Reliable Pseudo-Supervision for Unsupervised Domain Adaptive Person Search

Qixian Zhang, Duoqian Miao, Qi Zhang, Xuan Tan, Hongyun Zhang,
and Cairong Zhao, *Senior Member, IEEE*

***Abstract*—Unsupervised Domain Adaptation (UDA) person search aims to adapt models trained on labeled source data to unlabeled target domains. Existing approaches typically rely on clustering-based proxy learning, but their performance is often undermined by unreliable pseudo-supervision. This unreliability mainly stems from two challenges: (i) spectral shift bias, where low- and high-frequency components behave differently under domain shifts but are rarely considered, degrading feature stability; and (ii) static proxy updates, which make clustering proxies highly sensitive to noise and less adaptable to domain shifts. To address these challenges, we propose the Reliable Pseudo-supervision in UDA Person Search (RPPS) framework. At the feature level, a Dual-branch Wavelet Enhancement Module (DWEM) embedded in the backbone applies discrete wavelet transform (DWT) to decompose features into low- and high-frequency components, followed by differentiated enhancements that improve cross-domain robustness and discriminability. At the proxy level, a Dynamic Confidence-weighted Clustering Proxy (DCCP) employs confidence-guided initialization and a two-stage online–offline update strategy to stabilize proxy optimization and suppress proxy noise. Extensive experiments on the CUHK-SYSU and PRW benchmarks demonstrate that RPPS achieves state-of-the-art performance and strong robustness, underscoring the importance of enhancing pseudo-supervision reliability in UDA person search. Our code is accessible at https://github.com/zqx951102/RPPS**

***Index Terms*—Person search, unsupervised domain adaptation, frequency components, clustering proxy, wavelet transform.**

I. Introduction

PERSON search [1], [2], [3], [4], [5] aims to jointly perform pedestrian detection and person re-identification (ReID) [6], [7], [8] in complex images or videos, thereby enabling accurate retrieval of target individuals. This task has important applications in intelligent surveillance and public security. In recent years, supervised learning methods have achieved remarkable success on labeled datasets. However, their generalization ability is often severely constrained by cross-scene variations such as changes in viewpoint, camera configuration, and environment. Moreover, obtaining large-scale cross-domain annotations is prohibitively expensive, which further limits practical deployment. Consequently, unsupervised domain adaptation (UDA) has emerged as a key research direction for person search in real-world scenarios.

Existing UDA person search methods [9], [10], [11], [12] mainly rely on source-domain pretraining and pseudo-label generation in the unlabeled target domain. Representative works include Li et al. [9], who first introduced UDA into person search by employing a domain-alignment module to mitigate task discrepancies between detection and ReID, together with an epoch-wise clustering strategy for generating pseudo labels; DDAM [10], which enhances knowledge transfer through hybrid domain representations; and FOUS [11], which employs prototype-guided pseudo-label assignment to reduce redundancy and alleviate noisy supervision. While these approaches have significantly advanced the field, they focus primarily on spatial-domain alignment and clustering-based pseudo-labeling. Consequently, the role of frequency-domain features has remained largely unexplored, and proxy optimization remains vulnerable to noise and domain shifts.

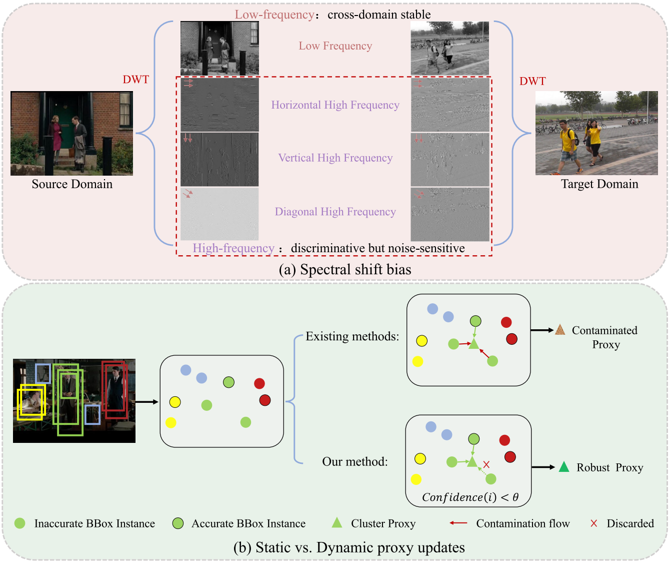
Specifically, two core challenges remain unsolved. First, spectral shift bias has not been explicitly modeled. As illustrated in Fig. 1(a), low-frequency components mainly encode coarse structural cues such as human contours, which are relatively invariant to camera styles and illumination changes, and thus remain stable across domains. In contrast, high-frequency components capture fine-grained identity-related details such as clothing textures and edges, which are more discriminative for distinguishing visually similar individuals [13], [14]. However, these high-frequency signals are also more sensitive to cross-domain noise, including illumination variation, background clutter, and detection misalignment. This trade-off between discriminability and robustness motivates the need for explicitly decomposing and differentially enhancing low- and high-frequency components, rather than treating them uniformly. Second, static proxy update strategies remain limited. As shown in Fig. 1(b), pseudo labels in the target domain are clustering-dependent, but inaccurately localized bounding boxes often introduce additional background noise, contaminating proxy representations. Moreover, existing methods lack fine-grained quantification of instance quality, and their updates are typically static, making it difficult to adapt to domain shifts and limiting long-term adaptability.

Received 27 September 2025; revised 21 December 2025; accepted 7 January 2026. Date of publication 21 January 2026; date of current version 27 January 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62576251, Grant 62376198, and Grant 62576247; and in part by the National Key Research and Development Program under Grant 2022YFB3104700. The associate editor coordinating the review of this article and approving it for publication was Dr. Huafeng Li. *(Corresponding author: Duoqian Miao.)*

The authors are with the School of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: zhangqx@tongji.edu.cn; dqmiao@tongji.edu.cn; zhangqi_cs@tongji.edu.cn; xtan@tongji.edu.cn; zhanghongyun@tongji.edu.cn; zhaocairong@tongji.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2026.3654373, provided by the authors.

Digital Object Identifier 10.1109/TIP.2026.3654373



(a) Spectral shift bias

(b) Static vs. Dynamic proxy updates

Fig. 1. Challenges in UDA person search. (a) Spectral shift bias: low-frequency components remain stable across domains, whereas high-frequency ones are discriminative but highly noise-sensitive. (b) Static vs. Dynamic proxy updates: existing methods rely on fixed clustering that is easily disrupted by noisy bounding boxes, whereas our confidence-guided strategy filters out low-quality instances to yield more robust proxies.

To address these challenges, we propose a novel framework termed **R**eliable **P**seudo-supervision in UDA **P**erson **S**earch (RPPS). Specifically: (1) To effectively mitigate spectral shift bias, we design a Dual-branch Wavelet Enhancement Module (DWEM). By applying the discrete wavelet transform (DWT) [15], features are decomposed into low- and high-frequency components. The low-frequency branch performs fine-grained enhancement to exploit cross-domain stability, whereas the high-frequency branch applies aggregated enhancement to emphasize discriminative details, thereby fully leveraging frequency-domain cues for robust representation learning. (2) To overcome the inherent limitations of static proxy updates, we introduce a Dynamic Confidence-weighted Clustering Proxy (DCCP). A dynamic confidence function, combining bounding-box IoU and feature entropy, quantifies instance quality to guide reliable proxy initialization. Furthermore, we develop a two-stage online–offline update strategy: in the online phase, confidence-weighted updates assign stronger updates to high-confidence instances and weaker ones to low-confidence instances; in the offline phase, confidence-guided stratification and asynchronous correction further purify proxies and integrate cross-epoch features, thereby significantly enhancing long-term adaptability to domain shifts. Extensive experiments on two challenging benchmark datasets, CUHK-SYSU and PRW, demonstrate that RPPS substantially outperforms state-of-the-art UDA person search methods, validating the effectiveness of frequency-domain enhancement and confidence-driven proxy optimization.

The main contributions are summarized as follows:

- We propose a novel framework, RPPS, which, to our knowledge, is the first to jointly tackle spectral shift bias and static proxy updates, thereby improving the reliability of pseudo-supervision in UDA person search.
- We design a DWEM that leverages DWT to decompose features into low- and high-frequency components, followed by differentiated enhancement strategies. This explicitly mitigates spectral shift bias by exploiting frequency-domain information.
- We introduce a DCCP to overcome static proxy update limitations, quantifying instance quality via a dynamic confidence function for robust initialization and employing a two-stage online–offline update strategy to ensure long-term adaptability under domain shifts.
- We conduct extensive experiments on the CUHK-SYSU and PRW benchmarks, demonstrating that RPPS consistently surpasses state-of-the-art UDA methods and delivers superior robustness to domain shifts and noisy.

The remainder of this paper is organized as follows. Section I reviews related work. Section III introduces the overall architecture of the proposed RPPS framework and elaborates on its key modules. Section IV presents the experimental setup and analyzes the results. Finally, Section V concludes the paper and outlines potential directions for future research.

II. Related Work

A. Person Search

With the rapid development of deep learning and the availability of large-scale benchmark datasets [2], [3], person search [1], [16], [17], [18] has become a prominent research topic in computer vision. This task requires the joint accomplishment of pedestrian detection and person re-identification (ReID). Existing fully supervised person search models can be broadly classified into two-step and one-step frameworks.

Two-step frameworks train the detection and ReID models independently. For example, Chen et al. [19] proposed the Mask-Guided Two-Stream (MGTS) method, which separates the detector and the ReID model to extract richer features. Han et al. [20] introduced an ROI transformation layer to provide more precise detection boxes for person search. Wang et al. [21] developed the Task-Consistent Two-Stage (TCTS) framework to alleviate inconsistencies between the two tasks.

In contrast, one-step frameworks have recently become the mainstream due to their efficiency. The key idea is to construct a unified model that jointly optimizes both tasks in an end-to-end manner. Xiao et al. [2] proposed the first one-step person search model, optimized using an online instance matching loss. Chen et al. [22] introduced an embedding decomposition approach to resolve task conflicts effectively. Yan et al. [23] designed a feature alignment network to address scale, region, and objective misalignments. Li and Miao et al. [24] developed a sequential structure to reduce low-quality candidate boxes. Yu et al. [25] employed a Transformer encoder to mix image patches of different pedestrians, significantly enhancing generalization to unseen scenarios. Chen et al. [26] leveraged human priors to generate pseudo-semantic labels, injecting richer semantics into feature learning. Kim et al. [27] proposed a prototype-guided attention distillation method to highlight identity-specific regions across diverse poses. Yang et al. [4] introduced an Efficient Tri-Hybrid framework that balances the conflicts between detection and ReID via hierarchical feature design. Cai et al. [28] developed a guided multi-task training strategy to achieve more balanced end-to-end learning.

Nevertheless, all these fully supervised methods rely heavily on large-scale labeled datasets and suffer from severe performance degradation in cross-domain scenarios due to domain discrepancies. More recently, several studies [29], [30], [31], [32] have explored weakly supervised settings (e.g., without identity labels). However, they still require a small amount of labeled data, which remains impractical in real-world scenarios where annotations are unavailable in the target domain.

B. Domain Adaptation for Person ReID

Unsupervised Domain Adaptation (UDA) has emerged as an effective paradigm for cross-domain adaptation. In the person ReID task, the core idea is to train a model on labeled source-domain data and transfer it to an unlabeled target domain. Mainstream UDA ReID approaches can be broadly categorized into two groups: (1) GAN-based methods [33], [34], [35], which mitigate style discrepancies such as illumination and background shifts to better adapt to target-domain scenarios; and (2) clustering-based methods [36], [37], [38], [39], which assign pseudo labels to target-domain samples to supervise training.

Beyond conventional UDA settings, several works have explored more challenging scenarios with scarce or missing cross-camera paired samples. Li et al. [6] proposed a domain-adaptive ReID framework without cross-camera paired supervision, while Li et al. [7] further leveraged large-scale unpaired samples for domain generalization. To improve pseudo-label reliability under camera-style variations, Li et al. [40] introduced logical relation inference and multiview information interaction to generate more robust pseudo labels via a two-stage intra- and inter-camera framework. In parallel, multi-modal ReID has been actively studied. Zhang et al. [41] proposed a weakly supervised visible–infrared ReID method based on heterogeneous expert collaboration, while Li et al. [5] incorporated language guidance into video-based visible–infrared ReID using CLIP. From a geometric perspective, Leng et al. [42] explored dual-space video ReID by jointly modeling Euclidean and hyperbolic representations. In addition, CLIP-based text–image ReID methods [8], [43] further enhanced cross-modal alignment through fine-grained correspondence mining and prompt learning.

However, conventional UDA ReID frameworks operate on cropped pedestrian images and therefore cannot effectively address detection noise in person search, where bounding-box annotations are absent in the target domain. In addition, clustering-based proxies in ReID methods are typically updated in a static manner, rendering them highly sensitive to noisy samples. In contrast, our approach explicitly tackles both detection noise and domain shift. The proposed DCCP quantifies instance quality by combining IoU and feature entropy to select reliable samples, and further employs a two-stage online–offline update strategy that dynamically refines proxies across epochs, thereby surpassing the static clustering schemes commonly adopted in ReID-focused UDA methods.

C. Domain Adaptation for Person Search

The core of UDA person search is to transfer knowledge from labeled source-domain data to unlabeled target-domain data, and its research value has attracted increasing attention. Several studies have recently explored this direction, yet notable limitations remain. For example, Li et al. [9] first introduced UDA into person search by incorporating a domain-alignment module and dynamic clustering to generate pseudo labels. However, implicit alignment often degrades the quality of regions of interest, while dynamic clustering tends to fail in scenarios with visually similar pedestrians, resulting in unreliable pseudo labels. Almansoori et al. [10] proposed an explicit bridging mechanism to mitigate domain discrepancy, but its complexity introduces model redundancy and parameter sensitivity, which can easily add extra noise. Cui et al. [11] developed a prototype-guided labeling strategy to reduce redundant computation, yet its strong reliance on prototype accuracy and coarse handling of noisy labels frequently cause erroneous assignments. Qi et al. [12] introduced a perception-driven filter and a Clustering Proxy Representation (CPR) strategy; however, the filter struggles to suppress pseudo bounding boxes under occlusion, while CPR suffers from delayed updates in large datasets, impairing classification accuracy.

In summary, although methods such as DSCA [12] aim to improve pseudo-label reliability, they still rely on prototype accuracy or static update strategies. Such designs overlook both instance-level quality estimation and frequency-domain discrepancies. To bridge these gaps, we introduce the DWEM to reinforce low-frequency stability and high-frequency discriminability, and the DCCP to dynamically stabilize proxy optimization under noisy supervision.

III. Methodology

This section first introduces the problem formulation and provides an overview of the proposed RPPS framework. We then detail its two core components: the Dual-branch Wavelet Enhancement Module (DWEM), which leverages frequency-domain information by separately enhancing low-frequency and high-frequency features, and the Dynamic Confidence-weighted Clustering Proxy (DCCP), which enhances proxy robustness and adaptability through confidence-guided initialization and a two-stage update strategy.

A. Problem Formulation

Given a source domain D_s and a target domain D_t, the goal of UDA person search is to transfer knowledge learned from the source domain, including pedestrian detection, identification, and their associations, to the target domain D_t. This enables improved performance in joint pedestrian detection and cross-image identity matching. The source domain dataset is defined as $D_s = \left\{\left(x_s^i, y_s^i\right)\right\}_{i=1}^{N_s}$, where x_s^i denotes the i-th training sample containing pedestrian scenes and y_s^i represents its corresponding annotations (e.g., bounding box positions and identity labels). Here, N_s is the total number of source domain images. The target domain dataset is denoted as $D_t = \left\{x_t^i\right\}_{i=1}^{N_t}$, where x_t^i is the i-th image containing pedestrian

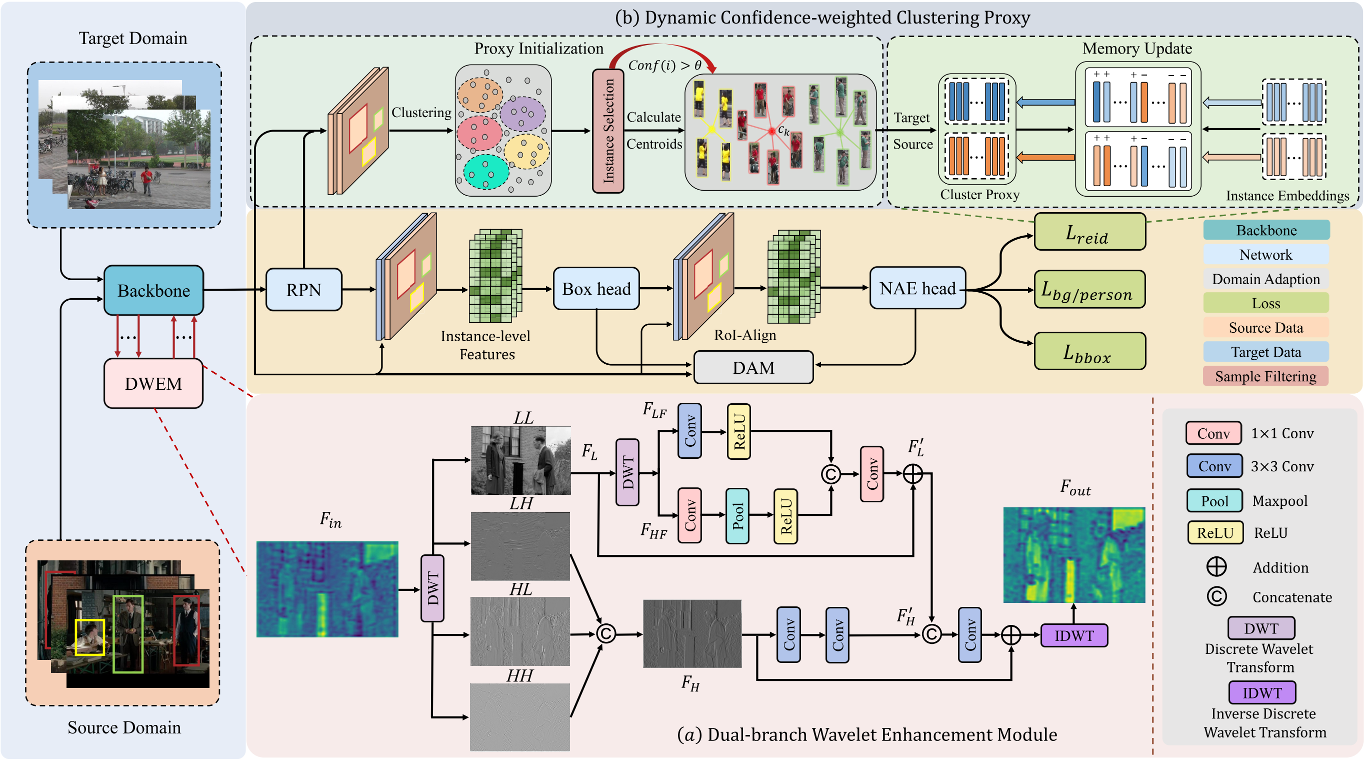


Fig. 2. Illustration of the proposed RPPS framework. (a) The backbone network equipped with the Dual-branch Wavelet Enhancement Module (DWEM), which applies discrete wavelet transform (DWT) to decompose features and separately enhance low-frequency stability and high-frequency discriminability, thereby mitigating spectral shift bias in cross-domain representation learning. (b) The Dynamic Confidence-weighted Clustering Proxy (DCCP), which employs confidence-guided initialization and a two-stage online–offline update strategy to stabilize proxy optimization under domain shifts, mitigating the impact of noisy pseudo labels and improving long-term adaptability.

scenes in the target domain, but no pedestrian annotations such as bounding boxes or identity labels are provided. N_t denotes the total number of target domain images.

In summary, the model is trained on labeled source domain data for joint learning of detection and ReID, and is expected to generalize to the unlabeled target domain. The key challenge lies in mitigating domain discrepancies, which often cause substantial performance degradation in accurate pedestrian localization and cross-image identity matching.

B. Framework Overview

The overall workflow of the proposed RPPS framework is illustrated in Fig. 2. It is built upon the DAPS [9] network, which incorporates an implicit Domain Alignment Module (DAM) to effectively reduce feature discrepancies between the source and target domains. For both source- and target-domain inputs, the first four convolutional layers of ResNet-50 [44] serve as the backbone to extract basic image-level features. On top of this backbone, a Dual-branch Wavelet Enhancement Module (DWEM) is embedded at each layer. Specifically, DWEM applies a Discrete Wavelet Transform (DWT) to decompose and recombine features in the frequency domain, thereby reinforcing the cross-domain stability of low-frequency contours while enhancing the discriminability of high-frequency textures, resulting in enriched image-level feature maps. These enhanced features are then fed into a Region Proposal Network (RPN) [45] to generate candidate bounding boxes, which are further processed by an RoI-Align [46] layer to obtain instance-level feature maps for subsequent detection and re-identification tasks, implemented through the NAE [22] head. In addition, the DAM module mitigates domain discrepancies at both the image and instance levels, further improving cross-domain feature adaptation.

To stabilize proxy optimization and enhance robustness against noisy pseudo labels, we propose the Dynamic Confidence-weighted Clustering Proxy (DCCP) module. Specifically, pseudo labels for target-domain instances are first generated using a clustering algorithm (e.g., DBSCAN [47]). A dynamic confidence function, combining bounding-box IoU and feature entropy, is then employed to quantify instance quality, and high-confidence samples satisfying $Conf(i) > \theta$ are selected for robust proxy initialization. Based on these reliable samples, clustering centers c_k are computed as initial proxies. To further refine proxies, we design a two-stage update strategy. In the online phase, confidence-weighted losses apply strong updates to high-confidence instances and weak updates to low-confidence ones. In the offline phase, confidence-guided stratification and asynchronous correction purify proxies and integrate cross-epoch features. Together, these steps stabilize proxy optimization and enable dynamic adaptation to domain shifts, while mitigating the negative impact of noisy pseudo labels.

C. Dual-Branch Wavelet Enhancement Module

To address the core challenge in UDA person search, namely spectral shift bias, we propose the DWEM. Low-frequency components (e.g., human contours) are relatively stable across domains, whereas high-frequency components (e.g., clothing textures) provide strong discriminability but are highly noise-sensitive. DWEM applies the DWT to decompose features into four subbands, performs differentiated enhancement in two branches, and reconstructs them via the IDWT with the Haar wavelet basis. This process explicitly mitigates spectral shift bias by leveraging frequency-domain cues, yielding more robust representations for subsequent domain alignment and proxy optimization. The overall architecture of DWEM is illustrated in Fig. 2(a).

1) DWT and IDWT: The DWT is applied to the image-level features $x \in \mathbb{R}^{B\times C\times H\times W}$ output by the backbone, where B, C, H, and W denote the batch size, number of channels, height, and width, respectively. Using a 2×2 sliding window, x is downsampled into four subbands: the low-frequency component (LL) preserving global structures, and three high-frequency components (LH, HL, HH) capturing horizontal, vertical, and diagonal edges with local details. The computation of each subband is defined as follows:

$$\begin{aligned}
LL &= \frac{1}{4}\left(x_{i,j}+x_{i,j+1}+x_{i+1,j}+x_{i+1,j+1}\right),\\
LH &= \frac{1}{4}\left(-x_{i,j}+x_{i,j+1}-x_{i+1,j}+x_{i+1,j+1}\right),\\
HL &= \frac{1}{4}\left(-x_{i,j}-x_{i,j+1}+x_{i+1,j}+x_{i+1,j+1}\right),\\
HH &= \frac{1}{4}\left(x_{i,j}-x_{i,j+1}-x_{i+1,j}+x_{i+1,j+1}\right),
\end{aligned} \tag{1}$$

here, $x_{i,j}$ denotes the pixel value at position (i, j), $i, j \in \{0, 1\}$ within the 2×2 sliding window of the feature map. After decomposition, the output features are concatenated along the channel dimension into four subbands: dwt_out = $[LL, LH, HL, HH] \in \mathbb{R}^{B\times 4C\times H/2\times W/2}$, which serve as the input for the subsequent dual-branch enhancement.

To reconstruct the enhanced frequency features back to their original spatial resolution for compatibility with the RPN and ROI-Align layers, the IDWT is employed. Specifically, the four decomposed subbands $LL, LH, HL, HH \in \mathbb{R}^{B\times C\times H/2\times W/2}$ are recombined as follows to produce a feature map consistent with the input size:

$$\begin{aligned}
x_{i,j} &= \frac{1}{2}\left(LL{-}LH{-}HL+HH\right),\\
x_{i,j+1} &= \frac{1}{2}\left(LL+LH{-}HL{-}HH\right),\\
x_{i+1,j} &= \frac{1}{2}\left(LL{-}LH+HL{-}HH\right),\\
x_{i+1,j+1} &= \frac{1}{2}\left(LL+LH+HL+HH\right).
\end{aligned} \tag{2}$$

Theoretical analysis of DWT, including its frequency-domain decomposition, energy conservation, and stability, is provided in the **Supplementary Material**.

2) Dual-Branch Differentiated Enhancement: To address the distinct characteristics of different frequency subbands produced by the DWT, we design a *low-frequency fine-grained enhancement branch* and a *high-frequency aggregated enhancement branch*, as illustrated in Fig. 2(a). These branches perform differentiated transformations and fusion to strengthen cross-domain feature representations.

For the low-frequency component F_L (corresponding to human contour information), which carries stable cross-domain structural cues, we apply a second-level DWT to obtain fine-grained low- and high-frequency subbands F_{LF} and F_{HF}. Local structural features are then extracted using a 3×3 convolution:

$$F'_{LF} = \mathrm{ReLU}\left(W^{LF}_{3\times 3}\otimes F_{LF}+b_{LF}\right), \tag{3}$$

where $W^{LF}_{3\times 3}$ and b_{LF} denote the convolutional weights and bias, respectively, and $\otimes$ represents convolution.

For the high-frequency component F_{HF}, convolution and pooling are employed to suppress noise while preserving discriminative cues:

$$F'_{HF} = \mathrm{ReLU}\left(\mathrm{MaxPool}\left(W^{HF}_{1\times 1}\otimes F_{HF}+b_{HF}\right)\right), \tag{4}$$

where $\mathrm{MaxPool}(\cdot)$ denotes global max pooling, and $W^{HF}_{1\times 1}$ implements channel compression. After convolution and pooling, F'_{HF} is resized via upsampling and concatenated with F'_{LF} along the channel dimension. A 1×1 convolution and upsampling operation are then applied, followed by residual addition with the original low-frequency component F_L:

$$F'_L = \mathrm{Up}\Big(W^{\mathrm{fuse}}_{1\times 1}\otimes\left[F'_{LF}\,\|\,\mathrm{Up}(F'_{HF})\right]+b_{\mathrm{fuse}}\Big)+F_L, \tag{5}$$

where $[\|]$ denotes channel concatenation and $\mathrm{Up}(\cdot)$ represents a $2\times$ upsampling operation.

For the high-frequency subbands LH, HL, HH (corresponding to the high-frequency branch F_H), the three components are first concatenated along the channel dimension, followed by two successive 3×3 convolutions:

$$F'_H = W^2_{3\times 3}\otimes\left(W^1_{3\times 3}\otimes[LH\,\|\,HL\ \|\,HH]+b_1\right)+b_2, \tag{6}$$

where $W^1_{3\times 3}$ and $W^2_{3\times 3}$ denote the weights of the two convolutional layers, which progressively suppress background noise while enhancing discriminative details such as clothing textures and edge patterns.

Finally, the enhanced low-frequency feature F'_L and high-frequency feature F'_H are concatenated along the channel dimension, fused with F_H through a 3×3 convolution, and replicated four times to match the input format of the IDWT:

$$F_{\mathrm{out}} = \mathrm{IDWT}\Big(\big(W^{\mathrm{final}}_{3\times 3}\otimes[F'_L\,\|\,F'_H]+F_H+b_{\mathrm{final}}\big)^{\otimes 4}\Big), \tag{7}$$

where $(\cdot)^{\otimes 4}$ denotes feature replication and concatenation across four subbands. The final output F_{out} is reconstructed into a feature consistent with the original spatial resolution.

D. Dynamic Confidence-Weighted Clustering Proxy

To address noisy instances and the limitations of static proxy updates in UDA person search, we propose the DCCP. DCCP stabilizes proxy optimization under domain shifts through confidence-guided initialization and a two-stage online–offline update strategy. Specifically, a dynamic confidence function that integrates bounding-box IoU and feature entropy quantifies instance reliability across multiple dimensions, enabling robust initialization and adaptive updates. Unlike static schemes, DCCP adaptively emphasizes reliable instances, reduces contamination risk, achieves precise quality estimation, and ensures long-term adaptability.

1) Proxy Initialization: The DCCP mechanism first selects high-quality instances for proxy initialization to construct more robust clustering proxies, as illustrated in Fig. 2(b).

a) Dynamic Confidence Computation: For each instance i, we calculate the overlap between the pedestrian search box (predicted by the source-pretrained model) and the pseudo bounding box using the Intersection-over-Union (IoU), and combine it with feature entropy to define the confidence score:

$$Conf(i) = \alpha \cdot IoU(i) + (1-\alpha)\cdot\left[1 - Entropy(f_i)\right], \tag{8}$$

where $IoU(i) \in [0,1]$ measures the accuracy between the pseudo and ground-truth bounding boxes, and α is a balancing factor. The feature entropy $Entropy(f_i)$ reflects the uncertainty of the instance feature distribution, formulated as:

$$Entropy(f_i) = -\sum_{d=1}^{D} p(f_{i,d})\log p(f_{i,d}), \tag{9}$$

where D is the feature dimension, and $p(f_{i,d})$ denotes the probability of the d-th feature of instance i. A higher entropy indicates greater feature uncertainty, whereas lower entropy implies more discriminative features.

b) Instance Selection and Proxy Initialization: We first set a confidence threshold (e.g., $Conf(i) > \theta$) to select high-confidence instances, thereby avoiding the contamination of clustering proxies by unreliable samples. The selected instances are then used to construct clustering proxies, improving the robustness of proxy initialization. For the selected instances, the proxy initialization is formulated as:

$$c_k = \frac{1}{N_k}\sum_{f_i \in F_k} f_i, \tag{10}$$

where F_k denotes the set of instance features belonging to the k-th cluster, N_k is the number of samples in the k-th cluster, and c_k represents the proxy of the k-th cluster. Finally, the initialized proxies are stored in a memory dictionary $M = \{(c_1, c_2, \ldots, c_K)\}$, where K is the number of clusters.

2) Memory Update: Memory update serves as a core component in UDA person search for adaptively refining clustering proxies. The DCCP mechanism employs a two-stage strategy, consisting of an **online update** phase and an **offline update** phase, both guided by dynamic confidence weighting. The key idea is to assign confidence scores as explicit weights during updates, thereby reinforcing reliable instances and mitigating the influence of noisy ones, as illustrated in Fig. 2(b).

a) Online Update: During online updates, the memory $M = \{c_1, c_2, \ldots, c_K\}$ is treated as a lookup table (LUT) storing class-level proxies. For a query instance feature f extracted by the ReID branch, its confidence-weighted similarity to each proxy is computed to optimize the model. The confidence-weighted loss is defined as:

$$L_{\text{online}} = -\log\frac{\exp\left(f\cdot c_{+}/\tau\right)}{\sum_{k=0}^{K}\exp\left(f\cdot c_k/\tau\right)}, \tag{11}$$

where c_{+} denotes the proxy of the class that f belongs to, and τ is a temperature factor controlling the smoothness of the probability distribution.

During backpropagation, the consistency between the confidence of f and its assigned proxy c_k is evaluated:

$$\text{Consist}(f, c_k) = \frac{\left|\text{Conf}(f) - \text{Conf}_{\text{history}}(c_k)\right|}{\text{Conf}_{\text{history}}(c_k)} < \delta. \tag{12}$$

If the condition is satisfied, the proxy is updated using confidence-weighted aggregation:

$$c_k \leftarrow (1-\gamma\cdot\text{Conf}(f))\cdot c_k + \gamma\cdot\text{Conf}(f)\cdot f, \tag{13}$$

where γ is a momentum factor controlling the update rate.

b) Offline Update: The offline stage focuses on the long-term optimization of clustering proxies. We design a *confidence stratification and asynchronous correction* mechanism to remove outdated proxies and integrate cross-epoch features.

The global confidence of each proxy c_k is computed as:

$$\text{GlobalConf}(c_k) = \frac{1}{N_k}\sum_{f\in F_k}\text{Conf}(f), \tag{14}$$

where F_k is the set of instance features assigned to proxy c_k. Proxies are divided into high-confidence ($\text{GlobalConf}(c_k) \geq \epsilon$) and low-confidence groups. For low-confidence proxies, we adopt a replacement strategy by finding the most similar high-confidence proxy based on cosine similarity:

$$\text{Sim}(c_{\text{low}}, c_{\text{high}}) = \frac{c_{\text{low}}^{\top} c_{\text{high}}}{\|c_{\text{low}}\|\cdot\|c_{\text{high}}\|}, \tag{15}$$

and replace c_{low} with its nearest high-confidence proxy.

To ensure the long-term adaptability of proxies during UDA, DCCP incorporates asynchronous learning. At the beginning of each epoch, proxies are re-synchronized with pseudo labels and refined using instance features from the previous epoch via Exponential Moving Average (EMA):

$$c_k = \frac{1}{N_k}\sum_{f_i\in F_k}\left[m\cdot\text{Match}(F_k^{t-1}, f_i) + (1-m)\cdot f_i\right], \tag{16}$$

where f_i^t denotes the i-th feature at epoch t, F_k^{t-1} is the feature set of cluster k at epoch $t-1$, $\text{Match}(\cdot)$ retrieves the matched instance from the previous epoch, and m is the EMA smoothing factor.

For clarity, we present the detailed procedure of the DCCP in Algorithm 1.

E. Discussion

Although DWEM and DCCP operate at the feature level and proxy level respectively, they exhibit a strong synergistic relationship and jointly enhance the reliability of pseudo-supervision in unsupervised domain adaptive person search. DWEM explicitly separates low- and high-frequency components through wavelet decomposition and applies differentiated enhancement, resulting in cross-domain features that are more stable and more discriminative. Such cleaner and more robust representations directly improve the quality of clustering in the subsequent proxy learning stage. Building on these enhanced features, DCCP measures sample reliability through dynamic confidence estimation and adopts an online–offline dual-stage proxy update strategy to effectively suppress the drift caused by noisy pseudo-labels. Since DWEM reduces spectral noise

Algorithm 1 Dynamic Confidence-Weighted Clustering Proxy

```
Input: Target-domain features $\{f_i\}$, pseudo labels $\{y_i\}$ from clustering, total epochs $E$
Output: Proxy dictionary $M = \{c_k\}_{k=1}^{K}$
Initialization: Compute confidence $\mathrm{Conf}(i)$ by Eq. (8); select high-confidence instances $\{f_i \,|\, \mathrm{Conf}(i) > \theta\}$. Initialize proxies $c_k$ by averaging selected features per cluster by Eq. (10).
for $e = 1$ to $E$ do
    Online update: Compute confidence-weighted loss $L_{\text{online}}$ by Eq. (11); update proxies with momentum: $c_k \leftarrow (1 - \gamma \cdot \mathrm{Conf}(f))\, c_k + \gamma \cdot \mathrm{Conf}(f) \cdot f$.
    Offline update: Compute $\mathrm{GlobalConf}(c_k)$ by Eq. (14); replace unreliable proxies with nearest high-confidence ones by Eq. (15); refine with EMA-based asynchronous correction by Eq. (16).
end for
return $M = \{c_k\}_{k=1}^{K}$
```

and strengthens identity-related high-frequency cues, DCCP can assess sample confidence more accurately and achieve more robust proxy optimization.

Therefore, the two modules form a complementary mechanism: DWEM determines whether the features used for clustering are stable, while DCCP determines whether the proxies obtained from clustering are reliable. This feature-to-proxy synergistic design enables RPPS to achieve significant performance gains across different benchmarks.

IV. Experiments

In this section, we conduct comprehensive experiments on two widely used person search benchmarks. We first introduce the datasets, evaluation metrics, and implementation details. We then compare our approach with state-of-the-art methods. Subsequently, ablation studies are performed to examine the effectiveness of each proposed module. Finally, we provide additional experiments and qualitative analysis to further demonstrate the advantages of our framework.

A. Datasets and Metrics

1) Datasets: Our method is evaluated on two large-scale benchmark datasets, namely CUHK-SYSU [2] and PRW [3].

CUHK-SYSU [2] is a large-scale person search dataset containing 8,432 identities across 18,184 images, with a total of 96,143 bounding boxes. During training, 5,532 identities with 11,206 images are available, while the remaining 2,900 identities with 6,978 images are reserved for evaluation. The dataset consists of two distinct sources: (1) street images, which exhibit variations in viewpoint, illumination, resolution, and occlusion; and (2) movie and TV frames, which introduce diverse indoor and outdoor challenges. This diversity makes CUHK-SYSU highly representative of real-world scenarios. For evaluation, 2,900 query persons are selected, and the remaining 6,978 images serve as the gallery set.

PRW [3] is another widely used dataset, consisting of 932 identities, 11,816 images, and 43,110 bounding boxes, captured from six outdoor surveillance cameras on a university campus. The training set includes 482 identities with 5,702 images, while the test set contains 2,057 query persons and a gallery of 6,112 images.

2) Evaluation Metrics: In the dataset configuration, the source-domain dataset is annotated with bounding boxes and identity labels, whereas the target-domain dataset lacks such annotations. For detection, we adopt Average Precision (AP) and Recall as evaluation metrics. For the re-identification (ReID) sub-task, we report mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) top-1 accuracy. Since ReID directly reflects the ability to retrieve the correct identity of a query person, it is considered the most challenging metric in person search.

B. Implementation Details

We implement our proposed RPPS using PyTorch and train it on a Tesla V100 GPU with a batch size of 4. The model is optimized using stochastic gradient descent (SGD) with an initial learning rate of 0.0024 and a warm-up in the first epoch. We adopt ResNet-50 [44] pre-trained on ImageNet-1k [48] as the backbone network. During training, input images are resized to 1500 × 900 and augmented with random horizontal flipping. The hyperparameters are set as follows: balance coefficient $\alpha = 0.6$, confidence threshold $\theta = 0.7$, consistency threshold $\delta = 0.3$, and global confidence threshold $\epsilon = 0.5$, which are used for both online and offline proxy updates. To prevent overfitting, the number of training epochs is adjusted according to the variability of the target dataset. Specifically, when PRW is used as the target domain, our RPPS is first pre-trained on CUHK-SYSU for 7 epochs and then jointly trained for another 13 epochs. Conversely, when PRW serves as the source domain, we first pre-train on PRW for 2 epochs before conducting 7 epochs of joint training. Detailed hyperparameter analysis is in Section IV-E.

C. Comparison With State-of-the-Art Methods

Since research on unsupervised person search remains relatively limited, we conduct a comprehensive and objective evaluation of our proposed UDA method, RPPS. As shown in Table I, we first compare it against **fully supervised** person search approaches, covering both two-stage and one-stage paradigms. By jointly leveraging frequency-domain feature enhancement and confidence-driven proxy optimization, RPPS significantly improves adaptability to target-domain feature distributions and strengthens cross-domain generalization. Remarkably, RPPS even surpasses several classical fully supervised methods, such as OIM [2] and RCAA [50], on multiple metrics.

It is worth noting that the performance gap between RPPS and the latest fully supervised methods is expected. Fully supervised person search models rely on complete identity and detection annotations from the target domain, enabling them to learn domain-specific appearance and detection patterns directly. In contrast, our UDA setting restricts the use of any target-domain labels, requiring the model to overcome domain shift while simultaneously learning discriminative features,

TABLE I

Quantitative Comparison of Our Proposed UDA Method With State-of-the-Art Fully Supervised Approaches on the CUHK-SYSU and PRW Datasets, Evaluated in Terms of mAP and Top-1 Accuracy. The Gray-Highlighted Row Denotes Our Method

Methods	**Ref**	**Backbone**	**CUHK-SYSU**		**PRW**	
			mAP	top-1	mAP	top-1
Fully supervised Two-step methods:						
IDE [3]	CVPR17	ResNet50	-	-	20.5	48.3
MGTS [19]	ECCV18	VGG16	83.0	83.7	32.6	72.1
RDLR [20]	ICCV19	ResNet50	93.0	94.2	42.9	70.2
TCTS [21]	CVPR20	ResNet50	93.9	95.1	46.8	87.5
OR [49]	TIP21	ResNet50	92.3	93.8	52.3	71.5
Fully supervised One-step methods:						
OIM [2]	CVPR17	ResNet50	75.5	78.7	21.3	49.4
RCAA [50]	ECCV18	ResNet50	79.3	81.3	-	-
CTXG [23]	CVPR19	ResNet50	84.1	86.5	33.4	73.6
NAE+ [22]	CVPR20	ResNet50	92.1	92.9	44.0	81.1
AlignPS+ [51]	CVPR21	ResNet50	94.0	94.5	46.1	82.1
SeqNet [24]	AAAI21	ResNet50	94.8	95.7	47.6	87.6
PSTR [52]	CVPR22	PVTv2-B2	95.2	96.2	56.5	89.7
COAT [25]	CVPR22	ResNet50	94.2	94.7	53.3	87.4
SAT [53]	WACV23	ResNet50	95.3	96.0	55.0	89.2
SOLIDER [26]	CVPR23	Swin-S	95.5	95.8	59.8	86.7
SeqNeXt [54]	WACV23	ConvNeXt-B	96.1	96.5	57.6	89.5
PAD [27]	TPAMI24	ConvNeXt-B	95.9	96.4	58.6	89.9
SEAS [55]	IJCAI24	ConvNeXt-B	97.1	97.8	60.5	89.5
Yang *et al.* [4]	TCSVT24	ResNet50	94.9	95.2	58.3	89.7
GMT [28]	TCSVT25	ResNet50	94.8	95.4	51.3	87.0
PS-DFSI [16]	INFFU25	ResNet50	95.5	95.9	55.2	88.6
RPPS (Ours)	-	ResNet50	80.2	81.2	39.4	81.5

TABLE II

Comparison With Weakly Supervised and Unsupervised Methods on the CUHK-SYSU and PRW Datasets in Terms of mAP and Top-1 Accuracy. The Unsupervised Setting Corresponds to Two Cross-Domain Scenarios: PRW → CUHK-SYSU and CUHK-SYSU → PRW. Bold Indicates the Highest Score in Each Group

Methods	**Ref**	**Backbone**	**CUHK-SYSU**		**PRW**	
			mAP	top-1	mAP	top-1
Weakly supervised:						
R-SiamNet [30]	ICCV21	ResNet50	86.0	87.1	21.4	75.2
CGPS [29]	AAAI22	ResNet50	80.0	82.3	16.2	68.0
SSL [31]	ICCV23	ResNet50	87.4	88.5	30.7	80.6
DICL [32]	PR24	ResNet50	87.4	88.8	35.5	80.9
OLA [56]	AAAI25	ResNet50	87.8	89.3	38.1	82.0
Unsupervised Domain Adaptive:						
DAPS [9]	ECCV22	ResNet50	77.6	79.6	34.7	80.6
DDAM [10]	WACV24	ResNet50	79.5	81.3	36.7	81.2
FOUS [11]	IJCAI24	ResNet50	78.7	80.5	35.4	80.8
DSCA* [12]	AAAI25	ResNet50	79.3	81.2	38.8	80.9
RPPS (Ours)	-	ResNet50	80.2	81.2	39.4	81.5
RPPS (Ours)	-	ViT-B/16	**81.4**	**82.6**	**40.3**	**82.2**

* Reproduced results using the released code.

TABLE III

Ablation Study on CUHK-SYSU and PRW, Evaluating the Effectiveness of DWEM and DCCP Modules in Terms of mAP and Top-1 Accuracy

Components		**CUHK-SYSU**		**PRW**	
DWEM	DCCP	mAP	top-1	mAP	top-1
✗	✗	77.6	79.6	34.7	80.6
✓	✗	79.3	80.7	38.4	81.1
✗	✓	78.8	80.5	37.8	80.9
✓	✓	**80.2**	**81.2**	**39.4**	**81.5**

which is an inherently more challenging task. Therefore, fully supervised approaches should be regarded as an upper performance bound rather than a direct baseline. Despite this limitation, RPPS effectively narrows this gap and demonstrates strong cross-domain generalization under a much more constrained setting.

In Table II, we further compare RPPS with ResNet50 as the backbone against the latest weakly supervised and unsupervised methods. The results show that RPPS consistently outperforms all competing approaches on the PRW dataset. On CUHK-SYSU, RPPS surpasses the most advanced unsupervised method DSCA [12] by 0.9% in mAP and exceeds FOUS [11] by 0.7% in top-1 accuracy. Although RPPS has not yet surpassed weakly supervised methods on CUHK-SYSU, this outcome is understandable. Specifically, the PRW dataset contains significantly fewer images and identities than CUHK-SYSU, leading to inferior detection performance when CUHK-SYSU is treated as the target domain, which in turn affects the accuracy of subsequent search tasks.

It is also worth noting that in the UDA setting (i.e., labeled source and unlabeled target), the availability of source annotations already provides strong discriminative supervision, making the overall performance close to saturation. Consequently, the performance margin among different UDA approaches tends to be small, and further improvements become inherently limited. In this context, RPPS still achieves the highest mAP and competitive top-1 accuracy across both cross-domain directions, demonstrating that our frequency-enhancement design effectively enhances cross-domain generalization rather than relying solely on source-domain supervision.

Beyond the ResNet50 setting used for fair comparison with existing UDA person search methods, we additionally evaluate RPPS with a ViT-B/16 backbone. As reported in the last row of Table II, RPPS with ViT consistently achieves higher mAP and top-1 accuracy on both CUHK-SYSU and PRW. This demonstrates that the proposed frequency-aware enhancement and dynamic proxy regulation are not restricted to convolutional backbones, and can effectively benefit from the stronger global modeling capability of Vision Transformers.

D. Ablation Study

In this section, we present ablation studies to assess the effectiveness of each component in RPPS. For fair comparison, all ablation experiments adopt ResNet50 as the backbone unless otherwise stated. We compare the baseline model with different combinations of the proposed modules and report results on the CUHK-SYSU and PRW datasets. In addition, we conduct a detailed analysis of the subcomponents within the core modules to identify the most critical design choices.

TABLE IV

Ablation Study of DWEM Branches on CUHK-SYSU and PRW Datasets

Methods	CUHK-SYSU		PRW	
	mAP	top-1	mAP	top-1
DWT+IDWT	77.9	79.8	35.5	80.6
DWT+Low-Freq	78.5	80.3	37.3	80.8
DWT+High-Freq	78.8	80.1	37.8	81.0
DWEM	**79.3**	**80.7**	**38.4**	**81.1**

TABLE V

Ablation Study of DCCP Submodules, Showing the Role of Each in Overall Performance

Methods	CUHK-SYSU		PRW	
	mAP	top-1	mAP	top-1
Confidence Only	78.0	79.6	36.0	80.3
Confidence + Online	78.4	80.1	37.2	80.6
Confidence + Offline	78.3	80.2	36.7	80.5
DCCP	**78.8**	**80.5**	**37.8**	**80.9**

TABLE VI

Ablation Study on Different DWT Decomposition Depths

Decomposition Strategy	CUHK-SYSU		PRW	
	mAP	top-1	mAP	top-1
1DWT	78.9	80.1	37.4	80.0
1DWT + LL-DWT (Ours)	**80.2**	**81.2**	**39.4**	**81.5**
Deep-DWT	79.3	80.4	38.6	80.6

1) Ablation on Core Modules: We first evaluate the overall contribution of the core modules by comparing the baseline method with different module combinations. Table III reports the results on the CUHK-SYSU and PRW datasets. When both the DWEM and DCCP modules are incorporated, the model achieves 80.2% mAP and 81.2% top-1 accuracy on CUHK-SYSU, as well as 39.4% mAP and 81.5% top-1 accuracy on PRW. Compared with cases where only a single module or no module is used, the performance improves significantly, clearly demonstrating that the synergy between DWEM and DCCP plays a critical role in enhancing model performance.

2) Ablation of DWEM Branches: Building on the baseline model, we conduct submodule ablation experiments by incrementally adding components of DWEM. Specifically, "DWT+IDWT" refers to applying wavelet decomposition and reconstruction without differentiated enhancement of low-frequency and high-frequency components. "DWT+Low-Freq" retains DWT and the low-frequency enhancement branch while removing the high-frequency branch, whereas "DWT+High-Freq" retains DWT and the high-frequency enhancement branch while discarding the low-frequency branch. As reported in Table IV, the complete DWEM achieves an mAP of 79.3% and a top-1 accuracy of 80.7% on CUHK-SYSU, and an mAP of 38.4% and a top-1 accuracy of 81.1% on PRW. Compared with partial variants, the full design consistently delivers superior performance. These results highlight the importance of jointly leveraging both low-frequency and high-frequency enhancement branches in DWEM, as their synergy more effectively improves cross-domain feature representations.

3) Ablation of DCCP Submodules: Using the baseline model as a foundation, we further conduct detailed submodule ablation studies on the individual components of DCCP. In particular, "Confidence Only" denotes instance selection performed solely through the dynamic confidence mechanism; "Confidence + Online" integrates dynamic confidence with confidence-weighted online updates while deliberately omitting offline stratification and correction; and "Confidence + Offline" combines dynamic confidence with offline stratification and asynchronous correction while intentionally excluding online updates. As reported in Table V, the complete DCCP achieves an mAP of 78.8% and a top-1 accuracy of 80.5% on CUHK-SYSU, and an mAP of 37.8% and a top-1 accuracy of 80.9% on PRW. Compared with these partial variants, the full design consistently delivers more stable and superior results, clearly underscoring the importance of jointly leveraging confidence stratification, online updates, and offline correction to refine pseudo-label quality and ultimately enhance model performance in unsupervised cross-domain scenarios.

4) Ablation of DWT Depths: To validate the decomposition depth design in DWEM, we conduct ablation experiments under different configurations. Specifically, we examine three variants: "1DWT" which applies a single-level DWT on F_{in}; "1DWT + LL-DWT" which performs an additional decomposition on the low-frequency component (LL) after the initial step; and "Deep-DWT" which applies deeper decompositions on the low-frequency branch while keeping other settings unchanged. As shown in Table VI, single-level decomposition is insufficient to fully exploit frequency-domain features, while deeper decomposition causes a severe loss of high-frequency details. In contrast, our proposed "1DWT + LL-DWT" achieves the best performance on both benchmark datasets, effectively balancing global structural information and local texture details. These results confirm the rationality of the DWEM design and highlight its role in enhancing the robustness of person search.

E. Hyperparameter Analysis

Several hyperparameters in RPPS critically influence model performance. We evaluate four key parameters: the balance coefficient α, the confidence threshold θ, the consistency threshold δ, and the global confidence threshold ε.

1) Analysis of Balance Coefficient αAlpha: The balance coefficient α controls the relative contributions of bounding-box IoU and feature entropy in the dynamic confidence function, thereby influencing the selection of high-quality instances. As shown in Fig. 3, when $\alpha = 0.3$, excessive weight on feature entropy misclassifies well-localized but slightly dispersed instances as low-quality, yielding an mAP of 38.4% on PRW. Conversely, when $\alpha = 0.7$, IoU dominates, failing to suppress background-similar redundancies and reducing CUHK-SYSU mAP to 79.7%. The best balance is achieved at $\alpha = 0.6$, which optimally trades off localization precision and feature certainty, leading to superior results on both datasets.

2) Analysis of Confidence Threshold θTheta: The confidence threshold θ determines which high-confidence instances

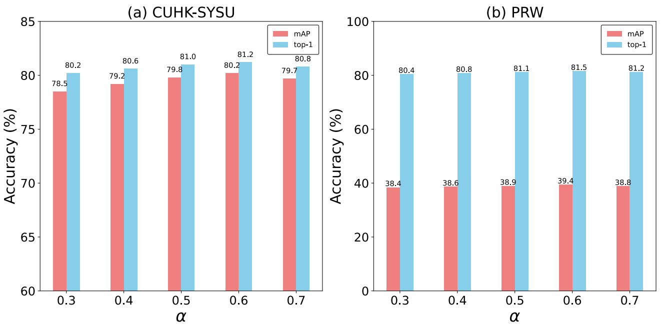


Fig. 3. Impact of balance coefficient α on CUHK-SYSU and PRW datasets.

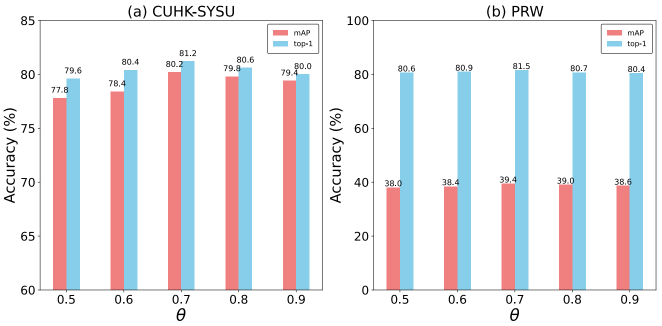


Fig. 4. Impact of threshold θ on CUHK-SYSU and PRW datasets.

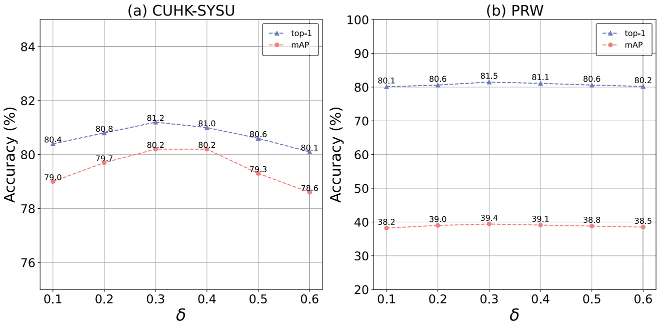


Fig. 5. Impact of threshold δ on CUHK-SYSU and PRW datasets.

are retained for proxy initialization. As shown in Fig. 4, setting $\theta = 0.5$ is too permissive, admitting many noisy samples, which lowers the PRW mAP to 38.0%. In contrast, setting $\theta = 0.9$ is overly strict, discarding too many valid instances, causing CUHK-SYSU top-1 accuracy to drop significantly to 79.4%. The optimal value of $\theta = 0.7$ strikes a balance between effectively filtering noisy data and retaining a sufficient number of high-quality samples, thereby delivering the best overall performance across both benchmark datasets.

3) Analysis of Consistency Threshold δDelta: The consistency threshold δ specifies whether the confidence gap between an instance and its proxy satisfies the criterion for strong updates. As shown in Fig. 5, when $\delta = 0.1$, the rule is overly restrictive, discarding useful instances and causing insufficient proxy updates, with PRW achieving 38.2% mAP. At $\delta = 0.6$, the rule is too lenient, allowing noisy instances to be wrongly updated and reducing CUHK-SYSU top-1 accuracy to 80.1%. The balanced setting $\delta = 0.3$ yields the most stable performance, effectively distinguishing reliable samples from noisy ones.

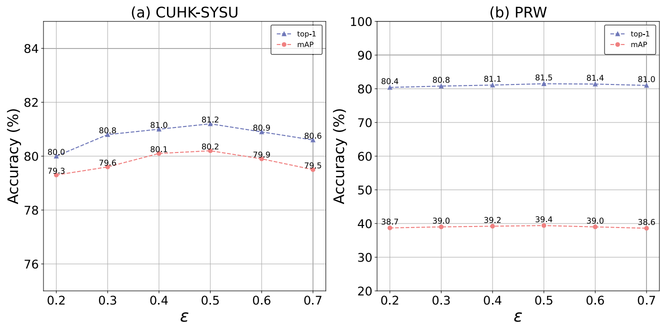


Fig. 6. Impact of threshold ε on CUHK-SYSU and PRW datasets.

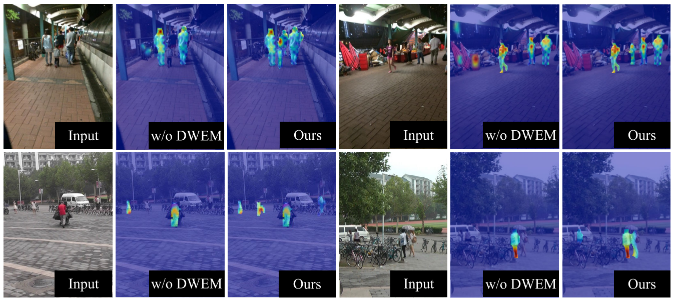


Fig. 7. Attention visualization with Layer-CAM on CUHK-SYSU (top row) and PRW (bottom row). Without DWEM, attention is scattered and spills into background regions; in contrast, it is concentrated on pedestrian areas, reducing irrelevant responses. Red regions indicate higher attention scores.

4) Analysis of Global Confidence Threshold εVarepsilon: The global confidence threshold ε is applied during the offline stage for proxy stratification, which determines whether low-confidence proxies should be replaced. As shown in Fig. 6, when $\varepsilon = 0.2$, the threshold is too low, allowing noisy proxies to persist and consequently reducing the PRW mAP to 37.8%. In contrast, when $\varepsilon = 0.7$, the threshold is too high, leading to the premature replacement of moderately reliable proxies, which causes information loss and reduces the CUHK-SYSU mAP to 79.5%. The optimal setting $\varepsilon = 0.5$ achieves an effective balance between preserving reliable proxies and eliminating unreliable ones, thus producing the most efficient stratification and improving overall model performance.

5) Sensitivity Discussion Across Datasets: It is worth noting that although the four hyperparameters exhibit different optimal values, their overall trends remain highly consistent across both CUHK-SYSU and PRW, as shown in Figs. 3–6. More importantly, each parameter demonstrates a broad stable region rather than a sharply peaked optimum. For example, α performs consistently well within 0.5-0.7 on both datasets, and similar stable intervals can be observed for θ (0.6-0.8), δ (0.2-0.4), and ε (0.4-0.6). These overlapping stable ranges indicate that RPPS is not overly sensitive to dataset variations and does not rely on dataset-specific tuning. This stability suggests that the proposed design confi-

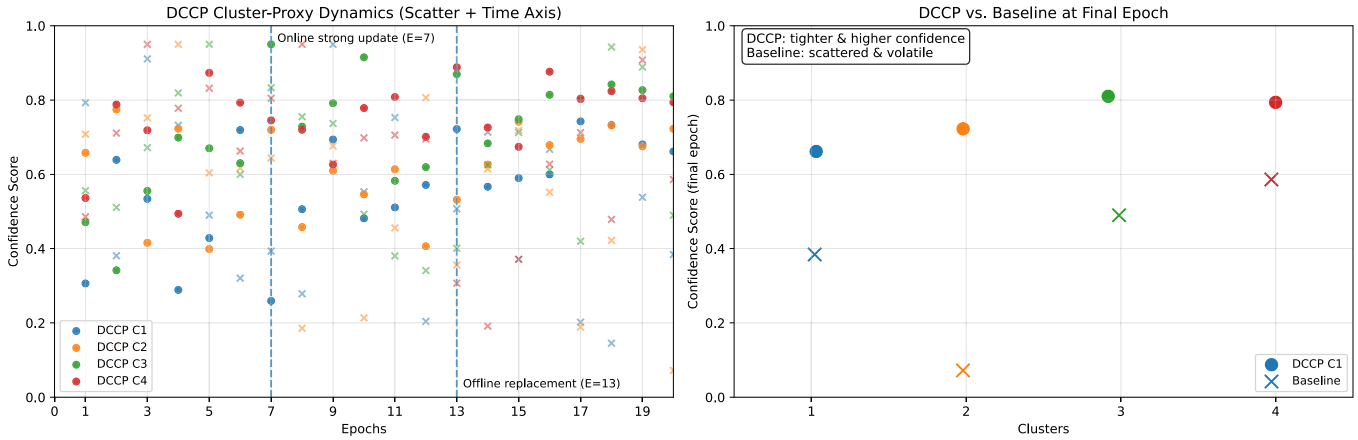


Fig. 8. Visualization of proxy dynamics under DCCP. Left: confidence distributions across epochs, showing improved stability with online updates at epoch 7 and offline replacements at epoch 13. Right: final-epoch comparison, where DCCP yields tighter, more reliable proxies than the scattered baseline.

TABLE VII
Comparison of Performance of Different Wavelet Basis Methods on the PRW Dataset

Methods	GFLOPs	Params	PRW	
			mAP	top-1
Daubechies Wavelet [58]	2.8	5.8M	38.2	79.3
Symlet Wavelet [59]	2.1	4.2M	38.9	81.1
Fourier Transform [60]	3.2	1.5M	37.8	78.6
Haar Wavelet (Ours)	1.5	2.3M	39.4	81.5

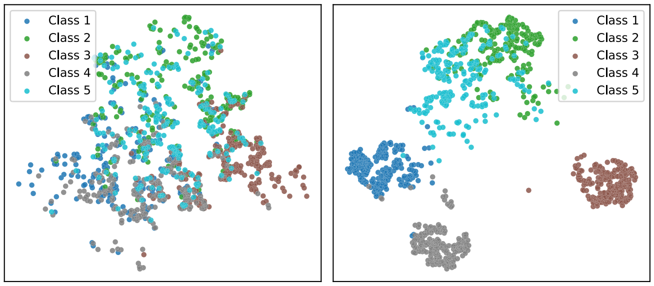
dence reliability, and proxy consistency-generalize well across different domains, reinforcing the practical applicability of RPPS.

F. Effectiveness Analysis of DWEM

To thoroughly assess the effectiveness of DWEM, we employ Layer-CAM [57] to visualize the output of the final layer in the backbone network. As illustrated in Fig. 7, without DWEM, the model's attention is scattered and easily leaked into background regions, making it susceptible to noise interference. In contrast, incorporating DWEM strengthens pedestrian contours through low-frequency enhancement and emphasizes clothing textures via high-frequency enhancement, enabling the model to focus more precisely on pedestrian regions while capturing fine discriminative details. On CUHK-SYSU (top row), DWEM improves the model's focus on target pedestrians, whereas on PRW (bottom row), it helps maintain stable attention under complex backgrounds. Overall, DWEM effectively suppresses cross-domain noise and highlights target-specific features, thereby enhancing both the robustness and reliability of person search representations.

G. Comparison of Different Wavelet Basis Methods

In the DWEM, we adopt the Haar wavelet basis for high–low frequency decomposition. To evaluate its effectiveness, we compare it with Daubechies, Symlet, and the Fourier transform on the PRW dataset, as shown in Table VII. The Haar wavelet provides the best trade-off between computational efficiency and model accuracy. While Daubechies [58] enhances high-frequency discriminability, its complexity and resource demands limit practical use. Symlet [59] shows moderate complexity but lower performance than Haar, with limited adaptability in frequency-domain processing. The Fourier transform [60] emphasizes global frequency analysis, but weak localization and high cost lead to poor performance in person search. Overall, the Haar wavelet achieves the most balanced performance in frequency-domain enhancement and decomposition, making it well-suited for person search tasks.



(a) Baseline (b) RPPS

Fig. 9. t-SNE visualization of target-domain feature embeddings. Compared to the baseline, RPPS yields more compact intra-class clusters and clearer inter-class separations.

H. Effectiveness Analysis of DCCP

To evaluate the effectiveness of the DCCP module in cross-domain adaptation, we visualize the dynamic updating of clustering proxies during training. As shown in Fig. 8, the left panel illustrates the global confidence distribution of proxies across epochs using a scatter–timeline plot, where different colors represent different clusters. As training progresses, most DCCP proxies gradually converge toward higher confidence regions (> 0.7). A sharp rise occurs at epoch 7 during the online strong update stage, followed by further stabilization at epoch 13 when low-confidence proxies are replaced with more reliable ones through offline

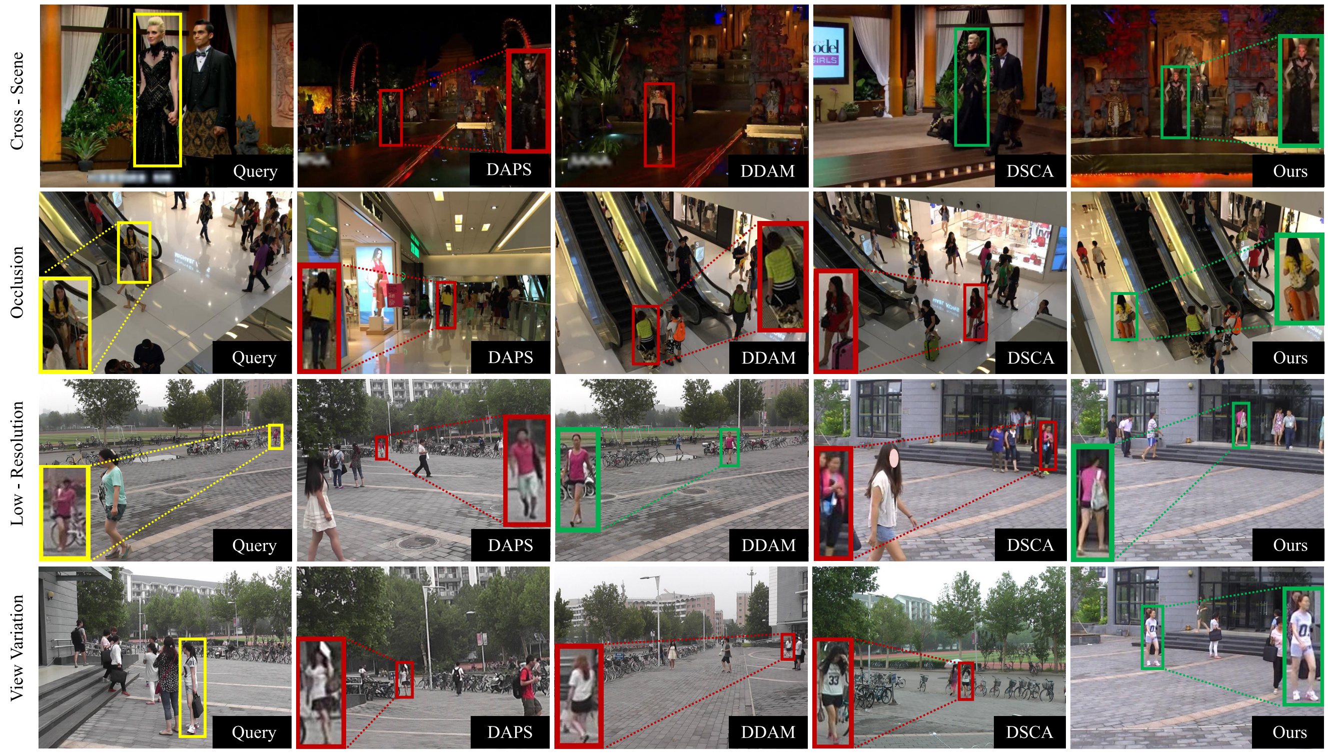


Fig. 10. Qualitative analysis on CUHK-SYSU (first two rows) and PRW (last two rows). Top-1 retrieval results of different UDA person search methods are presented. Our method effectively reduces domain gaps and achieves more accurate retrieval. Yellow boxes mark query persons, while green and red boxes denote correct and incorrect matches, respectively.

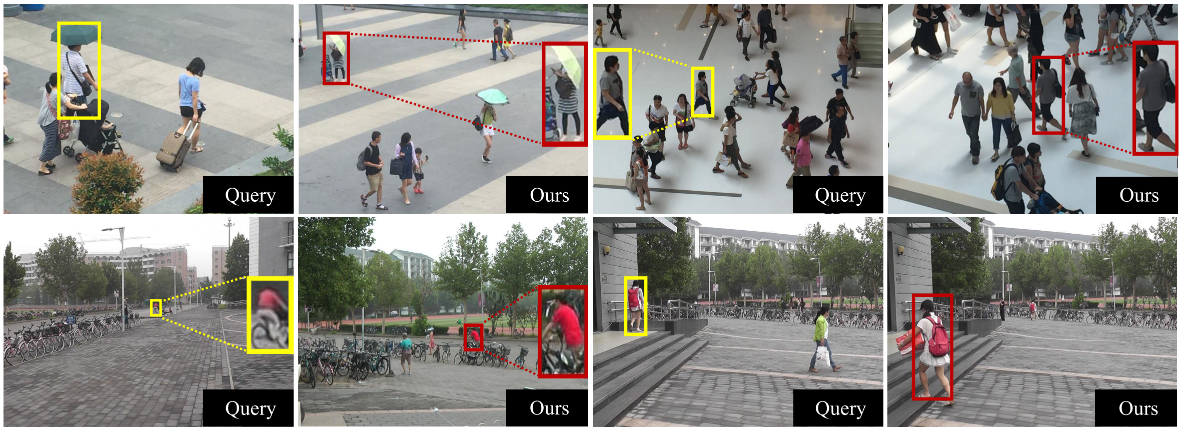


Fig. 11. Examples of failure cases on CUHK-SYSU (first row) and PRW (second row), where domain shifts and complex conditions lead to incorrect matches.

replacement. In contrast, the baseline DAPS [9] model (right panel) shows scattered proxies with large confidence fluctuations, failing to form stable cluster centers. These results demonstrate that DCCP effectively suppresses noisy proxies and enhances robustness through its dynamic updating strategy.

I. Computational Complexity

Finally, we compare our method with the latest weakly supervised and unsupervised approaches in terms of GFLOPs, parameters, and per-epoch training time. We further break down the computational cost of RPPS. The DWEM module introduces additional DWT and inverse DWT operations; however, since these operations are applied at intermediate feature levels rather than on high-resolution input images, they mainly involve lightweight filtering and element-wise computations. As a result, their overhead is significantly lower than complex spatial-domain transformations, keeping the overall computational cost manageable. The DCCP module introduces confidence estimation and proxy updates during training, but avoids large-scale matrix operations and only involves small-scale vector updates, resulting in a minor impact on overall complexity. Notably, DCCP is only activated during training and is completely removed during inference. Therefore, RPPS introduces no additional inference-time computation or GPU memory overhead.

TABLE VIII
COMPARISON OF MODEL COMPLEXITY AND TRAINING COST WITH WEAKLY SUPERVISED AND UNSUPERVISED METHODS ON PRW DATASET

Methods	**GFLOPs**	**Params**	**Time**	**PRW**	
				mAP	top-1
CGPS [29]	281	49.2M	1.1h	16.2	80.6
R-SiamNet [30]	320	52.5M	1.0h	21.4	75.2
DAPS [9]	774	53.4M	1.6h	34.7	80.6
DDAM [10]	784	56.4M	1.6h	36.7	81.2
DSCA [12]	776	56.2M	1.2h	38.8	80.9
RPPS (Ours)	755	57.3M	1.3h	39.4	81.5

Experiments are conducted under the same setting using an NVIDIA Tesla V100 GPU with an input image size of 1500 × 900. As shown in Table VIII, RPPS requires 755 GFLOPs, which is lower than the main baseline DAPS [9] and DDAM [10]. The parameter count remains within a reasonable range, and the per-epoch training time is comparable to existing methods. Overall, RPPS maintains inference-time latency and GPU memory usage comparable to DAPS, while achieving superior performance on the PRW dataset.

J. Analysis of Visualized Feature Distribution

To further illustrate the effectiveness of RPPS in the target domain, we visualize the feature distribution using t-SNE [61]. Specifically, we extract image-level embeddings from the ReID branch and compare the Baseline with the complete RPPS model. As shown in Fig. 9, the Baseline features exhibit significant overlap among different categories, with blurred inter-class boundaries and weak discriminability. In contrast, RPPS produces more compact intra-class clusters and clearer inter-class separations, benefiting from the frequency robustness reinforced by DWEM and the stabilized proxy optimization introduced by DCCP. These results demonstrate that RPPS effectively enhances the discriminability of target-domain features, alleviates feature confusion caused by domain shifts and proxy noise, and provides more reliable representations for subsequent identification.

K. Qualitative Performance

As shown in Fig. 10, we present qualitative comparisons on the CUHK-SYSU and PRW datasets under a variety of challenging conditions, including cross-scene variations, occlusion, low resolution, and significant viewpoint changes. Compared with DAPS [9], DDAM [10], and DSCA [12], our method more accurately localizes and identifies query persons in complex environments. In particular, RPPS demonstrates superior robustness in scenes with heavy background clutter, partial occlusion, and long-range cross-camera matching, highlighting its advantage in cross-domain feature alignment and discriminability preservation. It is worth noting that although our evaluation is conducted on the standard CUHK-SYSU and PRW benchmarks, these datasets inherently involve diverse and realistic domain shifts. Specifically, they cover different camera types, indoor and outdoor environments, illumination variations (e.g., daytime and nighttime), occlusion, low-resolution pedestrians, and large viewpoint changes. Under these heterogeneous conditions, RPPS consistently achieves more reliable retrieval results, indicating strong robustness to practical cross-domain variations and supporting its generalization capability in real-world person search applications.

Furthermore, Fig. 11 presents representative failure cases. These failures mainly occur under extreme domain shifts and highly complex scenes. For example, severe low-resolution conditions lead to the loss of informative texture details, weakening high-frequency discriminative cues, while heavy occlusion or high visual similarity among pedestrians reduces the distinctiveness of proxy representations, making it difficult to maintain stable embeddings. These cases highlight remaining challenges and suggest potential directions for future improvement.

V. CONCLUSION

In this paper, we address the challenge of unreliable pseudo-supervision in UDA person search, which primarily stems from spectral shift bias and static proxy updates. To overcome these issues, we propose the Reliable Pseudo-supervision in UDA Person Search (RPPS) framework, which comprises two key modules: the Dual-branch Wavelet Enhancement Module (DWEM) for frequency-domain enhancement and the Dynamic Confidence-weighted Clustering Proxy (DCCP) for robust proxy optimization. Extensive experiments on CUHK-SYSU and PRW demonstrate that RPPS achieves state-of-the-art performance and strong robustness, confirming the effectiveness of integrating frequency-domain enhancement with dynamic proxy optimization.

Limitations and Future Work: Despite its strong performance, RPPS still faces two limitations: (i) DWEM may retain domain-specific noise under extreme style shifts (e.g., indoor vs. outdoor), reducing discriminability; and (ii) applying DWT to full-resolution feature maps incurs high computational overhead, limiting real-time deployment. Future work will focus on incorporating domain-aware adaptive weighting and developing lightweight frequency decomposition methods to further improve robustness and efficiency.

REFERENCES

[1] Y. Xu, B. Ma, R. Huang, and L. Lin, "Person search in a scene by jointly modeling people commonness and person uniqueness," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 937–940, doi: 10.1145/2647868.2654965.

[2] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3415–3424.

[3] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1367–1376.

[4] X. Yang, M. Tian, N. Wang, and X. Gao, "Unleashing the feature hierarchy potential: An efficient tri-hybrid person search model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 11, pp. 11551–11563, Nov. 2024, doi: 10.1109/TCSVT.2024.3424261.

[5] S. Li, J. Leng, C. Kuang, M. Tan, and X. Gao, "Video-level language-driven video-based visible-infrared person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 5505–5520, 2025, doi: 10.1109/TIFS.2025.3573672.

[6] H. Li, Y. Mao, Y. Zhang, G. Qi, and Z. Yu, “Domain-adaptive person re-identification without cross-camera paired samples,” *Eng. Appl. Artif. Intell.*, vol. 145, Apr. 2025, Art. no. 110171, doi: 10.1016/j.engappai.2025.110171.

[7] H. Li, Y. Liu, Y. Zhang, J. Li, and Z. Yu, “Breaking the paired sample barrier in person re-identification: Leveraging unpaired samples for domain generalization,” *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 2357–2371, 2025, doi: 10.1109/TIFS.2025.3543040.

[8] S. Yan, N. Dong, L. Zhang, and J. Tang, “CLIP-driven fine-grained textimage person re-identification,” *IEEE Trans. Image Process.*, vol. 32, pp. 6032–6046, 2023, doi: 10.1109/TIP.2023.3327924.

[9] J. Li, Y. Yan, G. Wang, F. Yu, Q. Jia, and S. Ding, “Domain adaptive person search,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 302–318.

[10] M. K. Almansoori, M. Fiaz, and H. Cholakkal, “DDAMPS: Diligent domain adaptive mixer for person search,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 6674–6683.

[11] S. Cai, Y. Mo, L. Peng, Y. Xie, T. Tong, and X. Zhu, “MCD-CLIP: Multiview chest disease diagnosis with disentangled CLIP,” in *Proc. Thirty-ThirdInternational Joint Conf. Artif. Intell.*, Aug. 2024, pp. 702–710, doi: 10.24963/ijcai.2024/79.

[12] L. Qi, H. Wang, J. Zhang, J. Peng, and Y. Wang, “Unsupervised domain adaptive person search via dual self-calibration,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, 2025, pp. 6550–6558, doi: 10.1609/aaai.v39i6.32702.

[13] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness,” in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–22.

[14] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, “Leveraging frequency analysis for deep fake image recognition,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3247–3258.

[15] M. Vishwanath and R. M. Owens, “A common architecture for the DWT and IDWT,” in *Proc. Int. Conf. Appl. Specific Syst., Architectures Processors*, Aug. 1996, pp. 193–198, doi: 10.1109/ASAP.1996.542814.

[16] Q. Zhang et al., “Dynamic frequency selection and spatial interaction fusion for robust person search,” *Inf. Fusion*, vol. 124, Dec. 2025, Art. no. 103314, doi: 10.1016/j.inffus.2025.103314.

[17] Q. Zhang et al., “Learning adaptive shift and task decoupling for discriminative one-step person search,” *Knowledge-Based Syst.*, vol. 304, Nov. 2024, Art. no. 112483, doi: 10.1016/j.knosys.2024.112483.

[18] Q. Zhang, J. Wu, D. Miao, C. Zhao, and Q. Zhang, “Attentive multigranularity perception network for person search,” *Inf. Sci.*, vol. 681, Oct. 2024, Art. no. 121191, doi: 10.1016/j.ins.2024.121191.

[19] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, “Person search via a mask-guided two-stream CNN model,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 764–781.

[20] C. Han et al., “Re-ID driven localization refinement for person search,” in *Proc. Int. Conf. Comput. Vis.*, Oct. 2019, pp. 9814–9823.

[21] C. Wang, B. Ma, H. Chang, S. Shan, and X. Chen, “TCTS: A task-consistent two-stage framework for person search,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11949–11958.

[22] D. Chen, S. Zhang, J. Yang, and B. Schiele, “Norm-aware embedding for efficient person search,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12612–12621.

[23] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, “Learning context graph for person search,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2153–2162.

[24] Z. Li and D. Miao, “Sequential end-to-end network for efficient person search,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 2011–2019, doi: 10.1609/aaai.v35i3.16297.

[25] R. Yu et al., “Cascade transformers for end-to-end person search,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7257–7266.

[26] W. Chen et al., “Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15050–15061.

[27] H. Kim, J. Lee, and K. Sohn, “Prototype-guided attention distillation for discriminative person search,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 1, pp. 99–115, Jan. 2025, doi: 10.1109/TPAMI.2024.3461778.

[28] B. Cai, H. Wang, M. Yao, and X. Fu, “Focus more on what? Guiding multi-task training for end-to-end person search,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 7, pp. 7266–7278, Jul. 2025, doi: 10.1109/TCSVT.2025.3540089.

[29] Y. Yan et al., “Exploring visual context for weakly supervised person search,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, pp. 3027–3035, Jun. 2022, doi: 10.1609/aaai.v36i3.20209.

[30] C. Han et al., “Weakly supervised person search with region Siamese networks,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11986–11995.

[31] B. Wang, Y. Yang, J. Wu, G.-J. Qi, and Z. Lei, “Selfsimilarity driven scale-invariant learning for weakly supervised person search,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1813–1822.

[32] J. Wang, Y. Pang, J. Cao, H. Sun, Z. Shao, and X. Li, “Deep intra-image contrastive learning for weakly supervised one-step person search,” *Pattern Recognit.*, vol. 147, Mar. 2024, Art. no. 110047, doi: 10.1016/j.patcog.2023.110047.

[33] Y. Huang, Q. Wu, J. Xu, and Y. Zhong, “SBSGAN: Suppression of inter-domain background shift for person re-identification,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9527–9536.

[34] D. H. Pham, A. D. Nguyen, and H. N. Nguyen, “GAN-based data augmentation and pseudo-label refinement with holistic features for unsupervised domain adaptation person re-identification,” *Knowledge-Based Syst.*, vol. 288, Mar. 2024, Art. no. 111471, doi: 10.1016/j.knosys.2024.111471.

[35] Y. Zou, X. Yang, Z. Yu, B. V. K. V. Kumar, and J. Kautz, “Joint disentangling and adaptation for cross-domain person re-identification,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 87–104.

[36] Y. Ge, D. Chen, F. Zhu, R. Zhao, and H. Li, “Self-paced contrastive learning with hybrid memory for domain adaptive object re-ID,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 11309–11321.

[37] D. Wang and S. Zhang, “Unsupervised person re-identification via multilabel classification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10978–10987.

[38] Y. Cho, J. Jung, J. Kim, W. Kim, and S.-E. Yoon, “Generalizable person re-identification via balancing alignment and uniformity,” in *Proc. Adv. Neural Inf. Process. Syst. 37*, 2024, pp. 47069–47093.

[39] C. Zou, Z. Chen, Z. Cui, Y. Liu, and C. Zhang, “Discrepant and multi-instance proxies for unsupervised person reidentification,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 11024–11034.

[40] S. Li et al., “Logical relation inference and multiview information interaction for domain adaptation person re-identification,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 14770–14782, Oct. 2024, doi: 10.1109/TNNLS.2023.3281504.

[41] Y. Zhang, L. Kong, H. Li, and J. Wen, “Weakly supervised visible-infrared person re-identification via heterogeneous expert collaborative consistency learning,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2025, pp. 12659–12669.

[42] J. Leng, C. Kuang, S. Li, J. Gan, H. Chen, and X. Gao, “Dual-space video person re-identification,” *Int. J. Comput. Vis.*, vol. 133, no. 6, pp. 3667–3688, Jun. 2025, doi: 10.1007/s11263-025-02350-5.

[43] L. Zhou, S. Li, N. Dong, Y. Tai, Y. Zhang, and H. Li, “Hierarchical prompt learning for image-and text-based person re-identification,” 2025, *arXiv:2511.13575*.

[44] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[45] S. Ren, K. He, R. Girshick, and J. Sun, “Faster RCNN: Towards real-time object detection with region proposal networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 39, 2016, pp. 1137–1149.

[46] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[47] M. Ester, H. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proc. kdd*, 1996, pp. 226–231.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[49] H. Yao and C. Xu, “Joint person objectness and repulsion for person search,” *IEEE Trans. Image Process.*, vol. 30, pp. 685–696, 2021, doi: 10.1109/TIP.2020.3038347.

[50] X. Chang, P.-Y. Huang, Y.-D. Shen, X. Liang, Y. Yang, and A. G. Hauptmann, “RCAA: Relational context-aware agents for person search,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 86–102.

[51] Y. Yan et al., "Anchor-free person search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7686–7695.

[52] J. Cao et al., "PSTR: End-to-end one-step person search with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9448–9457.

[53] M. Fiaz, H. Cholakkal, R. M. Anwer, and F. Shahbaz Khan, "SAT: Scale-augmented transformer for person search," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4809–4818.

[54] L. Jaffe and A. Zakhor, "Gallery filter network for person search," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1684–1693.

[55] Y. Jiang, H. Wang, J. Peng, X. Fu, and Y. Wang, "Scene-adaptive person search via bilateral modulations," in *Proc. Int. Joint Conf. Artif. Intell.*, Aug. 2024, pp. 938–946.

[56] H. Zhu, X. Yang, and N. Wang, "Optimizing label assignment for weakly supervised person search," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, Apr. 2025, pp. 10941–10949, doi: 10.1609/aaai.v39i10.33189.

[57] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," *IEEE Trans. Image Process.*, vol. 30, pp. 5875–5888, 2021, doi: 10.1109/TIP.2021.3089943.

[58] C. Vonesch, T. Blu, and M. Unser, "Generalized Daubechies wavelet families," *IEEE Trans. Signal Process.*, vol. 55, no. 9, pp. 4415–4429, Sep. 2007, doi: 10.1109/TSP.2007.896255.

[59] E. Noskova and D. Tumakov, "Analysis of wavelet transform application for filtering real ECG signals from high-frequency noise," in *Proc. 26th Int. Conf. Digit. Signal Process. Appl. (DSPA)*, Mar. 2024, pp. 1–5, doi: 10.1109/dspa60853.2024.10510072.

[60] W. Fan et al., "FilterNet: Harnessing frequency filters for time series forecasting," in *Proc. Adv. Neural Inf. Process. Syst. 37*, 2024, pp. 55115–55140.

[61] L. V. D. Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

Qixian Zhang received the B.S. degree from Heilongjiang Institute of Technology, Harbin, China, in 2018, and the M.S. degree from Jilin University, Changchun, China, in 2021. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Tongji University, Shanghai, China. He has published multiple papers in leading journals, including IEEE TRANSACTIONS ON IMAGE PROCESSING, *Information Fusion*, *Knowledge-Based Systems*, and *Information Sciences*. His research interests include person search, person re-identification, and computer vision.

Duoqian Miao is currently a Professor and a Ph.D. Supervisor with the School of Computer Science and Technology, Tongji University, Shanghai, China. He has published more than 200 articles in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON FUZZY SYSTEMS, *Information Sciences*, and *Knowledge-Based Systems*. His research interests include machine learning, data mining, big data analytics, granular computing, artificial intelligence, and text and image processing. His representative awards include the Second Prize of Wu Wenjun Artificial Intelligence Science and Technology Award in 2018, the First Prize of the Chongqing Natural Science Award in 2010, the First Prize of the Shanghai Technical Invention Award in 2009, and the First Prize of the Ministry of Education Science and Technology Progress Award in 2007. He serves as an Associate Editor for *International Journal of Approximate Reasoning*, *Information Sciences*, and *CAAI Transactions on Intelligence Technology*. He is a fellow of the International Rough Set Society (IRSS) and the Chinese Association for Artificial Intelligence (CAAI). He serves as the President of the International Rough Set Society, the Chair of the CAAI Technical Committee on Granular Computing and Knowledge Discovery, the Vice Director of the MOE Key Laboratory of Embedded System and Service Computing, and the Vice President of Shanghai Computer Federation and Shanghai Association for Artificial Intelligence.

Qi Zhang received the Ph.D. degree from Beijing Institute of Technology, Beijing, China, and the University of Technology Sydney, Sydney, NSW, Australia, in 2020. He is currently an Associate Professor with the School of Computer Science and Technology, Tongji University, Shanghai, China. He has authored high-quality papers in premier conferences and journals, including AAAI, IJCAI, SIGIR, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and *ACM Transactions on Information Systems*. His research interests include collaborative filtering, sequential recommendation, learning to hash, and multivariate time series (MTS) analysis.

Xuan Tan received the B.S. degree from Anhui Agricultural University, Hefei, China, in 2021, and the M.S. degree from Southwest Jiaotong University, Chengdu, China, in 2025. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Tongji University, Shanghai, China. He has published multiple papers in leading journals, including IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. His research interests include person re-identification, multimodal learning, and unsupervised learning.

Hongyun Zhang received the Ph.D. degree in pattern recognition and intelligent systems from Tongji University, Shanghai, China, in 2005. She is currently an Associate Professor with the School of Computer Science and Technology, Tongji University, Shanghai, China. She is the author or co-author of nearly 60 journal and conference papers in *Principal Curves*, *Pattern Recognition*, *Machine Learning*, *Granular Computing*, and *Rough Sets*. Her current research interests include principal curves, pattern recognition, data mining, and image retrieval.

Cairong Zhao (Senior Member, IEEE) received the B.S. degree from Jilin University, Changchun, China, in 2003, the M.S. degree from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, in 2006, and the Ph.D. degree from Nanjing University of Science and Technology, Nanjing, China, in 2011. He is currently a Professor and a Ph.D. Supervisor with the School of Computer Science and Technology, Tongji University, Shanghai, China. He was the Chair of the Computer Vision Technical Committee of Shanghai Computer Society. He is a Distinguished Member of China Computer Federation (CCF), and he has been selected for the "Dongfang Yingcai (Top-notch)" Program. He serves on the editorial boards of *Pattern Recognition* and *Journal of Image and Graphics*, and has served as a Guest Editor for IEEE TRANSACTIONS ON MULTIMEDIA. He has published more than 50 papers in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON MULTIMEDIA, *International Journal of Computer Vision*, and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, as well as top-tier conferences including CVPR, ICML, NeurIPS, and ICLR. His research interests include computer vision, pattern recognition, and visual surveillance. His Representative awards include the First Prize of the Shanghai Science and Technology Progress Award in 2022, ranked 4/13 and the Second Prize of Shanghai Natural Science Award in 2023, ranked 1/4.