Instance-Guided Scene Adaptation for Unsupervised Person Search

Linfeng Qi[1], Huibing Wang[1]*, Jinjia Peng[2], Xianping Fu[1], Jiqing Zhang[1]

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

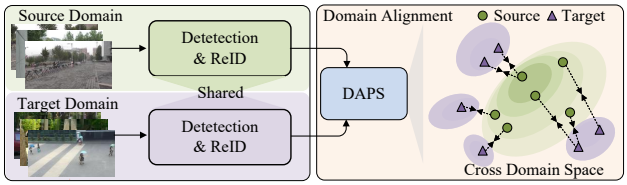
Abstract

Unsupervised Domain Adaptation (UDA) is a challenging task in person search. It adapts a well-trained model from a labeled source domain to an unlabeled target domain for privacy and efficiency. Currently, most of the state-of-the-art UDA person search methods adopt multi-scale feature alignment techniques to learn domain-invariant representations. However, person search is a multi-granularity task, and such an indiscriminate method of bridging the differences between domains misleads the identity learning process, which significantly limits the model's performance. In this paper, we propose an Instance-Guided Scene Adaptation (IGSA) framework by eradicating scene disparities and focusing the tasks on instances, effectively eliminating the contradiction between person search and domain adaptation. In IGSA, a Scene-Aware Bidirectional Filter (SABF) is designed to divide the image features into background and foreground to perform bidirectional modulations, thereby achieving simultaneous scene elimination and instance enhancement. To further improve the reliability of identity learning, we also propose an Instance Consistency Contrastive Learning (ICCL) method. By performing cross-epoch updates on the instance-level memory bank and re-initializing the cluster-level memory bank, the problem of inconsistent training across epochs caused by instance identity drift can be alleviated. Through the above designs, our method can achieve state-of-the-art performance on two benchmark datasets, with 82.1% mAP and 83.8% top-1 on the CUHK-SYSU dataset and 41.1% mAP and 82.3% top-1 on the PRW dataset, which is even better than some supervised methods.

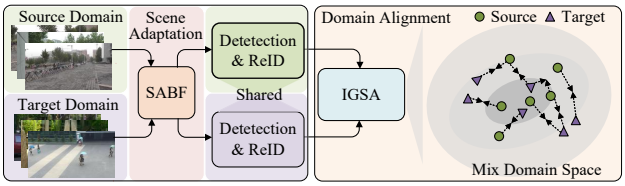
Code — https://github.com/whbdmu/IGSA

Introduction

Person search aims to locate and identify specific individuals from real-world scenes, which can essentially be regarded as a joint operation of person detection (Girshick et al. 2014; Ren et al. 2015) and ReID (Ye et al. 2023; Peng, Jiang, and Wang 2023; Peng, Zhang, and Wang 2025). Most person search methods follow the supervised paradigm to handle these two tasks in an end-to-end manner (Li and Miao 2021; Jaffe and Zakhor 2023; Jiang et al. 2024). However, due to domain disparities that stem from elements such as alterations in scene or camera configurations, a model trained within a particular domain often struggles to be effectively applied in an uncharted domain, which greatly limits the practical applicability of supervised person search.



(a) Domian Adaptive Person Search

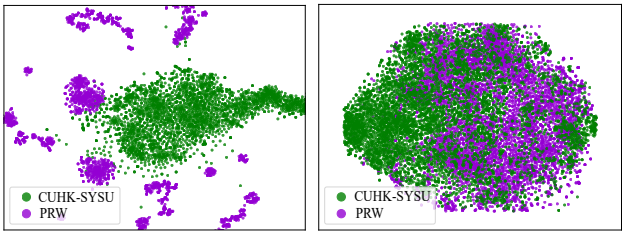


(b) Our Proposed Instance-Guided Scene Adaptation

Figure 1: Differences in operation between our proposed method and mainstream methods. (a) The DAPS framework utilizes alignment techniques to reduce domain disparities. However, due to the complexity of real-world scenes, this indiscriminate approach to bridging domain disparities may considerably limit the model's performance. (b) In contrast, we propose the IGSA framework, which enables the model to mitigate the impact of scene variations and simultaneously achieve foreground-target knowledge transfer.

Unsupervised Domain Adaptation (UDA) (Ganin and Lempitsky 2015; Kang et al. 2019) technology shows great potential in real-world person search. UDA focuses on narrowing the gap between the training environment and practical applications, transferring the prior knowledge learned by the model from the well-annotated source domain data to the unlabeled target domain. However, this technology is still in its infancy in the field of person search. Current approaches are to employ pre-trained models to generate pseudo-labels in the target domain and then conduct joint training with the source domain. Specifically, Li et

*Corresponding Author.



(a) Scene-wise distributions (b) Instance-wise distributions

Figure 2: Visualization of 2D distributions of samples from CUHK-SYSU (Xiao et al. 2017) PRW (Zheng et al. 2017) by t-SNE. a shows the distributions of the complete scene images, while b presents the distributions of the target instances cropped from images.

al. (Li et al. 2022) first proposed the UDA person search task and designed the DAPS framework. As shown in Figure 1a, DAPS includes an implicit alignment module that promotes the network to learn domain-invariant representations. FOUS (Cui et al. 2024), from the perspective of sample quality, proposed an Attention-based Domain Alignment method to eliminate noisy pseudo-labels. DDAM (Almansoori, Fiaz, and Cholakkal 2024) further advanced on the basis of DAPS, enhancing domain adaptability by generating mixed-domain representations. Though the aforementioned methods have reaped remarkable success, they fail to take into account the complexity of cross-domain scene disparities, as illustrated in Figure 2a. Merely relying on domain alignment modules in an attempt to address such disparities inevitably renders significant diversities among the instances with the same identity (Feng et al. 2018; Yao et al. 2024), resulting in performance degradation of person search.

To overcome the dilemma of existing methods, this paper proposes a novel Instance-Guided Scene Adaptation (IGSA) framework. As illustrated in Figure 1b, IGSA specifies the objects of feature extraction to foreground instances, alleviating the barriers to identity recognition caused by cross-domain scene alignments. To achieve this, IGSA incorporates a Scene-Aware Bidirectional Filter (SABF) for bidirectional modulation of image features. Specifically, SABF differentiates between instances and backgrounds in an image based on the instance sensitivity of the model trained on the source domain and then performs enhancement and filtering, respectively. This ensures that the domain alignment targets at all scales always concentrate on the foreground instances, avoiding interference from irrelevant information. In addition, considering the problem of inconsistent cross-epoch identity learning caused by instance identity drift in the target domain, this paper further proposes an Instance Consistency Contrastive Learning (ICCL) method. ICCL constructs memory banks at two levels. By performing cross-epoch updates on the instance-level memory bank and re-initializing the cluster-level memory bank based on the update results, the identity of instances is more stable throughout the training process. A large number of experiments conducted on different target domain datasets have verified the effectiveness of the proposed IGSA framework, and its performance is significantly better than state-of-the-art UDA methods.

In summary, this paper makes the following contributions:

- This paper proposes a novel Instance-Guided Scene Adaptation framework for UDA person search, which adheres to the principle of "Instance Priority" to eliminate the contradiction between person search and domain adaptation, which achieves state-of-the-art performance.
- A novel Scene Aware Bidirectional Filter is designed to simultaneously perform background elimination and foreground enhancement, which overcomes the cross-domain scene disparities for unsupervised person search.
- A novelty Instance Consistency Contrastive Learning method is designed to effectively solve the problem of inconsistent optimization directions across epochs caused by instance identity drift.

Method

In this section, we will introduce the Instance-Guided Scene Adaptation (IGSA) framework. Firstly, an overview of the framework's process will be presented. Secondly, the Scene-Aware Bidirectional Filter (SABF) proposed in this paper will be elaborated upon in detail. Finally, the working principle of the Instance Consistency Contrastive Learning (ICCL) will be illustrated.

Framework Overview

Based on DAPS (Li et al. 2022), IGSA is also designed with multi-scale implicit domain alignment modules in an end-to-end structure, which aims to reduce the disparities between different domains. However, unilaterally reducing disparities between domains from the global perspective neglects the foreground instances, which are decisive factors for person search. To adhere to the principle of "Instance Priority" to focus foreground instances, we proposed a Scene-aware Bidirectional Filters to the backbone and designed a new contrastive learning method. The complete training flow of IGSA consists of two parts: source domain pre-training and cross-domain joint training.

In the source domain pre-training stage, the model is trained on the accurately annotated source domain dataset to acquire the ability to perform person search tasks. The detailed process is as follows: The backbone first extracts the image features of source domain samples, then performs the ROIAlign (Ren et al. 2015) operation according to the GT bounding boxes to obtain the target instance features. Subsequently, the target instance features are directly classified based on GT-IDs, and the corresponding category centers are calculated to initialize the source domain's online memory bank M^s_{on}. Finally, multiple training epochs are carried out according to the supervised learning mode.

As shown in Figure 3, in cross-domain joint training, each epoch has two steps. During the target domain offline processing phase, the model is applied to evaluate the target domain dataset, and the resultant detection outcomes along with the corresponding instances are utilized to update the target domain's offline memory bank M_{off}. Subsequently, the updated offline memory bank M_{off} is clustered to assign

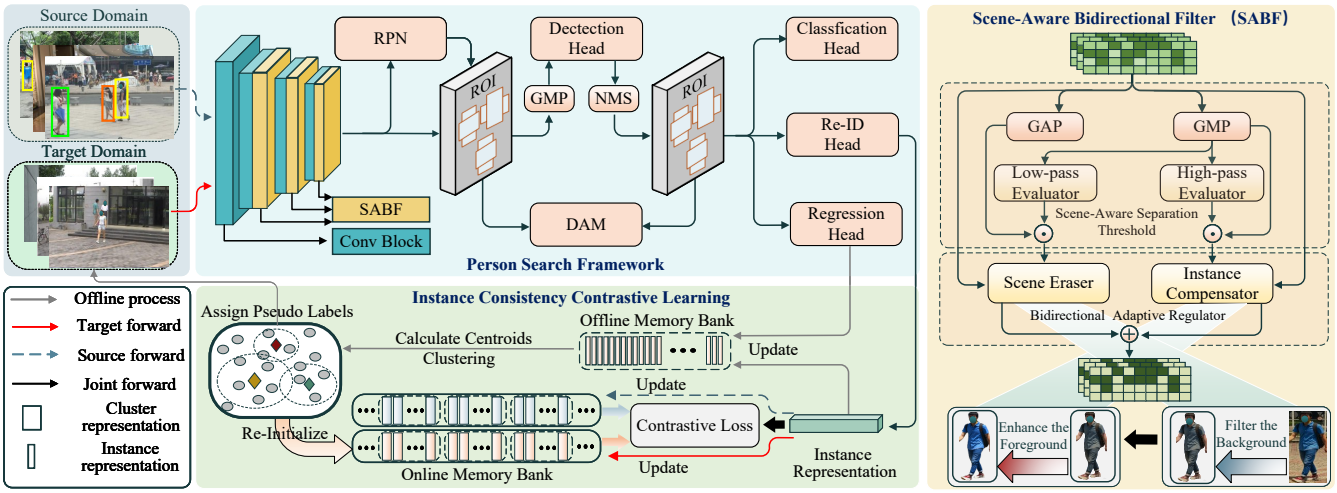


Figure 3: The process of IGSA in the cross-domain joint training stage is as follows: (a) Offline processing stage. In this stage, unlabeled samples in the target domain are relabeled, and the online memory bank of the target domain is reinitialized based on the updated samples. (b) Cross-domain online training stage. The image-level features with only effective instance information remaining after being refined by SABF will be applied to subsequent domain alignment and person search tasks.

pseudo-IDs. These pseudo-IDs, in conjunction with the detection boxes, re-annotate the target domain dataset. Concurrently, the instance features calculate the cluster centers and employ them to re-initialize the target domain's online memory bank M_{on}^{t}. In the cross-domain online training step, the target domain dataset with pseudo-labels is jointly learned with the source domain in a supervised way and uses the DAM module for domain alignment.

Scene-Aware Bidirectional Filter

Analysis of Inter-Domain Disparities. The design of DAPS (Li et al. 2022) has augmented the model's generalization capabilities and enhanced its performance. Nevertheless, despite the separate optimizations for detection and Re-ID, the outcomes of the ablation study reveal that the elevation in the Re-ID accuracy is predominantly due to the enhancement of detection performance, which supplies a greater number of effective training samples. The substantial cross-domain disparities still result in restricted fine-grained knowledge transfer (Wei et al. 2021). To overcome this obstacle, we initially examined the essential manifestations of the inter-domain disparities present among the datasets. As illustrated in Figure 2, by employing t-SNE (Van der Maaten and Hinton 2008) for dimensionality reduction of the datasets, it is observed that in contrast to the person instances possessing similar structural and semantic characteristics, the distinctions between complete scene images tend to be more conspicuous and burdensome to bridge.

Based on the above analysis, this paper proposes a Scene-Aware Bidirectional Filter (SABF). This module mainly consists of two crucial components: Scene-Aware Separation Threshold (SAST) and Bidirectional Adaptive Regulator (BAR). It can explicitly separate the target people from the background based on the sensitivity of the model towards instances and provide discriminative feature embeddings for the ReID task. Resolves the impairment of identity information caused by cross-domain scene disparities.

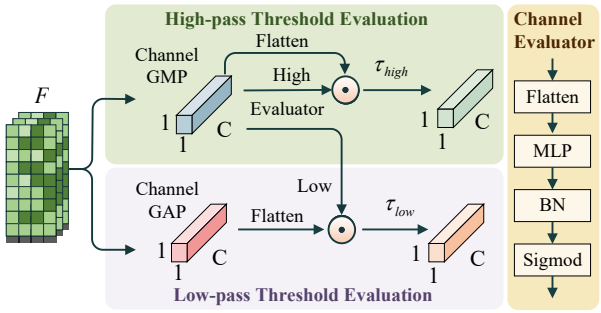


Figure 4: Details of our Scene-Aware Separation Threshold.

Scene-Aware Separation Threshold To achieve the functions of background elimination and instance enhancement, this paper proposes SAST to distinguish the information contained in the input features based on the instance sensitivity of the model trained on the source domain. As illustrated in Figure 4, SAST consists of two branches, which are respectively used to estimate the low-pass threshold for finding irrelevant backgrounds and the high-pass threshold for distinguishing reliable instances. To fulfill this objective, we use channel pooling to compress spatial information.

Through channel max pooling and average pooling operations, we compress the spatial information of feature $F \in \mathcal{R}^{H\times W\times C}$ from different approaches and then obtain two spatial context representations: $F_{max}^{c} \in \mathcal{R}^{1\times 1\times C}$ and $F_{avg}^{c} \in \mathcal{R}^{1\times 1\times C}$. Consequently, SAST respectively designates these two as the fundamental thresholds for the high-pass threshold and the low-pass threshold.

Regarding these two basic thresholds, SAST is guided by

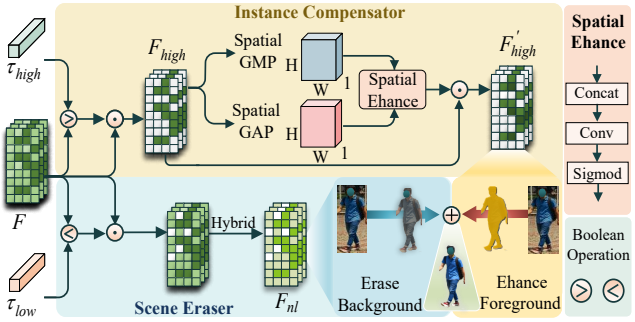


Figure 5: Details of our Bidirectional Adaptive Regulator.

F^c_{max}, which represents foreground instances. The intention is to leverage the characteristic that instance information is less sensitive to scene changes, thereby endowing the thresholds with enhanced scene adaptability. Specifically, we design a channel evaluator dedicated to processing F^c_{max}. For the high-pass threshold, the calculation process is as follows:

$$\begin{aligned} \alpha^c_{high} &= \sigma\left(BN\left(MLP_1\left(F^c_{max}\right)\right)\right), \\ \tau_{high} &= \alpha_{high} \odot F^c_{max}, \end{aligned} \tag{1}$$

where σ represents the sigmoid function and $\alpha^c_{high} \in \mathcal{R}^{1\times1\times C}$ is the channel adaptive scaling factor for the high-pass threshold, and $\tau_{high} \in \mathcal{R}^{1\times1\times C}$ is the high-pass threshold. For the low-pass threshold, the calculation process is given by:

$$\begin{aligned} \alpha^c_{low} &= \sigma\left(BN\left(MLP_2\left(F^c_{max}\right)\right)\right), \\ \tau_{low} &= \alpha_{low} \odot F^c_{avg}, \end{aligned} \tag{2}$$

where $\alpha^c_{low} \in \mathcal{R}^{1\times1\times C}$ is the channel adaptive scaling factor for the low-pass threshold and $\tau_{low} \in \mathcal{R}^{1\times1\times C}$ is the low-pass threshold.

Bidirectional Adaptive Regulator BAR separates the input features $F \in \mathcal{R}^{H\times W\times C}$ according to the threshold calculated by SAST and modulates the separated features, respectively. As illustrated in Figure 5, during this process, the Scene Eraser filters out the low-attention scene information in the $F \in \mathcal{R}^{H\times W\times C}$ according to the low-pass threshold $\tau_{low} \in \mathcal{R}^{1\times1\times C}$. Meanwhile, the Instance Compensator operates based on the high-pass threshold $\tau_{high} \in \mathcal{R}^{1\times1\times C}$. It compensates for the impairment to the feature representation of the human instance caused by the Scene Eraser and further enhances it.

Scene Eraser. The filtering of scene information hinges on an effective and appropriate function. However, traditional threshold functions (Donoho and Johnstone 1994; Donoho 1995), have limitations in practice. Specifically, the truncation in hard threshold functions disrupts feature space continuity, hindering the backbone network's feature extraction. The translation method of soft threshold functions, though smoother, induces a bias that restricts instance expression. To address these issues, this paper designs a hybrid piecewise function, which could be described as:

$$f(x,\tau) = \begin{cases} 0 & 0 \le x < \tau \\ \lambda\left(x^n - \tau^n + \epsilon\right)^{\frac{1}{n}} + (1-\lambda)(x-\tau) & x \ge \tau, \end{cases} \tag{3}$$

where x represents the elements of input feature $F \in \mathcal{R}^{H\times W\times C}$, τ refers to the low-pass threshold τ_{low}, $\lambda \in \mathcal{R}^{1\times1\times C}$ is a set of learnable parameters, which plays a role in smoothing processing. ϵ is a stabilizing decimal with a value of 1×10^{-6}. After filtering by the hybrid function $f(x,\tau)$, we obtain the feature without background called $F_{nl} \in \mathcal{R}^{H\times W\times C}$.

The function structure is similar to that of the soft threshold, ensuring the smoothness of processing. Meanwhile, it combines the advantages of the hard threshold. This can be described as follows:

$$\lim_{x\to+\infty} \frac{f(x,\tau)}{x} = 1. \tag{4}$$

As depicted in Equation 4, provided that the instance features extracted by the model are sufficiently prominent, the impairment inflicted by this function thereon will be minimal. Furthermore, since $f(x,\tau)$ is nonlinear, The derivative can be expressed as:

$$f'(x,\tau) \begin{cases} 0 & 0 \le x < \tau \\ 1 + \lambda\left(\left(x^n - \tau^n + \epsilon\right)^{\frac{1}{n}-1} x^{n-1} - 1\right) & x \ge \tau, \end{cases} \tag{5}$$

with the initial value of the parameter λ set to 0, the function $f(x,\tau)$ is equivalent to the soft threshold function during the initialization stage. Based on this, combined with the effect of the stabilizing decimal ϵ, it can mitigate the gradient explosion situation that may occur when x approaches τ^+. It provides a solid foundation for the stable training and effective operation of the model, enabling the model to perform more reliably in complex scene information filtering tasks.

Instance Compensator. Although Scene Eraser has achieved certain optimizations with the problem that impacts instance representation as Equation 4. The filtering operation of Scene Eraser on those instances with insufficiently salient representations might potentially undermine the model's perceptual ability.

To address this issue, this paper proposes an Instance Compensator to simultaneously enhance the foreground instances while the Scene Eraser is performing filtering, so as to ensure the model's effective acquisition and accurate perception of instance features. In detail, the Instance Compensator filters the input feature $F \in \mathcal{R}^{H\times W\times C}$ using a high-pass threshold τ_{high} and only retains potential instance features $F_{high} \in \mathcal{R}^{H\times W\times C}$. This can be expressed as:

$$F_{high} = \begin{cases} 0 & x < \tau_{high} \\ x & x \ge \tau_{high}, \end{cases} \tag{6}$$

where x represents the elements of the input feature $F \in \mathcal{R}^{H\times W\times C}$.Then, spatial pooling is used to compress the channel information to obtain $F^s_{max} \in \mathcal{R}^{H\times W\times1}$ and $F^s_{avg} \in \mathcal{R}^{H\times W\times1}$. Based on these, compensation is performed on $F_{high} \in \mathcal{R}^{H\times W\times C}$. This can be expressed as:

$$\begin{aligned} \alpha^s &= \sigma\left(Conv\left(Concat\left(F^s_{max}; F^s_{avg}\right)\right)\right), \\ F'_{high} &= \alpha^s \odot F_{high}, \end{aligned} \tag{7}$$

where σ represents the sigmoid function and $\alpha^s \in \mathcal{R}^{H\times W\times1}$ is the spatial adaptive scaling factor and $F'_{high} \in \mathcal{R}^{H\times W\times C}$ is the compensation of instance.

Algorithm 1: Offline Update

Input:

Target domain dataset, $D_t = \{G_1, G_2, \ldots\}$
Offline memory bank, M_{off}
Person search model, PM
Smoothing factor, γ

Output:

Update the elements of the offline memory bank M_{off}

IOU threshold, τ = 0.7
for *each* $G_i \in D_t$ **do**
  $instances,\ boxes = PM.eval(G_i)$
  Set $Pairs \leftarrow []$
  for $instance,\ box \in (instances,\ boxes)$ **do**
    for $x_i^j,\ b_i^j \in M_{off}[\ i\ ]$ **do**
      if IOU(box, x_i^j) $\geq \tau$ **then**
        $instance = (1-\gamma)\, instance + \gamma\, x_i^j$
        $box = (1-\gamma)\, box + \gamma\, b_i^j$
        Append $[\ instance\ , box\ ]$ in $Pairs$
      end if
    end for
  end for
  $M_{off}[\ i\ ] = Pairs$
end for
return M_{off}

Finally, BAR adds the results of Scene Eraser and Instance Compensator to obtain the final output result $F_{out} \in \mathcal{R}^{H\times W\times C}$ as follows:

$$F_{out} = F_{nl} + F'_{high}. \tag{8}$$

Instance Consistency Contrastive Learning

Through leveraging the bidirectional modulation function of the SABF module, the misleading identity learning caused by cross-scene alignment can be effectively eliminated. However, during the cross-domain joint training stage, since the pseudo-IDs of the target domain are assigned through clustering, there is a problem that the instance identity may change in different epochs, which poses a challenge to the model's cross-epoch identity learning.

In response, this paper proposed the Instance Consistency Contrastive Learning (ICCL) strategy. ICCL designs two distinct types of memory banks to integrate the same instances from different epochs to generate more stable pseudo-labels. Thus, the model's learning of the same target is always consistent, ensuring the stability of training.

Target Domain Offline Processing. ICCL utilizes an offline memory bank and supplies it with a dedicated instance update strategy. Let $M_{off} = \{(x_i^j, b_i^j) \mid (i,j) \in \{(1,1), \cdots, (N,k)\}\}$ denotes the state of the offline memory bank prior to the start of the t-th epoch. In this context, x_i^j represents the j-th instance within the i-th image of the target domain dataset, and b_i^j is the corresponding bounding box. Additionally, k represents the number of instances in image i, and N represents the number of images. We summarize the update of M_{off} in Algorithm 1.

After the update operation of the M_{off} is completed, the ICCL algorithm will utilize the DBSCAN (Ester et al. 1996) algorithm to perform clustering processing on the instances stored in the offline memory bank and then a series of pseudo-IDs. Subsequently, these pseudo-IDs are combined with the already updated bounding boxes to re-annotate the target domain dataset. Meanwhile, based on the assignment results of these pseudo-IDs, the corresponding cluster centroids in the offline memory bank are calculated. This process can be described as follows:

$$c_k = \frac{1}{N_k} \sum_{f_i \in F_k} f_i, \tag{9}$$

where c_k denotes the cluster centroids for the k-th cluster, F_k represents a set of instances with the same pseudo-IDs, $N_k$is the size of this set, and f_i is the feature of each instance. After calculating the clustering centers corresponding to all the pseudo-IDs, we re-initialize the target domain online memory bank $M_{on}^t = \{c_1^t, c_2^t, ..., c_n^t\}$. Note that the source domain also calculates centers, but it is based on its GT annotations, and its online memory bank $M_{on}^s = \{c_1^s, c_2^s, ..., c_n^s\}$ just initialized once at the source domain pre-training stage, and no initialization will occur in other phases.

Cross-Domain Online Training. ICCL regards both M_{on}^s and M_{on}^t as a unified online memory bank $M_{on} = \{c_1, c_2, ..., c_K\}$ to conduct the training of the ReID task. The objective function of it can be expressed as follows:

$$L = -\log \frac{\exp\left(f \cdot c_+/\tau\right)}{\sum_{k=0}^{K} \exp\left(f \cdot c_k/\tau\right)}, \tag{10}$$

in this formula, f represents the instance feature extracted for the ReID task, and c_+ is the center of the cluster or class to which f belongs. τ as a temperature factor, aims to accelerate the convergence of the model. During backward, if the target label is k, then M_{on} will update the k-th column by:

$$c_k \longleftarrow (1-\alpha)\, c_k + \alpha f, \tag{11}$$

where $\alpha \in [0,1]$ is the momentum factor used to control the update speed of the M_{on}.

Experiments

Experimental Settings

Datasets. We evaluate our IGSA on two benchmark 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, consisting of 18,184 images. It contains 8,432 distinct person identities and 96,143 annotated bounding boxes. The PRW dataset contains 11,816 scene images with annotations for 932 different person identities and 43,110 bounding boxes. Its training set is composed of 483 identities and 5,704 images, and the test set contains 2,057 query persons and 6,112 scene images.

Implementation Details. We implement the IGSA using PyTorch (Paszke et al. 2019) and train it on an NVIDIA A800 GPU with a batch size of 4. We use the Stochastic Gradient Descent (SGD) optimizer and set the learning rate to 2.4×10^{-3}. The learning rate is warmed up in the first

Method	Venue	Backbone	PRW		CUHK-SYSU	
			mAP	top-1	mAP	top-1
Fully-Supervised						
OIM (Xiao et al. 2017)	CVPR2017	ResNet-50	21.3	49.4	75.5	78.7
MGTS(Chen et al. 2018)	ECCV2018	VGG-16	32.6	72.1	83.0	83.7
RDLR (Han et al. 2019)	ICCV2019	ResNet-50	42.9	70.2	93.0	94.2
NAE+ (Chen et al. 2020)	CVPR2020	ResNet-50	44.0	81.1	92.1	92.9
SeqNet (Li and Miao 2021)	AAAI2021	ResNet-50	46.7	83.4	93.8	94.6
PSTR (Cao et al. 2022)	CVPR2022	PVTv2-B2	56.5	89.7	95.2	96.2
SeqNeXt (Jaffe and Zakhor 2023)	WACV2023	ConvNeXt-B	57.6	89.5	96.1	96.5
SEAS(Jiang et al. 2024)	IJCAI2024	ConvNeXt-B	60.5	89.5	97.1	97.8
Weakly-Supervised						
CGPS(Yan et al. 2022)	AAAI2022	ResNet-50	16.2	68.0	80.0	82.3
SSL(Wang et al. 2023)	ICCV2023	ResNet-50	33.9	82.7	87.6	89.0
DICL(Wang et al. 2024)	PR2024	ResNet-50	35.5	80.9	87.4	88.8
Unsupervised						
DAPS(Li et al. 2022)	ECCV2022	ResNet-50	34.7	80.6	77.6	79.6
FOUS(Cui et al. 2024)	IJCAI2024	ResNet-50	35.4	80.8	78.7	80.5
DDAM(Almansoori, Fiaz, and Cholakkal 2024)	WACV2024	ResNet-50	36.7	81.2	79.5	81.3
Mos(Kim, Kim, and Sohn 2025)	CVPR2025	ResNet-50	37.1	81.9	80.1	81.5
IGSA(Ours)		ResNet-50	**41.1**	**82.3**	**82.1**	**83.8**

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. Pre-trained ResNet50 (He et al. 2016) serves as the model's backbone. During training, input images are resized to 1500×900 and randomly flipped horizontally for data augmentation. For Equation 5, we set n as 2. For the online memory bank update, the momentum factor is 0.2. Considering the distribution characteristics of the CUHK-SYSU and PRW datasets, we initialize the smoothing factors γ as 0.6 for CUHK-SYSU and 0.3 for PRW. To prevent over-fitting, the number of training epochs is flexibly adjusted based on the target domain dataset. When PRW is the target domain, the IGSA has 8 pre-training epochs on CUHK-SYSU and then 12 joint training epochs. Conversely, when CUHK-SYSU is the target domain, it has 3 pre-training epochs on PRW followed by 7 joint training epochs.

Comparison with State-of-the-Art Methods

To demonstrate the effectiveness of the IGSA, we compare the IGSA with SOTA methods that adopt various training strategies, covering supervised, weakly supervised, and unsupervised methods. As shown in Table 1, the proposed IGSA exhibits outstanding performance on both benchmark datasets. Specifically, IGSA reaches 82.1% mAP on the CUHK-SYSU dataset and 41.1% mAP on the PRW dataset, outperforming the runner-up by 2.6% and 4.4%, respectively. Moreover, as shown in Figure 7, a comparison between our IGSA and other SOTA UDA methods is conducted on the CUHK-SYSU test set with gallery sizes varying from 50 to 4,000. More distracting persons in the gallery set pose a challenge for all compared methods. However, our IGSA consistently outperforms all the UDA methods. From the qualitative analysis perspective, Figure 6 further displays the advantages of IGSA under complex and challenging conditions such as cross-scene transitions, low-resolution images, object occlusions, and viewpoint changes. It highlights its strong adaptability in complex environments.

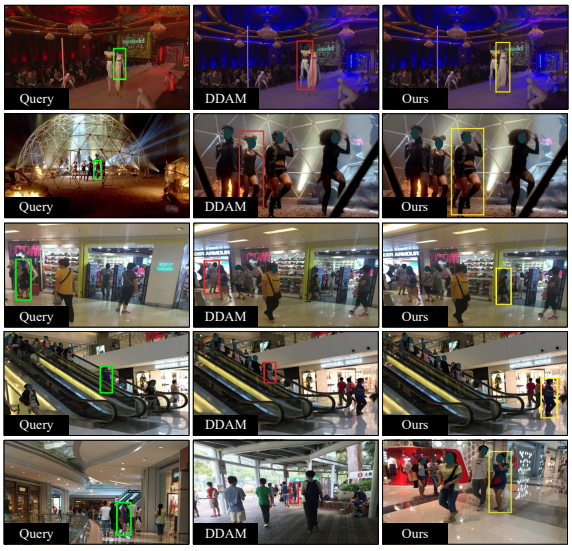


Figure 6: Qualitative comparison of IGSA with DDAM on the CUHK-SYSU test set.

Ablation Study

Effectiveness of Scene-Aware Bidirectional Filter. Our proposed SABF aims to effectively eliminate the interference of scene information and enable the model to focus on foreground instances. It mainly consists of two key components: Scene Eraser (SE) and Instance Compensator (IC).

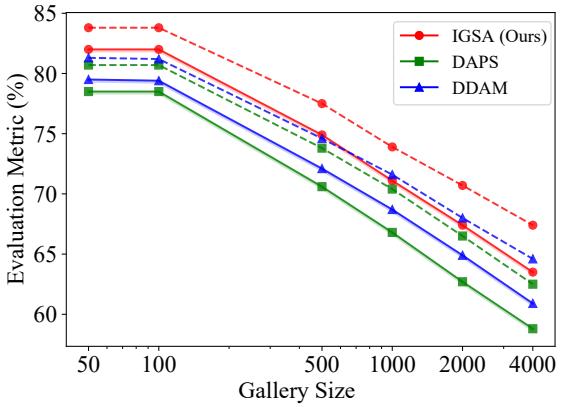


Figure 7: Comparison of performance on CUHK-SYSU across various gallery sizes. The x-axis uses a logarithmic scale. The solid lines represent mAP(%) and the dashed lines represent top-1 accuracy(%).

Components			PRW		CUHK-SYSU	
SE	IC	ICCL	mAP	top-1	mAP	top-1
✗			37.2	80.7	78.0	79.6
	✗		40.7	81.4	81.2	82.7
✗	✗		36.2	81.1	79.1	80.9
		✗	40.2	81.7	80.8	83.0
✗	✗	✗	34,7	80.6	77.6	79.6
			41.1	**82.3**	**82.1**	**83.8**

Table 2: Validating the effectiveness of different components on the CUHK-SYSU and PRW datasets. SE: Scene Eraser. IC: Instance Compensator. ICCL: Instance Consistency Contrastive Learning.

To comprehensively validate the effectiveness of SABF, we meticulously designed and carried out a series of experiments: (i) without SE; (ii) without IC; and (iii) without both SE and IC simultaneously. As shown in #1 and #2 of Table 2, whether SE or IC is removed, the model performance on both datasets will significantly decline. This result confirms that the bidirectional modulation plays a crucial role in ensuring the accuracy of the UDA person search. When both SE and IC are removed, as indicated in #3, the performance additionally drops, which further demonstrates the effectiveness of our designed SABF architecture.

To validate the effectiveness of the thresholds on which SABF performs bidirectional modulation, we conducted additional ablation experiments on the Scene-Aware Separation Threshold (SAST). Specifically, different pooling strategies were used as input for SAST. As the results presented in Table 3 show, the max pooling strategy we applied consistently outperforms other pooling schemes. This demonstrates that the foreground instance extracted by max pooling has a more consistent representation across different scenes, ensuring the reliability of the threshold.

Impact of Instance Consistency Contrastive Learning. ICCL is another essential component of our method, focusing on ensuring the consistency of the training process by generating stable pseudo-labels. As shown in Table 2 #4, removing ICCL reduces both mAP and top-1 scores on the PRW and CUHK-SYSU datasets. Moreover, as shown in Table 4, the combination of online and offline updates significantly improves model performance, which confirms the effectiveness of our proposed method.

Pooling Strategy		PRW		CUHK-SYSU	
α^c_{low}	α^c_{high}	mAP	top-1	mAP	top-1
GAP&GMP	GAP&GMP	39.1	81.4	81.1	82.8
GAP	GMP	39.0	81.1	79.5	80.8
GAP	GAP	40.1	82.3	78.1	80.1
GMP	GAP	39.8	81.0	81.0	82.6
GMP	GMP	**41.1**	**82.3**	**82.1**	**83.8**

Table 3: Results of the different pooling strategies for the channel adaptive scaling factor.

ICCL	PRW		CUHK-SYSU	
	mAP	top-1	mAP	top-1
w/o Offline Update	40.0	81.7	79.8	81.2
w/o Online Update	40.3	**82.6**	81.2	82.7
$w/$ Online & Offline	**41.1**	82.3	**82.1**	**83.8**

Table 4: Comparative results by employing different update strategies for ICCL.

n	PRW		CUHK-SYSU	
	mAP	top-1	mAP	top-1
1	39.1	81.6	80.6	82.3
2	**41.1**	**82.3**	**82.1**	**83.8**
3	38.6	81.6	80.2	81.8

Table 5: Performance on the benchmark CUHK-SYSU and PRW datasets with different n of Scene Eraser.

Analysis on Hyper-Parameter. For the hyper-parameter n in Equation 3. As shown in Table 5, setting $n = 2$ leads to the optimal mAP and top-1 in tests on both datasets, outperforming the soft threshold function (i.e., when $n = 1$). This verifies that the hybrid piecewise function is more suitable for the UDA person search task than the traditional threshold, effectively meeting its complex cross-domain scene adaptation and enhancing the model's performance.

Conclusion

This paper proposes an Instance-Guided Scene Adaptation framework to achieve unsupervised cross-domain scene adaptation within an end-to-end manner. By separating the person instances and scene information, IGSA enhances the features from foreground instances while filtering the background noises simultaneously, which effectively overcomes cross-domain scene disparities for UDA person search. A large number of experiments have verified the superior performance of our method, further narrowing the gap between supervised and unsupervised person search.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China Grant 2024YFB4710800 and 2022YFE0132600, National Natural Science Foundation of China Grant 62576067, 62501226 and 62176037, Liaoning Provincial Natural Science Foundation Grant 2025-YQ-01 and 2024-MS-012, Natural Science Foundation of Hebei Province F2025201037, Dalian Science and Technology Talent Innovation Support Plan Grant 2024RY010.

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 Tai, Y. 2018. 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. 2020. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12615–12624.

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

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.

Feng, L.; Wang, H.; Jin, B.; Li, H.; Xue, M.; and Wang, L. 2018. Learning a distance metric by balancing kl-divergence for imbalanced datasets. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(12): 2384–2395.

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

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.; 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.; 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.

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, 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.

Kim, M.; Kim, S.; and Sohn, K. 2025. Mixture of Submodules for Domain Adaptive Person Search. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13990–14001.

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.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Peng, J.; Jiang, G.; and Wang, H. 2023. Adaptive memorization with group labels for unsupervised person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10): 5802–5813.

Peng, J.; Zhang, S.; and Wang, H. 2025. CDE-Learning: Camera Deviation Elimination Learning for Unsupervised Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6452–6460.

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.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

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

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.

Yao, M.; Wang, H.; Chen, Y.; and Fu, X. 2024. Between/within view information completing for tensorial incomplete multi-view clustering. *IEEE transactions on multimedia*.

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*.

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.