Localization-Anchored Instance Discrimination for Domain Adaptive Person Search

Linfeng Qi[1], **Huibing Wang**[1*], **Jinjia Peng**[2], **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, jqz}@dlmu.edu.cn, pengjinjia@hbu.edu.cn

Abstract

Domain-adaptive person search (DAPS) aims to transfer pedestrian detection and re-identification capabilities from a labeled source domain to an unlabeled target domain, yet faces critical challenges from domain shift: semantic confusion among overlapping instances, over-reliance on shallow features for look-alike targets, and poor discriminability of small-scale instances. To address these issues, we propose the Localization-Anchored Instance Discrimination (LAID) framework, which leverages spatial relationships between bounding boxes as auxiliary signals to enhance instance identity learning. LAID integrates three complementary strategies: 1) Cost-Aware Instance Matching (CAIM) uses IoU-based global optimal assignment to align current detections with historical identities, reducing overlap-induced misassociations; 2) Dual-Scope Contrastive Learning (DSCL) combines spatial separation constraints (for geometrically distant pairs) with global contrastive learning, prompting the model to learn deep discriminative features beyond superficial similarities; 3) Task-Sensitivity Alignment (TSA) aligns confidence distributions of detection and ReID heads via KL divergence, ensuring consistent pseudo-label generation. Extensive experiments on CUHK-SYSU and PRW datasets demonstrate that LAID outperforms state-of-the-art DAPS methods, validating its effectiveness in mitigating domain shift and narrowing the performance gap between supervised and domain-adaptive person search.

Code — https://github.com/whbdmu/LAID

Introduction

Person search is developed to localize and identify specific pedestrians in real-scene images, with its core challenge lying in simultaneously handling the joint tasks of object 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). Existing methods mostly follow a typical supervised learning paradigm (Li and Miao 2021; Jaffe and Zakhor 2023; Jiang et al. 2024), relying on precisely annotated bounding boxes and reliable identity labels for training. While supervised methods have made significant progress in this field, with performance on some datasets approaching or even surpassing human-level accuracy, their limitations have become increasingly prominent. Such models often overfit to domain-specific information in training data (e.g., scene-specific lighting, background distributions, camera styles), leading to substantial performance degradation in unseen new domains. This restricts their direct application in real-world scenarios where annotation costs are unaffordable or scenes change dynamically.

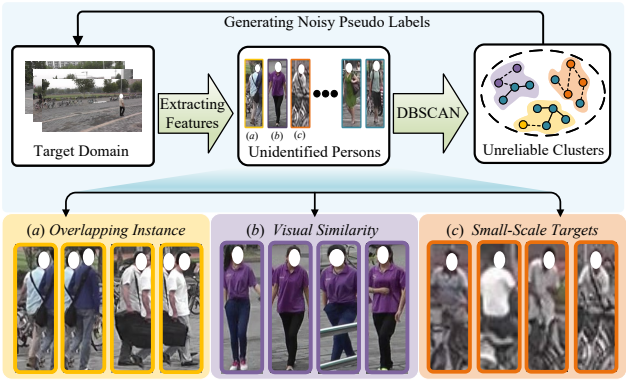


Figure 1: Typical scenario example: Domain shift leads to unreliable pseudo-labels. (a) Semantic confusion of overlapping instances; (b) Visual similarity instances are dominated by shallow features; (c) Insufficient discriminability for small targets.

To enhance the cross-domain generalization ability of person search models, (Li et al. 2022) pioneered the task setting of Domain-Adaptive Person Search (DAPS) and designed a framework of the same name. This method employs a domain alignment mechanism fusing multi-scale features to guide the model in learning domain-invariant representations, laying the foundation for cross-domain transfer. Building on this, (Almansoori, Fiaz, and Cholakkal 2024) and (Kim, Kim, and Sohn 2025) further addressed the optimization conflict in domain adaptation tasks, proposing to generate mixed-domain features to enhance the model's domain adaptability. Furthermore, (Qi et al. 2025) defined domain shift in person search as a form of scene adaptation problem, improving the model's generalization performance in new scenes by mitigating underlying scene discrepancies.

*Corresponding Author.

While advancing DAPS, existing methods face notable challenges in generating reliable pseudo-labels and learning robust instance features under domain shift. Their reliance on clustering for pseudo-label initialization inherently renders them sensitive to feature distribution distortions induced by domain gaps, resulting in ambiguous cluster boundaries (Yao et al. 2024). Furthermore, prevalent local matching strategies for temporal smoothing often produce suboptimal or erroneous links for highly overlapping instances. Critically, these approaches lack explicit mechanisms to leverage spatial context for disambiguation or to enforce consistency between the core detection and ReID tasks. Consequently, several failure modes persist, as shown in Fig. 1: Semantic confusion among overlapping instances: Spatially proximate but distinct identities are prone to incorrect linking or shared pseudo-labels, particularly under occlusion or for small targets; Dominance of shallow features: Models over rely on easily transferable yet non-discriminative superficial cues (e.g., color blocks) when processing visually similar distractors, failing to acquire robust identity semantics; Exacerbated challenges for small targets: Low resolution and detail loss in small targets compound both overlap-induced confusion and over reliance on shallow features. It is worth noting that these limitations often do not exist in isolation. Instead, they collectively degrade the quality of self-supervised signals, creating a cascading effect that hinders effective domain adaptation.

To overcome these challenges, we propose a novel Localization-Anchored Instance Discrimination (LAID) framework. LAID uniquely exploits the geometric properties of bounding boxes, a readily available and domain-agnostic signal, to anchor and enhance instance discrimination learning in the unlabeled target domain. It comprises three synergistic components: Cost-Aware Instance Matching (CAIM): Replaces error-prone local matching with a global optimal assignment based on IoU cost matrix. This provides a more accurate and stable association between current detections and historical identity proposals, directly tackling semantic confusion arising from spatial overlaps. Dual Scope Contrastive Learning (DSCL): Combines intra-scene spatial separation constraints with global contrastive learning. Specifically, it identifies high-confidence negative pairs as instances within the same image with extremely low IoU, leveraging the strong prior that they must represent different identities. Enforcing dissimilarity for these pairs, alongside standard global instance discrimination, compels the model to discover deep, robust semantics beyond misleading visual similarities. Task-Sensitivity Alignment (TSA): Addresses the misalignment between Region Proposal Network (RPN) objectness scores and ReID confidence scores. TSA minimizes the KL divergence between their confidence distributions for the same instance (Feng et al. 2018), fostering task-consistent pseudo-label selection. By anchoring on spatial localization and ensuring task consistency, LAID generates more reliable pseudo-labels and learns significantly more discriminative features for the unlabeled target domain. Extensive ablation studies on benchmark datasets validate the effectiveness of each component. Overall, experiments demonstrate that LAID achieves better domain generalization and outperforms state-of-the-art (SOTA) methods in DAPS settings, particularly in challenging scenarios with heavy occlusion, small or similar targets.

Our contributions can be summarized below:

- We design global matching via optimal transport with IoU-based costs. It solves overlap-induced identity confusion by ensuring consistent cross-epoch associations.
- Our method unifies geometric priors with contrastive learning. This compels deep feature mining beyond superficial similarities.
- We bridge detection-ReID confidence gaps via KL-divergence minimization. It enhances pseudo-label quality through task-consistent regularization.
- To the best of our knowledge, our localization-anchored approach is the first to exploit spatial relationships for DAPS. It anchors instance discrimination on bbox geometry, achieving SOTA performance.

Methodology

Framework Overview

The training framework of Localization-Anchored Instance Discrimination (LAID), illustrated in Fig. 2, adopts an iterative two-stage process per training epoch to address domain shift in person search. Given a labeled source domain $\mathcal{D}_s = \{(x_n, y_n)\}_{n=1}^{N}$ and unlabeled target domain $\mathcal{D}_t = \{x_m\}_{m=1}^{M}$, LAID progressively refines target pseudo-labels while enhancing feature discriminability through:

Stage 1: Pseudo-label Generation. The model infers on $\mathcal{D}_t$ to generate candidate bounding boxes and extracts instance features. Cost-Aware Instance Matching (CAIM) aligns current detections with historical proposals via global optimal assignment, followed by DBSCAN clustering to generate pseudo-identity labels.

Stage 2: Multi-constraint Training. Dual-Scope Contrastive Learning (DSCL) enhances discriminability through: (a) global contrastive loss with cluster-based pseudo-labels; (b) spatial separation loss for low-overlap in-scene pairs. Concurrently, Task-Sensitivity Alignment (TSA) enforces consistency between detection and ReID confidence distributions via KL divergence.

Cost-Aware Instance Matching

The reliability of pseudo-label generation in domain-adaptive person search fundamentally depends on accurate bounding box localization. Existing approaches typically employ temporal smoothing techniques that rely on greedy local matching strategies, associating each current detection with its maximally overlapping historical proposal. While computationally efficient, this local optimization strategy risks accumulating identity-binding errors in complex scenes characterized by multiple overlapping instances. Consider a typical failure case: a current detection of pedestrian A may be incorrectly associated with historical proposal B due to transient spatial proximity, even when a more consistent match exists with its true historical counterpart C. Such misassociations propagate through training

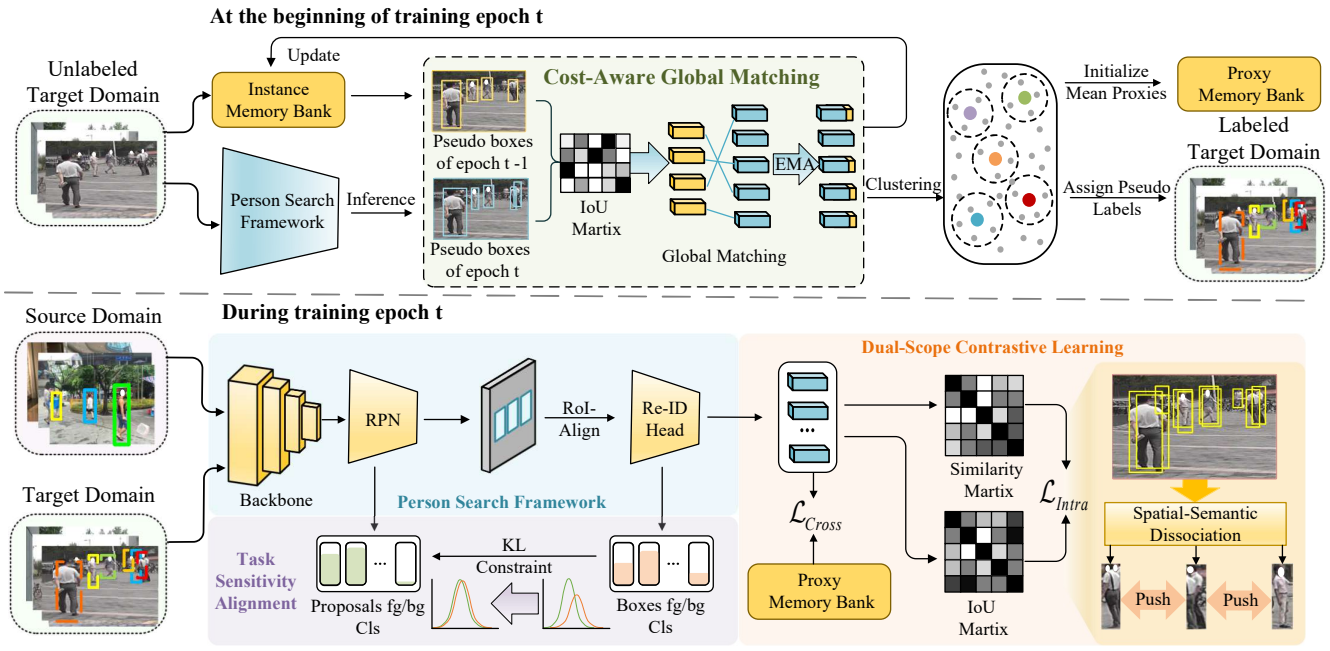


Figure 2: The training framework of Localization-Anchored Instance Discrimination (LAID). This framework adopts a two-stage training process per epoch. Stage 1 generates pseudo-labels via CAIM-based temporal alignment and clustering; Stage 2 enhances feature discriminability through DSCL and TSA.

epochs, progressively degrading pseudo-label quality and ultimately undermining model performance.

To address this critical limitation, we propose Cost-Aware Instance Matching (CAIM), a robust association strategy grounded in optimal transport theory. CAIM establishes reliable correspondence between current detections and historical proposals through a carefully designed pipeline that prioritizes global consistency over local optima. For each target domain image $x_m \in \mathbb{R}^{H\times W\times 3}$ at the training epoch t, we begin with two key inputs: the set of current detections $\mathcal{B}_t = \{b_i^t \in \mathbb{R}^4\}_{i=1}^N$ with corresponding instance features $\mathcal{F}_t = \{f_i^t \in \mathbb{R}^d\}_{i=1}^N$, and the historical proposals $\mathcal{H}_m^{t-1} = (\mathcal{B}_{t-1}, \mathcal{F}_{t-1})$ cached from the previous epoch $t-1$.

The CAIM process initiates with an inclusive candidate selection phase. To avoid prematurely eliminating potentially valuable matches, we employ a relaxed matching threshold $\epsilon_{\text{loose}} = \alpha \cdot \epsilon$, where ϵ represents the standard strict threshold and $\alpha = 0.95$ functions as the relaxation factor. This balanced strategy retains matching flexibility while upholding reasonable quality constraints, proving particularly advantageous for partially occluded instances or those with low confidence. These cases might otherwise be discarded but hold significance for maintaining identity continuity across epochs.

At the core of CAIM lies the global optimal assignment mechanism. We formulate the matching problem as a bipartite graph optimization where the cost matrix $\mathbf{C} \in \mathbb{R}^{|\mathcal{B}_{t-1}|\times|\mathcal{B}_t|}$ is defined by the spatial dissimilarity between proposals: $c_{ij} = 1 - \text{IoU}(b_i^{t-1}, b_j^t)$. To ensure meaningful associations, we impose a validity constraint requiring $\text{IoU}(b_i^{t-1}, b_j^t) > 0.7$, a threshold empirically determined to balance matching precision and recall. The optimal assignment $\mathcal{M}^*$ is then solved using the Jonker-Volgenant (Jonker and Volgenant 1983) algorithm, which minimizes the total matching cost while enforcing one-to-one correspondences:

$$\mathcal{M}^* = \arg\min_{\mathcal{M}} \sum_{(i,j)\in\mathcal{M}} c_{ij}, \quad s.t.\ \mathcal{M} \subseteq \mathcal{B}_{t-1} \times \mathcal{B}_t\ . \quad (1)$$

This global optimization strategy avoids the myopic decision-making inherent in greedy approaches. Instead, it considers the full configuration of proposals to identify the most coherent set of associations, thereby mitigating errors caused by local overlap biases.

Following optimal assignment, we enforce temporal consistency through exponential smoothing. For each matched pair $(i, j) \in \mathcal{M}^*$, the bounding box and feature representations are updated as:

$$\begin{cases} \tilde{b}_j^t = \lambda b_j^{t-1} + (1-\lambda) b_i^t \\ f_j^t = \lambda f_j^{t-1} + (1-\lambda) f_i^t\ , \end{cases} \quad (2)$$

where the momentum coefficient λ controls historical information retention, a value calibrated to balance stability against responsiveness to new observations. For detections without historical correspondences yet exhibiting high confidence (conf $> \epsilon$), they will be initialized in the updated proposal cache $\mathcal{H}_m^t$, enabling the model to adapt to emerging identities as follows:

$$\mathcal{H}_m^t \leftarrow \mathcal{H}_m^t \cup \left\{(b_t^i, f_t^i)\right\}. \quad (3)$$

The final stage transforms geometrically aligned features into semantic pseudo-labels. The smoothed feature set $\tilde{\mathcal{F}}_t$ serves as input to DBSCAN (Ester et al. 1996) clustering, which automatically discovers identity clusters while filtering noise. We configure the clustering with $\epsilon_{\text{DBSCAN}} = 0.5$ and min_samples $= 4$. The resulting cluster assignments yield pseudo-identity labels $\hat{y}_i$, while cluster centroids initialize and continuously update the ReID memory bank, creating a self-reinforcing cycle where improved features enable better clustering, which in turn enhances feature learning. CAIM reduces identity misassociation caused by bounding box overlaps through global optimal matching and historical feature smoothing. This provides more stable pseudo-labels for subsequent ReID tasks.

Dual-Scope Contrastive Learning

Domain-adaptive person search methods commonly employ cluster-based global contrastive learning to enhance instance discriminability. This approach constructs a memory bank containing identity proxies and utilizes variants of the InfoNCE loss, typically formulated as:

$$\mathcal{L}_{cross} = -\log \frac{\exp(f \cdot c_+/\tau)}{\sum_{i=1}^{N} \exp(f \cdot c_i^s/\tau) + \sum_{j=1}^{M} \exp(f \cdot c_j^t/\tau)} , \quad (4)$$

where f represents the embedding of a detected instance, c_+ its corresponding proxy, and c_i^s, c_j^t denote source and target domain proxies, respectively. While theoretically valid, this global contrastive learning strategy encounters notable limitations under domain shift: the model tends to over-rely on superficial visual similarities (e.g., dominant color blocks) rather than learning deep discriminative features, resulting in clustering outcomes that propagate and amplify initial pseudo-label errors across training iterations.

To address this fundamental challenge, LAID introduces a geometric-aware intra-scene constraint that leverages spatial separation as a domain-invariant supervisory signal. This approach begins with the observation that detection boxes with minimal spatial overlap must represent distinct identities, a reliable prior unaffected by domain discrepancies. For each input image containing N candidate detections, we first identify foreground pedestrian boxes $\mathcal{B} = \{b_i | y_i = 1\}_{i=1}^{M}$ using the RPN's classification output. Crucially, we establish group assignments $\mathbf{g} \in \mathbb{Z}^M$ through geometric clustering based solely on bounding box relationships, implemented via Algo 1 with a strict IoU threshold θ. This process satisfies the condition:

$$\mathbf{g}_i = \mathbf{g}_j, \quad s.t.\ \text{IoU}(b_i, b_j) > \theta , \quad (5)$$

effectively creating identity-agnostic groups where boxes with IoU $< \theta_{\text{group}}$ constitute high-confidence negative pairs.

Building on this geometric foundation, we design a spatial separation loss that enforces feature dissimilarity between distinct groups. For normalized embeddings $\mathbf{E} = [e_1, \ldots, e_M]^\top$ where $\|e_i\|_2 = 1$, the pairwise similarity matrix is:

$$\mathbf{S} = \mathbf{E}\mathbf{E}^\top, \quad S_{ij} = e_i^\top e_j . \quad (6)$$

Algorithm 1: Union-Find Grouping of Detection Boxes

Require: $\mathcal{B} = \{b_n\}_{n=1}^{N}$: bounding boxes ($b_n \in \mathbb{R}^4$)
$\theta \in [0, 1]$: IoU threshold
Ensure: $\mathbf{g} \in \mathbb{Z}^N$: group labels
Compute IoU matrix: $\mathrm{I} \in \mathbb{R}^{N \times N}$
Initialize Union-Find structure $\mathcal{U}$ with N elements
for $i = 1$ **to** $N - 1$ **do**
  for $j = i + 1$ **to** N **do**
    if $\mathrm{I}_{i,j} > \theta$ **then**
      $\mathcal{U}$.union(i, j)
    end if
  end for
end for
Assign group IDs to components in $\mathcal{U}$ as $\mathbf{g}$
return $\mathbf{g}$

The loss function then targets inter-group pairs $\Omega = \{(i, j) | \mathbf{g}_i \neq \mathbf{g}_j, i \neq j\}$:

$$\mathcal{L}_{\text{intra}} = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} w_{ij} \cdot \max(S_{ij} - \gamma, 0) , \quad (7)$$

with γ defining the separation margin. The adaptive weights w_{ij} focus learning on challenging cases through a temperature-controlled mechanism:

$$w_{ij} = \frac{\exp(\max(S_{ij} - \gamma, 0)/\tau)}{\sum_{(k,l)\in\Omega} \exp(\max(S_{kl} - \gamma, 0)/\tau)} , \quad (8)$$

where τ controls the hardness emphasis. This formulation ensures three critical properties: 1) Geometric grouping guarantees separation constraints apply only to truly distinct pedestrians; 2) The $\max(S_{ij} - \gamma, 0)$ term forces the model to explore semantics beyond superficial similarities; 3) Adaptive weighting prioritizes semantically ambiguous pairs that are visually similar yet spatially separated.

The complete Dual-Scope Contrastive Learning objective synergistically combines global and local constraints:

$$\mathcal{L}_{\text{DSCL}} = \mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{intra}} . \quad (9)$$

This integrated approach provides noise resilience against pseudo-label errors while compelling the model to discover robust identity semantics. Crucially, the spatial separation component remains valid even when clustering fails, creating a self-correcting mechanism that progressively improves feature discriminability throughout adaptation.

Task-Sensitivity Alignment

To enhance pseudo-label reliability in joint detection-ReID frameworks, we introduce a Task-Sensitive Alignment Loss that aligns confidence predictions between the RPN and ReID head. Specifically, in domain-adaptive person search, pseudo-label generation typically relies on confidence scores of detected bounding boxes. When detection and ReID tasks produce inconsistent confidence estimates for the same instance (e.g., high confidence from detection but low from ReID), the resulting pseudo-labels tend to be unreliable. By minimizing the KL divergence between the

Category	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
	HOIM(Chen et al. 2020a)	AAAI2020	ResNet-50	39.8	80.4	89.7	90.8
	NAE+ (Chen et al. 2020b)	CVPR2020	ResNet-50	44.0	81.1	92.1	92.9
	TCTS(Wang et al. 2020)	CVPR2020	ResNet-50	46.8	87.5	93.9	95.1
	AlignPS+ (Yan et al. 2021)	CVPR 2021	ResNet-50	46.1	82.1	94.0	94.5
	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
	R-SiamNet(Han et al. 2021)	ICCV2021	ResNet-50	21.4	75.2	86.0	87.1
	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
	OLA(Zhu, Yang, and Wang 2025)	AAAI2025	ResNet-50	38.1	82.0	87.8	89.3
Domain-Adaptation	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
	DSCA (Qi et al. 2025)	AAAI2025	ResNet-50	<u>39.9</u>	<u>81.6</u>	<u>80.2</u>	<u>81.7</u>
	LAID (Ours)	-	ResNet-50	**40.5**	**81.9**	**80.8**	**82.6**

Table 1: Comparison of mAP(%) and top-1 accuracy(%) with fully supervised, weakly supervised, and domain adaptation methods on the CUHK-SYSU and PRW test sets. The best and second-best results in domain adaptation methods are shown in bold and underlined, respectively.

two tasks' output probability distributions, we encourage their confidence estimates for the same instance to converge, thereby improving pseudo-label reliability. This loss is defined as follows:

$$\mathcal{L}_{TSA} = D_{KL}(\mathrm{Q}||\mathrm{P}) \,, \tag{10}$$

where $\mathrm{P} \in \mathbb{R}^{N\times 2}$ represents the foreground/background probability distribution from RPN classification scores, $\mathrm{Q} \in \mathbb{R}^{N\times 2}$ is the binary probability distribution from ReID confidence scores.

Experiments

Experimental Settings

Datasets. We evaluate LAID on two standard person search benchmarks:

- **CUHK-SYSU** (Xiao et al. 2017): Contains 18,184 images with 96,143 annotated bounding boxes and 8,432 distinct identities. The training set includes 5,532 identities across 11,206 images, while the test set comprises 6,978 gallery images with 2,900 queries.
- **PRW** (Zheng et al. 2017): Consists of 11,816 video frames with 43,110 bounding boxes and 932 identities. Its training set has 5,704 images covering all identities; the test set has 6,112 gallery images with 2,057 queries.

Metrics. We employ two standard metrics for quantitative evaluation of person search performance: mean Average Precision (mAP) and top-1 accuracy (top-1). All evaluations are conducted on the target domain test set without additional annotations, ensuring that results accurately reflect the model's performance in practical scenarios.

Implementation Details. We implement LAID using PyTorch (Paszke et al. 2019) and train it on two NVIDIA RTX 4090 GPUs with a batch size of 4. We adopt the batched Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 3×10^{-3}, momentum of 0.9, and weight decay of 5×10^{-4}. The learning rate is warmed up in the first epoch. ResNet-50 (He et al. 2016) pre-trained on ImageNet-1k (Deng et al. 2009) serves as the model backbone. During training, input images are resized to 1500×900 and randomly horizontally flipped for data augmentation. For CAIM, considering the distribution characteristics of the CUHK-SYSU and PRW datasets, we set the momentum coefficient λ in Eq. 2 to 0.6 for CUHK-SYSU and 0.2 for PRW. For DSCL, the IoU threshold θ in Eq. 5 is set to 0.3; the separation margin γ in Eq. 7 is set to 0.5 for the source domain and 0.7 for the target domain.

Comparison with State-of-the-Art Methods

As shown in Tab. 1, to verify the effectiveness of LAID, we compare it with state-of-the-art (SOTA) methods employing various training strategies, including supervised, weakly supervised, and unsupervised domain adaptation approaches.

Comparison on CUHK-SYSU $\Longrightarrow$ PRW. The 5th column of Tab. 1 shows the performance of all the methods on the PRW test set. With 40.5% mAP and 81.9% top-1 accuracy, our framework outperforms most other weakly supervised

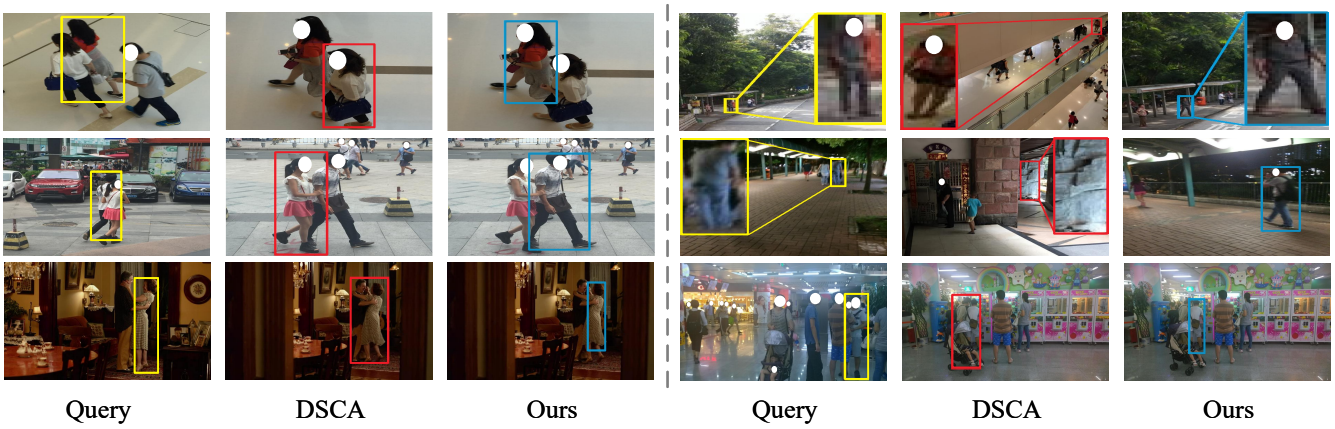


Figure 3: Qualitative comparison of **LAID** with **DSCA** on the **CUHK-SYSU** test set. The yellow bounding boxes denote the queries, while the red and blue bounding boxes denote incorrect and correct matches, respectively.

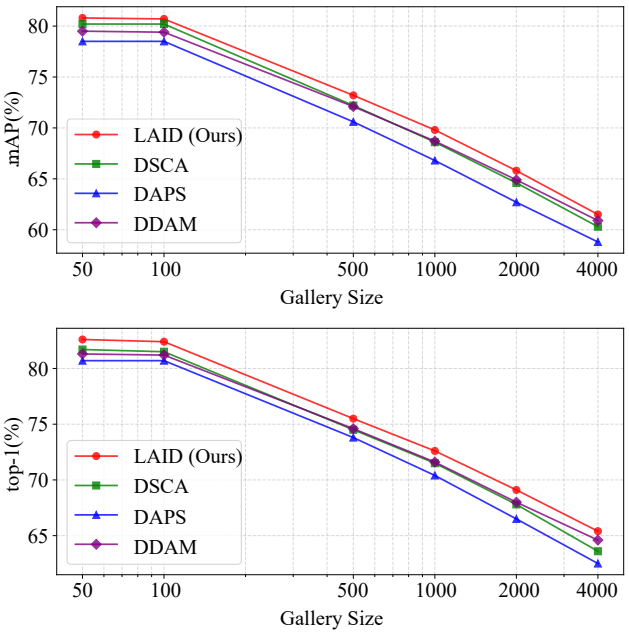


Figure 4: Comparison of performance on the **CUHK-SYSU** test set across various gallery sizes.

and domain adaptation methods. It is also competitive compared to some supervised learning methods.

Comparison on PRW $\Longrightarrow$ CUHK-SYSU. As shown in the 6th column of Tab. 1, our proposed LAID framework achieves excellent performance on the CUHK-SYSU test set, achieving 80.8% mAP and 82.6%. Compared with the SOTA method DSCA (Qi et al. 2025), it represents improvements of 0.6% and 0.9%, achieving SOTA performance.

To further validate the robustness of LAID, we compared it with other leading domain adaptation methods across a range of gallery sizes (50 to 4000) on the CUHK-SYSU test set. As illustrated in Fig. 4, with the increase in gallery size, all methods encounter greater interference from distractor

CAIM	PRW		CUHK-SYSU	
	mAP	top-1	mAP	top-1
w/o Loose Threshold	40.2	81.3	80.5	82.0
w/o Global Matching	39.6	80.8	79.9	81.8
	40.5	**81.9**	**80.8**	**82.6**

Table 2: Ablation studies of **CAIM** on both the **CUHK-SYSU** and **PRW** test set, evaluating the effectiveness of its match strategies.

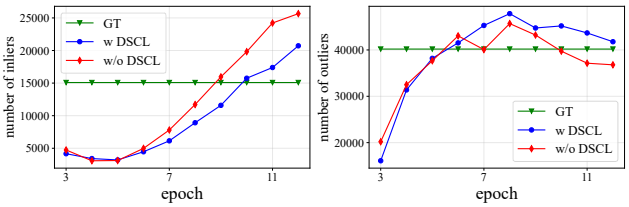
pedestrians. However, LAID maintains consistent superiority over competing methods across all size settings, confirming its robust discriminative capability even in cluttered scenarios.

Qualitative Results. Qualitatively, Fig. 3 highlights LAID's distinct advantages in complex, challenging scenarios. In practical domain-adaptive person search, two critical challenges often arise: (i) feature ambiguity caused by heavy overlap among pedestrians in crowded scenes; (ii) detail loss in small or occluded targets due to long-range imaging or partial occlusion.

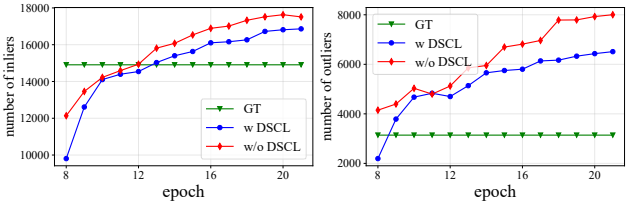
LAID addresses these by leveraging spatial relationships between bounding boxes as auxiliary supervisory cues, which mitigates over-reliance on shallow visual cues (e.g., color, posture). Instead, it prioritizes more discriminative features to mine the underlying semantic characteristics of pedestrians, enabling precise discrimination between distinct instances. This capability allows LAID to maintain robust performance in real-world scenarios with dense crowds and frequent occlusions, confirming its practical adaptability in complex environments.

Ablation Study

Effectiveness of Cost-Aware Instance Matching. To address misalignment between pseudo-labels and instances, CAIM employs a two-step strategy: it optimizes associations between current detections and historical proposals via cost



(a) Clustering results on the CUHK-SYSU dataset



(b) Clustering results on the PRW dataset

Figure 5: Comparison of clustering results with and without **DSCL** on the **CUHK-SYSU** and **PRW** training set.

DSCL	PRW		CUHK-SYSU	
	mAP	top-1	mAP	top-1
w/o Separation Loss	39.6	81.3	79.6	80.9
$w/$ Triplet Loss	40.2	81.1	80.0	81.6
	40.5	**81.9**	**80.8**	**82.6**

Table 3: Comparative results of the different configurations on **DSCL**.

TSA	PRW		CUHK-SYSU	
	mAP	top-1	mAP	top-1
w/o RPN $\Longleftarrow$ ReID	40.1	80.6	80.5	**82.7**
$w/$ RPN $\Longleftrightarrow$ ReID	37.4	79.3	76.2	78.4
	40.5	**81.9**	**80.8**	82.6

Table 4: Ablation studies of **TSA** on the **CUHK-SYSU** and **PRW** test set.

matrix construction, mitigating semantic confusion arising from misassociations caused by local overlap. The strategy comprises two core components: loose threshold-based candidate selection and global optimal assignment. The former retains more potential candidates by moderately relaxing the matching threshold, while the latter identifies true matches from the candidate set using a global optimal search algorithm. Their synergy effectively eliminates false association biases induced by scene interference and feature ambiguity.

We conducted ablation studies to isolate each component's contribution: (i) without loose threshold selection; (ii) without global optimal assignment; (iii) with both mechanisms enabled. As shown in the 1st and 2nd rows of Tab. 2, removing either component leads to notable drops in mAP and top-1 accuracy on both datasets. This confirms that both loose threshold selection and global optimal assignment are indispensable to CAIM, with their integrated design playing a critical role in ensuring strong domain-adaptive person search performance.

Effectiveness of Dual-Scope Contrastive Learning. DSCL serves as another critical component, introducing intra-scene spatial exclusion constraints to complement existing contrastive learning frameworks. It groups detected instances within a single scene based on geometric overlap and enforces differentiation via a separation loss, driving the model to explore more discriminative deep semantic features and enhancing pedestrian distinguishability. Tab. 3 shows that removing the separation loss leads to decreased mAP and top-1 accuracy on both the PRW and CUHK-SYSU datasets. We further replaced the separation loss with the batch-hard triplet loss, a common contrastive learning objective. While this replacement yields improvements over the baseline, results remain slightly inferior to those with the separation loss, confirming the efficacy of our designed loss function.

Clustering results in Fig. 5 further validate DSCL's advantages. Across training epochs, our method more closely aligns with ground truth in two key aspects: (i) For clustered instances, it exhibits a more stable growth trend and is closer to the true number of people with distinct identities, indicating more reliable pseudo-label generation for valid identities; (ii) For outliers, people without identity annotation, counts remain more consistent with actual values, particularly in later stages, reflecting a stronger capability to distinguish outliers from identity-bearing instances.

Overall, the closer alignment of both inlier counts and outlier counts with the ground truth when using DSCL confirms that the intra-scene spatial mutual exclusion constraints effectively enhance the model's capability to differentiate instances.

Effectiveness of Task-Sensitivity Alignment. Tab. 4 presents ablation studies on the TSA mechanism. Removing the alignment between RPN and ReID confidence predictions or adopting bidirectional alignment both result in notable performance drops on both PRW and CUHK-SYSU datasets. This confirms the effectiveness of aligning confidence predictions across tasks. Such alignment is critical because detection and ReID tasks naturally exhibit distinct confidence distribution patterns. TSA mitigates this discrepancy by calibrating their confidence scales, ensuring consistent pseudo-label reliability and thereby reinforcing the mutual promotion between detection accuracy and feature discriminability.

Conclusion

This paper proposes a Localization-Anchored Instance Discrimination framework for the domain-adaptive person search task. Anchored on spatial localization information within scenes, the framework constructs a multi-dimensional constraint mechanism, which systematically improves the reliability of pseudo-labels and the discriminability of instance features. Extensive experiments verify the superior performance of the proposed method, providing an effective solution to narrow the performance gap between supervised and domain adaptation methods.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China Grant 2024YFB4710800, National Natural Science Foundation of China Grant 62576067, 62501226, 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 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. 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. 2020b. 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*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

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.

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.; 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*.

Jonker, R.; and Volgenant, T. 1983. Transforming asymmetric into symmetric traveling salesman problems. *Operations Research Letters*, 2(4): 161–163.

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.

Qi, L.; Wang, H.; Zhang, J.; Peng, J.; and Wang, Y. 2025. Unsupervised Domain Adaptive Person Search via Dual Self-Calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6550–6558.

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, 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.

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.

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.

Zhu, H.; Yang, X.; and Wang, N. 2025. Optimizing Label Assignment for Weakly Supervised Person Search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10941–10949.