = \frac{1}{N_k} \sum_{f_i^t \in F_k^t} \left[(1-m)\, f_i^t + m \left(Match\left(F^{t-1}, f_i^t \right) \right) \right], \quad (10)$$

where t and $t-1$ represent the training epoch. F^{t-1} represents all the instance features extracted by the previous epoch. The $Match$ operation refers to returning the instance in F^{t-1} that coincides with the pseudo bounding box corresponding to the current instance f_i^t, and m represents the smoothing factor.

Experiments

Experimental Settings

Datasets. We evaluate our DCSA on two person search datasets: CUHK-SYSU (Xiao et al. 2017) and PRW (Zheng et al. 2017). The CUHK-SYSU dataset is a large-scale person search benchmark comprising a total of 18,184 images. The dataset contains 8,432 unique person identities and 96,143 annotated bounding boxes. It is divided into a training set with 5,532 identities and 11,206 images, and a test set with 2,900 query persons and 6,978 gallery images. The PRW dataset includes 11,816 scene images with annotations for 932 unique person identities and 43,110 bounding boxes. The training set comprises 932 identities with 5,704 images, while the test set contains 2,057 query persons and 6,112 scene images.

Metrics. We adopt two widely used metrics for quantitatively evaluating person search performance: the mean Average Precision (mAP) and top-1 accuracy. Note that all evaluations are performed on the test set of the target domain without any additional annotations.

Implementation Details. We implemented our DSCA using Pytorch and trained it on an NVIDIA A800 GPU with a batch size of 4. We adopted the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.0024, which is warmed up in the first epoch. We employed the pre-trained ResNet50 (He et al. 2016) as our backbone. During training, the input images are resized to 1500×900, and random horizontal flipping is applied for data augmentation. We set both the momentum factor γ and smoothing factor m to 0.2 for online and offline cluster updating, respectively. To prevent overfitting, we adjust the number of epochs based on the variations in the target domain dataset. Specifically, when PRW is used as the target domain, our DSCA undergoes pre-training on the source domain CUHK-SYSU for 7 epochs before commencing joint training for 13 epochs. Conversely, we first pre-train on the source domain PRW for 2 epochs, followed by beginning joint training for 7 epochs.

Comparison with State-of-the-Art Methods

To demonstrate the effectiveness of our method, we compare DSCA with SOTA approaches that employ various training strategies, including fully supervised, weakly supervised, and unsupervised methods. As shown in Table 1, our DCSA outperforms other compared unsupervised domain adaptive approaches on both two datasets in terms of both mAP and top-1 scores. For example, our method achieves 80.2% over-

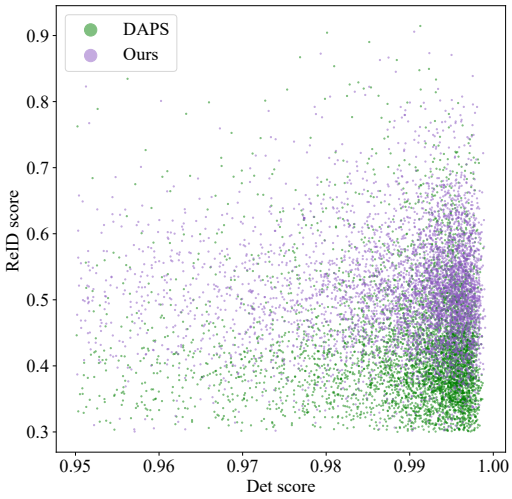

Figure 4: Visualization of the detection scores and re-identification scores on the CUHK-SYSU dataset. Purple and green dots indicate the results of DAPS and ours.

Components			PRW		CUHK-SYSU	
PDT	SCF	CPR	mAP	top-1	mAP	top-1
✗			39.0	80.8	76.9	78.3
	✗		38.4	80.8	79.6	81.2
✗	✗		36.2	79.7	78.2	80.5
		✗	39.2	81.1	79.5	80.7
✗	✗	✗	34,7	80.6	77.6	79.6
			39.9	**81.6**	**80.2**	**81.7**

Table 2: Validating the effectiveness of different components on the CUHK-SYSU and PRW datasets. CPR: Cluster Proxy Representation. PDT: Perception-Driven Threshold. SCF: Self-Calibrating Filter.

all mAP and 81.7% top-1 on the CUHK-SYSU dataset, outperforming the runner-up by 0.7% and 0.4%, respectively. This demonstrates that eliminating the disruptive influence of noisy pseudo-labels in domain adaptation can contribute to enhancing the robustness of UDA person search. Figure 5 further qualitatively demonstrates the benefits of our method in some challenging conditions, including cross-scene, low-resolution, occlusion, and viewpoint variation.

Surprisingly, Table 1 shows that our DSCA continues to deliver comparable performance in comparison to both fully supervised and weakly supervised approaches. Particularly on the PRW dataset, our method surpasses all compared weakly supervised methods by a large margin, outperforming the best one by 4.4% and 0.7% in mAP and top-1, respectively. Besides, the results show a noticeable disparity between DSCA and SOTA fully supervised methods, suggesting that there is still potential for improvement to bridge this gap. In this work, we prioritize minimizing data dependencies and achieving a favorable balance between performance and resource efficiency. We also anticipate that our approach will stimulate further investigation in this field.

Figure 5: Qualitative comparison of DSCA with DAPS on the CUHK-SYSU test set. The green bounding boxes denote the queries, while the red and orange bounding boxes denote incorrect and correct matches, respectively.

Ablation Study

Impact on Perception-Driven Adaptive Filter. Our proposed Perception-Driven Adaptive Filter (PDAF) aims to filter out the interference of erroneous pseudo-boxes on domain adaptation, which has two key components: Perception-Driven Threshold (PDT) module and Self-Calibrating Filter (SCF). We thus conduct the following experiments to validate the effectiveness of PDAF: (i) without PDT (*i.e.*, replacing PDT with the global average operation for input features) (ii) without SCF; (iii) without PDT and SCF. As shown in Table 2 #1 and #2, removing either PDT or SCF results in significant performance degradation on both datasets. This validates that the exclusion of noisy pseudo-boxes is essential for person search accuracy. Removing both PDT and SCF leads to an additional decrease in performance, as shown in #3, providing further evidence of the effectiveness of our design.

To further validate the effectiveness of PDT and SCF, we conduct additional ablation experiments: (i) replacing PDT with the Adaptive Mean Threshold (AMT) module (Zhao et al. 2019), which is a typical structure that mainly employs global average pooling to predict thresholds; (ii) replacing SCF with Hard and Soft Threshold schemes. As shown in Table 3, our proposed PDT consistently outperforms AMT counterparts in terms of both mAP and top-1 scores, suggesting the effectiveness of our PDT strategy to combine local foreground and global spatial information. Moreover,

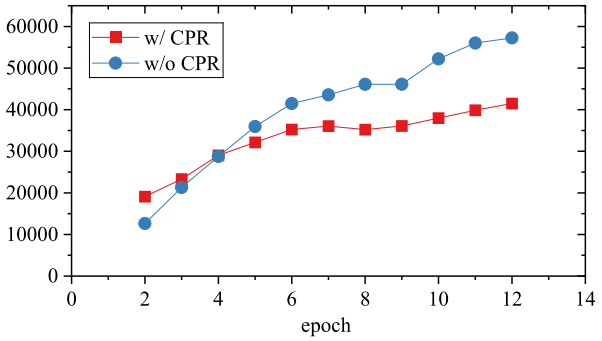

Figure 6: Number of representations stored in the memory dictionary, when the CUHK-SYSU dataset is used as the target domain.

PDAF		PRW		CUHK-SYSU	
Threshold	Filter	mAP	top-1	mAP	top-1
AMT	Hard	38.6	81.0	78.7	80.1
AMT	Soft	38.4	80.3	77.9	80.1
AMT	SCF	39.2	81.3	79.5	81.2
PDT	Hard	38.1	79.9	79.3	80.9
PDT	Soft	38.4	80.8	79.6	81.2
PDT	SCF	**39.9**	**81.6**	**80.2**	**81.7**

Table 3: Results of the ablation study on Perception-Driven Adaptive Filter, compare results of the different thresholds and filters.

Table 3 shows that the performance of our proposed filtering method SCF considerably exceeds the standard Hard and Soft threshold approaches. This is because conventional soft and hard thresholds are constrained by inter-domain disparities. Specifically, for the scene-rich SYSU-CUHK dataset, the hard threshold cannot adapt to scene variations. In contrast, soft thresholds are inhibited by monotonous scenes in the PRW dataset. Our SCF fully integrates the strengths of both to achieve the best overall performance.

To get more insights into PDAF, we adopt LayerCAM (Jiang et al. 2021) to visualize the attention maps of the backbone outputs. As shown in Figure 7, under the guidance of PDAF, our method pays more attention to the location of the target, greatly reducing noise interference in the domain adaptation process and thus enhancing the reliability of person search tasks.

Effectiveness of Cluster Proxy Representation. The Cluster Proxy Representation (CPR) scheme is another essential component of our approach, with a focus on enhancing training efficiency by reducing the impact of misidentified instances on cluster pollution. As shown in Table 2 #4, the removal of CPR results in a decrease in both mAP and top-1 scores on the PRW and CUHK-SYSU datasets, which serves as evidence of the effectiveness of our proposed cluster proxy. As shown in Figure 6, The use of CPR also greatly reduces GPU memory consumption.

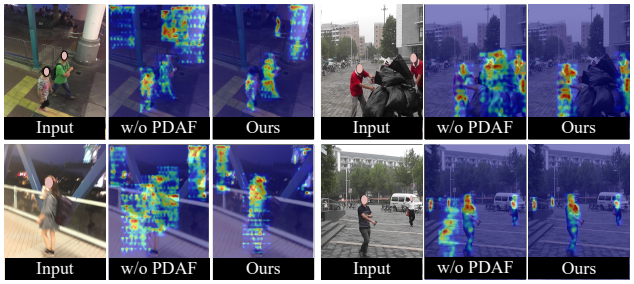

Figure 7: Visualization of the LayerCAM for the last layer of the backbone for each predicted object (images from CUHK-SYSU and PRW datasets). Evaluated separately with and without PDAF, the perceptual scores are differentiated by different colors, from blue to red, representing increasing perceptual scores.

Method	$\textbf{Prop}_{0.3}\uparrow$	$\textbf{Prop}_{0.5}\uparrow$	$\textbf{Prop}_{0.7}\uparrow$
Ours	**99.7%**	**52.8%**	**2.3%**
DAPS	97.0%	17.5%	1.8%
$\Delta(drop)$	2.7%	35.3%	0.5%

Table 4: Results obtained from the evaluation of the CUHK-SYSU dataset. $\textbf{Prop}_{0.3}$, $\textbf{Prop}_{0.5}$ and $\textbf{Prop}_{0.7}$ denote the proportion of re-identification scores greater than 0.3, 0.5, 0.7, respectively.

Performance on Person Detection and ReID. To further illustrate the effectiveness of our DSCA, we examine the performance of the two sub-tasks involved in person search: pedestrian detection and ReID. As shown in Figure 4, our method obtains significantly higher overall scores in both detection and ReID on the CUHK-SYSU dataset when contrasted with another SOTA unsupervised method DAPS. We conduct a more detailed examination of the performance of ReID at various thresholds. Table 4 shows our method significantly outperforms the DAPS across all evaluated thresholds of ReID scores. Specifically, with a threshold of 0.5, our DSCA achieves a significant rate of 52.8%, surpassing DAPS by a notable margin of 35.3%. The results demonstrate that our design for mitigating the impact of noisy pseudo-labels significantly enhances the overall performance of both the detection and ReID sub-tasks.

Conclusion

In this paper, we propose a Dual Self-Calibration (DSCA) framework to remove the interference caused by noisy pseudo-labels due to inter-domain disparities for UDA person search. This disturbance originates from both images and instances. For these two kinds of interference, we design perception-driven adaptive filter and cluster proxy representation to achieve the filtering of noisy pseudo-boxes and the elimination of misidentified instances, respectively.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant 62172136, U21A20470 and 72188101, by the Institute of Advanced Medicine and Frontier Technology under Grant 2023 IHM01080, by National Key Research and Development Program of China Grant 2024YFB4710800, Liaoning Provincial Natural Science Foundation Grant 2024-MS-012. This work was carried out in part using computing resources at the High Performance Computing Center of Dalian Maritime University.

References

Almansoori, M. K.; Fiaz, M.; and Cholakkal, H. 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.

Cao, J.; Pang, Y.; Anwer, R. M.; Cholakkal, H.; Xie, J.; Shah, M.; and Khan, F. S. 2022. PSTR: End-to-End One-Step Person Search With Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9458–9467.

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*, volume 34, 10518–10525.

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

Chen, D.; Zhang, S.; Yang, J.; and Schiele, B. 2020b. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12615–12624.

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2018b. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3339–3348.

Dai, Z.; Wang, G.; Yuan, W.; Zhu, S.; and Tan, P. 2022. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian conference on computer vision*, 1142–1160.

Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; and Jiao, J. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 994–1003.

Donoho, D. L. 1995. De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3): 613–627.

Donoho, D. L.; and Johnstone, I. M. 1994. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3): 425–455.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 226–231.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

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.

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

Han, C.; Su, K.; Yu, D.; Yuan, Z.; Gao, C.; Sang, N.; Yang, Y.; and Wang, C. 2021. Weakly supervised person search with region siamese networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12006–12015.

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*, 9814–9823.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

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

Huang, Z.; Ren, Y.; Pu, X.; Huang, S.; Xu, Z.; and He, L. 2023. Self-supervised graph attention networks for deep weighted multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7936–7943.

Jaffe, L.; and Zakhor, A. 2023. Gallery filter network for person search. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1684–1693.

Jiang, P.-T.; Zhang, C.-B.; Hou, Q.; Cheng, M.-M.; and Wei, Y. 2021. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30: 5875–5888.

Jiang, Y.; Wang, H.; Peng, J.; Fu, X.; and Wang, Y. 2024. Scene-Adaptive Person Search via Bilateral Modulations. In *Proceedings of the Thirty-Third International Conference on International Joint Conferences on Artificial Intelligence*.

Kang, G.; Jiang, L.; Yang, Y.; and Hauptmann, A. G. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4893–4902.

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

Li, Z.; and Miao, D. 2021. Sequential end-to-end network for efficient person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2011–2019.

Liu, H.; Wang, Y.; Qian, B.; Wang, M.; and Rui, Y. 2024. Structure Matters: Tackling the Semantic Discrepancy in Diffusion Models for Image Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8038–8047.

Peng, J.; Wang, Y.; Wang, H.; Zhang, Z.; Fu, X.; and Wang, M. 2020. Unsupervised vehicle re-identification with progressive adaptation. *arXiv preprint arXiv:2006.11486*.

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.

Wang, B.; Yang, Y.; Wu, J.; Qi, G.-j.; and Lei, Z. 2023. Self-similarity driven scale-invariant learning for weakly supervised person search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1813–1822.

Wang, C.; Ma, B.; Chang, H.; Shan, S.; and Chen, X. 2020. Tcts: A task-consistent two-stage framework for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11952–11961.

Wang, H.; Yao, M.; Chen, Y.; Xu, Y.; Liu, H.; Jia, W.; Fu, X.; and Wang, Y. 2024a. Manifold-based Incomplete Multi-view Clustering via Bi-Consistency Guidance. *IEEE Transactions on Multimedia*.

Wang, J.; Pang, Y.; Cao, J.; Sun, H.; Shao, Z.; and Li, X. 2024b. Deep intra-image contrastive learning for weakly supervised one-step person search. *Pattern Recognition*, 147: 110047.

Wang, Y.; Peng, J.; Wang, H.; and Wang, M. 2022. Progressive learning with multi-scale attention network for cross-domain vehicle re-identification. *Science China Information Sciences*, 65(6): 160103.

Wei, X.-S.; Song, Y.-Z.; Mac Aodha, O.; Wu, J.; Peng, Y.; Tang, J.; Yang, J.; and Belongie, S. 2021. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12): 8927–8948.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3415–3424.

Yan, Y.; Li, J.; Liao, S.; Qin, J.; Ni, B.; Lu, K.; and Yang, X. 2022. Exploring visual context for weakly supervised person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3027–3035.

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*, 7690–7699.

Ye, M.; Lan, X.; Leng, Q.; and Shen, J. 2020. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE Transactions on Image Processing*, 29: 9387–9399.

Ye, M.; Wu, Z.; Chen, C.; and Du, B. 2023. Channel augmentation for visible-infrared re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhao, M.; Zhong, S.; Fu, X.; Tang, B.; and Pecht, M. 2019. Deep residual shrinkage networks for fault diagnosis. *IEEE Transactions on Industrial Informatics*, 16(7): 4681–4690.

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1367–1376.