Information Fusion 124 (2025) 103314

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Dynamic frequency selection and spatial interaction fusion for robust person search ☆

Qixian Zhang [a,b], Duoqian Miao [a,b,*], Qi Zhang [a,b], Cairong Zhao [a,b], Hongyun Zhang [a,b], Ye Sun [c], Ruizhi Wang [a,b]

[a] School of Computer Science and Technology, Tongji University, Shanghai 201804, China
[b] Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China
[c] School of Computer Science, Fudan University, Shanghai 200438, China

ARTICLE INFO

Keywords:
Person search
Frequency decoupling
Feature fusion
Frequency selection
Spatial interaction

ABSTRACT

Person search aims to locate target individuals in large image databases captured by multiple non-overlapping cameras. Existing models primarily rely on spatial feature extraction to capture fine-grained local details, which is vulnerable to background clutter and occlusions and leads to unstable feature representations. To address the issues, we propose a Dynamic Frequency Selection and Spatial Interaction Fusion Network (PS-DFSI), marking the first attempt to introduce frequency decoupling and selection into person search. By integrating frequency and spatial features, PS-DFSI enhances feature expressiveness and robustness. Specifically, it comprises two core modules: the Dynamic Frequency Selection Module (DFSM) and the Spatial Frequency Interaction Module (SFIM). DFSM decouples feature maps into low-frequency and high-frequency components using learnable low-pass and high-pass filters, and a frequency selection modulator emphasizes key frequency components via channel attention. SFIM refines local details by fusing frequency-enhanced features with high-level semantic representations, leveraging multi-scale receptive fields and cross-feature attention for efficient spatial-frequency integration. Extensive experiments on CUHK-SYSU and PRW demonstrate that PS-DFSI significantly improves person search performance, validating its effectiveness and robustness.

1. Introduction

Person search [1–6] aims to locate and identify individuals across images from different camera views, involving two sub-tasks: person detection [7] and person re-identification (ReID). Person detection locates individuals and generates bounding boxes (BBoxes), while ReID [8–13] matches the same individual across cameras, ensuring consistent identity. With applications in surveillance, public safety, and crime tracking [14], person search has become a key research area in computer vision.

Existing methods for person search are categorized into two-step [15,16] and one-step approaches [17,18]. The two-step method first detects individuals and generates bounding boxes, then uses a re-identification network for matching. Although effective, it is less computationally efficient and consumes more resources. In contrast, the one-step method uses a shared feature extraction network for both localization and re-identification, reducing computation redundancy and improving its suitability for large-scale video data and real-time surveillance applications. Despite the efficiency of existing one-step methods, they typically rely on neural networks, e.g., Convolutional Neural Networks (CNNs), to capture spatial domain features. Their pixel-level modeling is vulnerable to noise, complex backgrounds, and occlusions and reduces feature robustness and discriminability. As illustrated in Fig. 1(a), complex backgrounds often lead to confusion with person features, while occlusions resulting in missing appearance features severely reduce feature discriminability. Additionally, existing spatial domain methods struggle to capture cross-scale details and global information, limiting adaptability in complex scenarios such as crowded or partially occluded environments.

Fortunately, frequency modeling provides a novel perspective for addressing challenges in person search, which capably decomposes image features into different frequency components to capture discriminative and robust features. Low-frequency components capture global structural information, facilitating an understanding of the overall layout of the image. Meanwhile, high-frequency components focus

☆ Code is accessible at https://github.com/zqx951102/DFSI.

* Corresponding author.
E-mail addresses: zhangqx@tongji.edu.cn (Q. Zhang), dqmiao@tongji.edu.cn (D. Miao), zhangqi_cs@tongji.edu.cn (Q. Zhang), zhaocairong@tongji.edu.cn (C. Zhao), zhanghongyun@tongji.edu.cn (H. Zhang), yesun23@m.fudan.edu.cn (Y. Sun), ruizhiwang@tongji.edu.cn (R. Wang).

https://doi.org/10.1016/j.inffus.2025.103314
Received 25 February 2025; Received in revised form 5 May 2025; Accepted 7 May 2025
Available online 28 May 2025

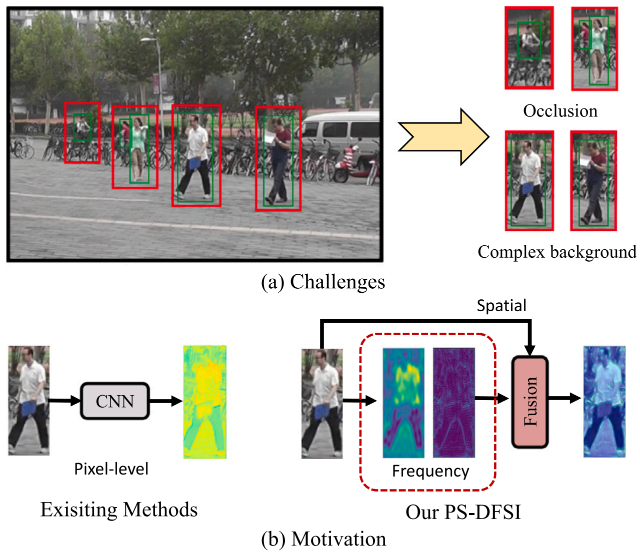

Fig. 1. Illustrations of person search challenges and the motivation behind the proposed PS-DFSI. (a) Challenges: The image shows difficulties caused by occlusion and complex backgrounds. (b) Motivation: Existing methods rely on pixel-level feature extraction (left), while PS-DFSI integrates spatial and frequency features (right) to improve robustness.

on local details, such as textures and edges, improving the accuracy of pedestrian identification. In addition, frequency modeling exhibits strong noise suppression capabilities, effectively filtering out high-frequency noise while preserving valuable low-frequency information, thus enhancing noise resistance. Naturally, low-frequency components in complex scenarios (e.g., occlusion) help extract unobstructed global information, while high-frequency components capture unobstructed local details, guaranteeing recognition accuracy. Although frequency modeling has shown success in person re-identification [19,20] and time-series forecasting [21,22], its potential to enhance the robustness and recognition accuracy of person search tasks has yet to be fully explored.

To address this, we propose a novel person search framework, the Dynamic Frequency Selection and Spatial Interaction Fusion Network (PS-DFSI). PS-DFSI strives to fully utilize the complementary advantages of frequency and spatial features to enhance detail perception and improve adaptability to complex scenarios. Its core idea is illustrated in Fig. 1(b) that PS-DFSI introduces frequency feature decoupling and selection effectively to enhance the expressiveness and robustness of features. Specifically, PS-DFSI comprises two key modules: the Dynamic Frequency Selection Module (DFSM) and the Spatial Frequency Interaction Module (SFIM). DFSM first decouples input features into low-frequency and high-frequency components using learnable low-pass and high-pass filters. Then, through a channel attention mechanism, it selects the key frequency components related to the person region and suppresses irrelevant information, avoiding background noise interference in feature extraction. SFIM further combines these frequency features with spatial features, enhancing detail perception through multi-scale receptive fields and achieving efficient fusion of spatial and frequency features via cross-feature attention interaction. PS-DFSI not only overcomes the robustness challenges of spatial domain methods in dealing with complex backgrounds and occlusions but also significantly enhances the overall performance of person search tasks. Extensive experimental results demonstrate that PS-DFSI surpasses existing methods in terms of robustness and accuracy on the CUHK-SYSU and PRW.

Our main contributions are summarized as follows:

- We propose a Dynamic Frequency Selection and Spatial Interaction Fusion Network (PS-DFSI), marking the first attempt to introduce frequency decoupling and selection into person search, significantly enhancing the expressiveness and robustness of person features.
- We design the Dynamic Frequency Selection Module (DFSM) to decouple and select frequency features, effectively emphasizing person features and suppressing background noise.
- We design the Spatial Frequency Interaction Module (SFIM) to effectively fuse frequency and spatial features, enabling PS-DFSI to perceive details.
- Extensive experiments on CUHK-SYSU and PRW show the superior robustness and accuracy of PS-DFSI on person search tasks, verifying significant improvements over state-of-the-art person search baselines.

The paper is structured as follows: Section 2 reviews related work, Section 3 describes the network architecture and principles of the PS-DFSI method, Section 4 presents and analyzes the experimental results, and Section 5 summarizes the paper and discusses future research directions.

2. Related work

2.1. Person search

Person search has gained attention in computer vision due to its real-world applications. Based on training workflows, existing methods are categorized into two-step and one-step approaches.

(1) Two-step methods: These methods [2,15,16,23,24] separate person search into two sub-tasks: person detection and re-identification (ReID), using separate models for training. Zheng et al. [2] explored detector-ReID combinations and proposed Confidence-Weighted Similarity (CWS) to reduce false positives. Lan et al. [23] addressed resolution diversity with a Cross-Level Semantic Alignment (CLSA) network for multi-scale matching. Chen et al. [24] identified and mitigated the optimization conflict between detection and ReID with a dual-stream model. Han et al. [15] introduced ReID loss supervision for more reliable bounding boxes, while Wang et al. [16] improved ReID training consistency by generating query-like bounding boxes.

(2) One-step methods: These methods [3,4,14,17,18] integrate person detection and ReID into a unified end-to-end framework for joint training, yielding fewer parameters and higher efficiency. Xiao et al. [3] first introduced an end-to-end person search framework based on Faster R-CNN [25], improving efficiency by sharing lower-level features with the ReID network and proposing the Online Instance Matching (OIM) loss for representation learning. This sparked interest in end-to-end approaches, which became mainstream [26,27]. Dong et al. [26] introduced a bidirectional interaction model to reduce unnecessary contextual interference. Chen et al. [27] proposed norm-aware embedding (NAE), which decouples detection and ReID by decomposing representations into norm and angle. Additionally, some studies further optimize person search by incorporating query images. Munjal et al. [28] proposed a query-guided mechanism to refine search areas, while Li and Miao [4] introduced SeqNet, using two sequential Faster R-CNNs for detection and ReID. Jaffe et al. [14] developed SeqNeXt, filtering irrelevant images before detection to reduce the search space. Han et al. [29] adopted a queue-style memory buffer for recent sample training, and Kim et al. [18] guided attention modules to highlight identity-invariant regions across different poses using prototypes. Jiang et al. [30] achieved scene adaptation by maintaining highly similar feature representations of the same person across different scenes. Yang et al. [31] proposed an efficient Tri-Hybrid model that improves similarity matching by integrating multi-level information into the pedestrian detection stage. Tian et al. [32] introduced a hybrid pre-training framework that leverages sub-task data for full-task person search, significantly enhancing generalization performance.

In addition, recent studies have begun exploring weakly-supervised and domain-adaptive approaches for person search [33,34], further advancing the field under more challenging real-world scenarios.

In addition to Faster R-CNN-based models, Yan et al. [35] developed AlignPS, an anchor-free person search model based on FCOS [36], which learns feature relationships without predefined anchor boxes. As the first anchor-free approach, it addresses misalignment at scale, region, and task levels. Given the success of Transformers [37,38] in computer vision, researchers have integrated them into person search [39–41]. Cao et al. [39] introduced PSTR, a DETR-based [38] model with a detection encoder–decoder and a discriminative ReID decoder. Yu et al. [40] developed COAT, a Cascade R-CNN-based [42] model that applies multi-scale convolution transformers across stages to capture occlusion features from coarse to fine. Fiaz et al. [41] proposed SAT, a scale-augmented transformer for handling scale variations and occlusions.

Mainstream person search methods primarily rely on spatial domain modeling, enhancing performance through local context augmentation (e.g., AlignPS) or global feature extraction (e.g., COAT, PSTR). However, they often introduce irrelevant information in complex backgrounds, reducing feature discriminability. In contrast, we first incorporate frequency modeling into person search, selecting key frequency features to emphasize person-related regions and suppress background interference. The global properties of frequency enhance robustness against occlusion, pose, and scale variations. By integrating frequency and spatial domains, we improve feature representation, enhancing the model's robustness, generalization, and efficiency in complex scenarios.

2.2. Frequency domain information

Frequency information is pivotal fo digital image processing. Some studies leverage frequency information to enhance visual tasks [19–22], while others use frequency-domain techniques for network acceleration [43,44]. For instance, Zhang et al. [19] proposed PHA, which enhances high-frequency feature representation using the Discrete Haar Wavelet Transform. Li et al. [20] introduced an adaptive high-frequency transformer to strengthen high-frequency learning. Lao et al. [45] proposed FSRU, utilizing Fourier transform for discriminative frequency spectrum features in multi-modal rumor detection. Yi et al. [21] developed FreTS, which employs MLP to learn frequency components for efficient time series prediction. Yang et al. [22] proposed SFFNet, decomposing features into low-frequency and high-frequency components using Haar wavelet transform for remote sensing segmentation. Oyallon et al. [43] proposed a wavelet scattering model for efficient image recognition, while Rao et al. [44] proposed GFNet, establishing long-term dependencies with logarithmic complexity from a frequency perspective.

Although frequency domain methods show potential in visual tasks, most studies overlook the distinction between high-frequency (textures and details) and low-frequency (smooth regions) features. To address this, we introduce frequency domain decoupling and dynamic selection to person search, dynamically selecting key frequency information and fusing it with spatial features to enhance robustness and feature expressiveness in complex scenarios.

2.3. Feature fusion and attention mechanisms

Feature fusion is essential for improving model performance in visual tasks [46,47], especially in complex scenarios where interaction between different feature domains is critical. Recent studies have explored combining spatial and frequency domain features to enhance robustness, such as applying Fourier transform for global frequency features and fusing them with local spatial features to improve detection accuracy. However, most approaches rely on simple concatenation and lack deep interaction modeling, particularly in person search tasks, where such studies are limited. Multi-scale receptive field methods (e.g., FPN [48], ASPP [49], HRNet [50]) improve detail capturing but are restricted to spatial domain processing, underutilizing global frequency information. This limitation hampers performance in handling occlusion, complex backgrounds, and scale variations. Combining global frequency characteristics with multi-scale detail enhancement can significantly improve robustness and adaptability. Attention mechanisms, such as ECA-Net [51] and Transformer, excel in feature selection and capturing cross-domain interactions. While some studies integrate attention mechanisms into feature fusion, most are confined to single-domain optimization, lacking efficient interaction modeling between frequency and spatial domains.

In summary, despite advancements in feature fusion and attention mechanisms, achieving efficient integration of frequency and spatial domains through dynamic interaction remains a challenge in person search. These findings motivate the design of our proposed DFSM and SFIM.

3. Methodology

This section outlines the problem definition and the overall structure of the proposed PS-DFSI model. We then detail its key components, including the Dynamic Frequency Selection Module (DFSM) with its learnable low-pass filter and the Spatial Frequency Interaction Module (SFIM). Finally, we explain the model's loss functions.

3.1. Problem formulation

Given a query person q in a query image Q and a gallery set $I = \{I_1, I_2, \dots, I_N\}$, person search aims to detect pedestrian bounding boxes B within I and identify the best match for q in B. During training, a model $F(\cdot)$ learns discriminative representations $f_i = F(x_i)$ from a labeled dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, ensuring robustness to occlusions and pose variations. In testing, the trained model extracts f_q from Q. For each gallery image I_i, a detector produces candidate boxes $B_i = \{B_{i1}, B_{i2}, \dots, B_{iM}\}$, and the network extracts features f_{ij} for each B_{ij}. The query feature f_q is then compared with f_{ij}, and the box with the highest similarity score is chosen as the final match.

Traditional methods focus on extracting spatial features to obtain boundary features f_{ij} but struggle with complex backgrounds, leading to unstable representations and performance variations. To address this, we propose PS-DFSI, which enhances feature robustness by using DFSM to extract essential frequency features and suppress noise, while SFIM integrates spatial-frequency domain features F_{spa} and F_{fre}, producing robust fusion features F_{out}. These improvements enable PS-DFSI to achieve superior person search performance.

3.2. Overall architecture

Our person search framework builds on SeqNet's [4] end-to-end architecture but adopts a "ReID-first" strategy, integrating novel methods for robust feature representation in complex scenarios. We eliminate detection-related parts from the ReID network and redesign the re-identification head with a Dynamic Frequency Selection Module (DFSM) and a Spatial-Frequency Interaction Module (SFIM), as shown in Fig. 2.

PS-DFSI comprises four components: a backbone network, a detection network, a DFSM, and a SFIM. The first four convolutional blocks of Swin-S [52] serve as the backbone, extracting scene feature maps from complete images. These maps are fed into the detection network, where the Region Proposal Network (RPN) generates region proposals, and the box predictor classifies and regresses them to produce bounding boxes. Pedestrian feature maps are then extracted via RoI-Align [53] and subsequently processed by DFSM, which encodes them into 512-dimensional embeddings, emphasizing key frequency domain information. To align feature dimensionality with the semantic branch and to enable effective attention-based interaction within the SFIM, a

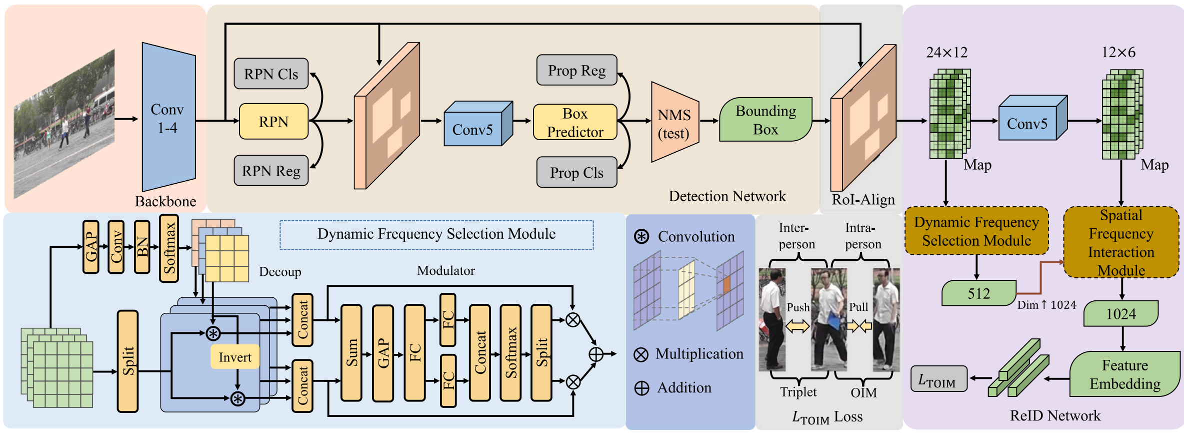

Fig. 2. An outline of the proposed PS-DFSI framework: The input image is first processed by the Backbone network to extract initial features. These features are then passed through the Region Proposal Network (RPN) to generate bounding box predictions, which are subsequently refined using Non-Maximum Suppression (NMS) during the test phase for detection. In the ReID phase, we introduce two novel modules: the Dynamic Frequency Selection Module (DFSM) and the Spatial Frequency Interaction Module (SFIM). The DFSM decouples and selects frequency features, enhancing robustness by suppressing irrelevant background noise. The SFIM fuses spatial and frequency features to improve detail perception and feature representation for ReID. The resulting 1024-dimensional feature embedding is supervised using the L_{TOIM} loss, which effectively balances intra-person similarity and inter-person discrimination.

1×1 convolution is applied to project the DFSM output from 512 to 1024 dimensions. These projected frequency-aware features are then fused with high-level semantic features, such as the 12×6 feature map produced by Conv5, the fifth convolutional block of Swin-S.

During training, we introduce the L_{TOIM} loss to supervise feature learning. This loss integrates the OIM and triplet losses. The triplet loss enhances inter-person discrimination, and the OIM loss forces intra-person similarity.

3.3. Dynamic frequency selection module

To select effective frequency domains for feature enhancement, DFSM involves two main components: Frequency Decoupling Operation and Frequency Selection Modulator. The decoupling operation dynamically decomposes features into different frequency components using learnable filters, while the modulator employs channel attention to emphasize components related to pedestrian regions. DFSM further split features along the channel dimension, using various sizes of filters to capture diverse frequency features, as shown in the lower left of Fig. 2.

3.3.1. Frequency decoupling operation

To dynamically decompose the feature map, we employ theoretically proven learnable low-pass filters (see Section 3.4) and corresponding high-pass filters to generate low-frequency and high-frequency maps. The learned filters are distributed across groups, ensuring a balance between complexity and feature diversity. Given a feature map $X \in \mathbb{R}^{H\times W\times C}$, we apply the filters to generate low-pass components for each group using the following formula:

$$F^{Low} = \text{Softmax}(\text{BN}(W(\text{GAP}(X)))) \tag{1}$$

where $F^{Low} \in \mathbb{R}^{g\times k\times k}$, with $k \times k$ as the size of low-pass filter kernels and g as the number of groups. BN, W, and GAP denote the batch normalization, the linear transformation, and the global average pooling, respectively. The Softmax function is applied to each group. The group operation constrains the kernel size within each group, reducing computational complexity.

High-pass filters are constructed by subtracting low-pass filters from a unit kernel, where the central value is set to 1 and others to 0. Given a group feature $X_i \in \mathbb{R}^{H\times W\times C_i}$ with $C_i = C/g$, low-pass filters F^{Low} and high-pass filters F^{High} are used to extract corresponding frequency components. The computation is as follows:

$$X^{low}_{i,h,w,c} = \sum_{p,q} F^{Low}_{i,p,q} X_{i,h+p,w+q,c} \tag{2}$$

$$X^{high}_{i,h,w,c} = \sum_{p,q} F^{High}_{i,p,q} X_{i,h+p,w+q,c} \tag{3}$$

where $X^{low}_{i,h,w,c}$ and $X^{high}_{i,h,w,c}$ denote the extracted low-frequency and high-frequency features, respectively. Here, c is the channel index, and h, w are spatial coordinates. $p, q \in \{-1, 0, 1\}$ define the local receptive field of the filter in the spatial domain. $F^{Low}_{i,p,q}$ and $F^{High}_{i,p,q}$ represent the weights of the low-pass and high-pass filters, respectively.

3.3.2. Frequency selection modulator

After frequency decomposition, a frequency-selective modulator dynamically adjusts feature weights across frequency components, emphasizing the most relevant features for target feature enhancement. Formally, given two frequency features X^{low} and X^{high}, the fused features are computed as follows:

$$\mathcal{Z} = W_{fc}\left(\text{GAP}\left(X^{low} + X^{high}\right)\right) \tag{4}$$

where W_{fc} defines the parameters of the fully connected layer. To compute the attention weights for high-frequency and low-frequency components, two 1×1 convolution layers process $\mathcal{Z}$. The results are then combined and normalized via the Softmax function, ensuring the sum of the weights to 1.

$$[W^{low}, W^{high}]_c = \frac{e^{[W_{low}(\mathcal{Z}), W_{high}(\mathcal{Z})]c}}{\sum_{j=1}^{2C} e^{[W_{low}(\mathcal{Z}), W_{high}(\mathcal{Z})]_j}} \tag{5}$$

where W^{low} and W^{high} are the channel attention weights for the two frequencies, and $W_{low}(\mathcal{Z})$ and $W_{high}(\mathcal{Z})$ are the weights of two fully connected layers. $[\cdot,\cdot]$ denotes concatenation. The total number of high-frequency and low-frequency channels is $2C$, with c as the current channel index. The Softmax computes weights for each channel c separately, iterating over all channels indexed by j.

The normalized weights are then split into high-frequency and low-frequency parts and applied to X^{high} and X^{low} for channel-wise weighting. Lastly, a 1×1 convolution layer further refines the fused features. The formula is:

$$X^{out} = W_{out}\left(X^{high} \odot W^{high} + X^{low} \odot W^{low}\right) \tag{6}$$

The above describes a single-branch dynamic frequency selection module. We extend it to a multi-branch structure with varying filter sizes, represented as:

$$\hat{X} = [\mathcal{S}_1(\mathcal{D}_1(X_1)), \ldots, \mathcal{S}_m(\mathcal{D}_m(X_m))] \tag{7}$$

where $\hat{X}$ denotes the feature set after multi-branch fusion, $\mathcal{D}$ and $\mathcal{S}$ represent the frequency domain decoupler and modulator, respectively, and X_m denotes the feature set after the mth branch partitioning.

3.4. Proof the low-pass filter

In DFSM, the filter is applied to a $k\times k$ region of the feature map via convolution, operating in the spatial dimension with a sliding window and a specific stride. Next, we prove its low-pass characteristics.

Theorem 1. *Given $W, D \in \mathbb{R}^{n\times n}$, and $W_i = softmax(D_i)$, $i \in \{0, 1, \dots, n-1\}$, then W is a low-pass filter, satisfying for any $m \in \mathbb{R}^n$, as $t \to \infty$:*

$$\lim_{t\to\infty} \frac{\|\mathcal{H}[W^t m]\|_2}{\|W^t m\|_2} = 0. \tag{8}$$

where $\mathcal{H}[\cdot]$ represents the high-frequency component extraction operation. In our case, $n = k^2$. Here, t denotes the number of times the filter W is applied recursively.

Proof. Through the Softmax operation, W is a non-negative matrix, and each row sums to 1, i.e. $We = e, \quad e = [1, 1, \dots, 1]^T \in \mathbb{R}^n$. Thus, the main eigenvalue of W is $\lambda_1 = 1$, corresponding to the eigenvector e. According to the **Perron–Frobenius** theorem [54], the spectral radius of W is 1, and the absolute values of the other eigenvalues satisfy $|\lambda_i| < 1, i \neq 1$. To compute the higher powers of W, we perform the Jordan canonical decomposition of W.

$$\begin{aligned} W &= PJP^{-1} \\ &= \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix} \begin{bmatrix} \lambda_1 & J_2(\lambda_2) & \dots & J_s(\lambda_s) \end{bmatrix} \begin{bmatrix} b_1^T \\ b_2^T \\ \vdots \\ b_n^T \end{bmatrix} \end{aligned} \tag{9}$$

where P is a non-singular matrix (a linearly independent combination of eigenvectors), and $J_i(\lambda_i)$ is the Jordan block. For W, after t iterations, we have: $W^t = PJ^tP^{-1}, \quad J^t = \mathrm{diag}(\lambda_1^t, J_2(\lambda_2)^t, \dots, J_s(\lambda_s)^t)$. Thus, for any vector $m \in \mathbb{R}^n$: $W^t m = \lambda_1^t a_1 + \sum_{i=2}^{s} J_i(\lambda_i)^t b_i, \quad a_1, b_i \in \mathbb{R}^n$. Where $\lambda_1 a_1$ represents the principal component, and $J_i(\lambda_i)^t$ represents the high-frequency components.

Since $|\lambda_i| < 1$ for all $i \neq 1$, the elements of the Jordan block $J_i(\lambda_i)^t$ decay exponentially:

$$\lim_{t\to\infty} J_i(\lambda_i)^t = 0, \quad i \neq 1. \tag{10}$$

Thus, the high-frequency component $\mathcal{H}[W^t m]$ is primarily contributed by $J_i(\lambda_i)^t$, and we have: $\|\mathcal{H}[W^t m]\|_2 \to 0$ as $t \to \infty$. At the same time, due to the eigenvalue $\lambda_1 = 1$, the principal component $\lambda_1^t a_1$ remains unchanged. Therefore, the overall expression for $W^t m$ is: $\|W^t m\|_2 \sim \|\lambda_1^t a_1\|_2 = \|a_1\|$. Finally, the ratio of the norms is:

$$\lim_{t\to\infty} \frac{\|\mathcal{H}[W^t m]\|_2}{\|W^t m\|_2} = \frac{0}{\|a_1\|_2} = 0. \tag{11}$$

Proved. □

Based on the above derivation, the matrix W, as a low-pass filter, can effectively preserve the low-frequency components, while the high-frequency components decay to zero over time, thus proving the result.

3.5. Spatial frequency interaction module

Although DFSM enhances frequency domain features, it struggles with capturing local details in complex scenarios like occlusion, multi-scale variations, or background interference. To address this, we introduce the SFIM, which integrates spatial and frequency features for improved local representation. Specifically, SFIM fuses frequency-modulated features from DFSM with high-level semantic features (e.g., Conv5), enhancing local detail perception through multi-scale receptive fields and efficient feature fusion via cross-feature attention. This integration improves the network's ability to capture both global semantics and local details, significantly enhancing its representational power. The SFIM structure, shown in Fig. 3, includes multi-scale mapping and cross-feature attention.

3.5.1. Multi-scale mapping operation

We apply dimensionality reduction to project the frequency and spatial features into a unified dimension. Local details are then extracted using horizontal and vertical convolutions of varying sizes. The results are concatenated and processed with 1×1 convolutions to generate a unified feature array, Q, K, and V. For the input feature $x \in \mathbb{R}^{C\times H\times W}$, multi-scale mapping operations are defined as:

$$\begin{aligned} Q, K, V &= f_{mm}(x) \\ &= Conv_{1\times 1}\Big(OC_3(LN(x)) + OC_5(LN(x)) + OC_7(LN(x))\Big) \end{aligned} \tag{12}$$

where LN denotes layer normalization, OC_k represents the kernel size of $k\times 1$ and $1\times k$ convolutions, and f_{mm} indicates the convolution operation. Q, K, and $V \in \mathbb{R}^{C\times H\times W}$ are the feature matrices obtained after multi-scale mapping.

For the SFIM, the input frequency-optimized feature $F_{fre} \in \mathbb{R}^{C\times H\times W}$ and spatial feature $F_{spa} \in \mathbb{R}^{C\times H\times W}$ are processed using multi-scale mapping as follows:

$$\begin{cases} Q_1, K_1, V_1 = f_{mm}(F_{fre}) \\ Q_2, K_2, V_2 = f_{mm}(F_{spa}) \end{cases} \tag{13}$$

3.5.2. Cross-feature attention interaction

Building on multi-scale mapping, cross-feature attention interaction (CAI) strengthens the features by performing interaction calculations across input branches, enabling effective fusion.

Specifically, given the Query and Key features Q_1, K_1 from multi-scale mapping, and Q_2, K_2 from other branches, the attention interaction process is as follows:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{Q\cdot K^T}{\sqrt{d_k}}\right)\cdot V \tag{14}$$

where d_k is the dimension of the Key feature and it serves to scale the attention weights and ensure numerical stability. The calculation process of CAI is as follows:

$$\begin{cases} F_{Fre} = Conv_{1\times 1}(\mathrm{Attention}(Q_2, K_1, V_1)) \\ F_{Spa} = Conv_{1\times 1}(\mathrm{Attention}(Q_1, K_2, V_2)) \end{cases} \tag{15}$$

The entire SFIM process can be defined as:

$$F_{out} = f_{sfim}(F_{fre}, F_{spa}) = \mathrm{Concat}(F_{Fre}, F_{Spa}) \tag{16}$$

where $F_{Fre}, F_{Spa} \in \mathbb{R}^{C/2\times H\times W}$ represent the frequency and spatial domain features, respectively, as the two outputs of CAI. f_{sfim} denotes the SFIM application, and $F_{out} \in \mathbb{R}^{C\times H\times W}$ denotes the output of the SFIM module.

3.6. Loss function

OIM [3] loss is a widely used method in person search for re-identification. It stores features of labeled identities in a lookup table (LUT), $V \in \mathbb{R}^{D\times L} = v_1, \dots, v_L$, while features of unmarked identities are stored in a circular queue, $U \in \mathbb{R}^{D\times Q} = u_1, u_2, \dots, u_L$. Here, D is the feature dimension, and L is the number of identities. For an identity labeled as I, the feature representation is $x \in \mathbb{R}^D$, and the OIM calculation probability is as follows:

$$p_i = \frac{\exp(v_i^T x/\tau)}{\sum_{j=1}^{L}\exp(v_j^T x/\tau) + \sum_{k=1}^{Q}\exp(u_k^T x/\tau)} \tag{17}$$

where $\tau = 1/30$ is a hyperparameter for controlling the smoothness of the probability distribution. OIM aims to minimize the negative log-likelihood of the target:

$$L_{\mathrm{OIM}} = -E_x(\log p_t), \quad t = 1, 2, \dots, L \tag{18}$$

Although OIM yields satisfactory results, its softmax function emphasizes inter-person similarity but is less sensitive to differences between pedestrians, limiting discriminative power. To address this, we

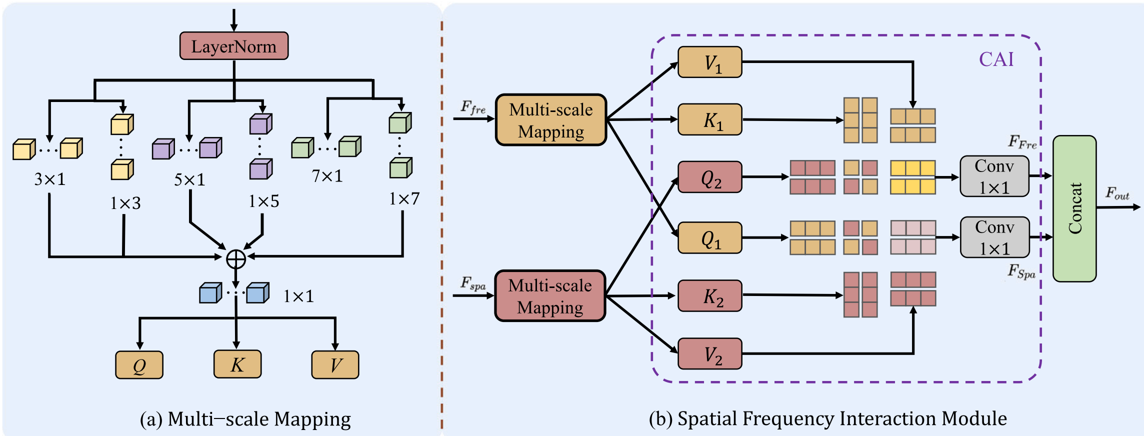

Fig. 3. Illustrations of the Multi-scale Mapping and Spatial Frequency Interaction Module. (a) Multi-scale Mapping: This module performs feature extraction at multiple scales using convolutions of varying kernel sizes to generate the query (Q), key (K), and value (V) feature matrices. LayerNorm standardizes the input feature maps before convolution. (b) Spatial Frequency Interaction Module (SFIM): This module integrates frequency-modulated features and spatial features through multi-scale mapping. It computes separate query, key, and value features (Q_1, K_1, V_1 from the frequency domain and Q_2, K_2, V_2 from the spatial domain) and performs cross-feature attention interaction. The results from both domains are combined through 1×1 convolutions and concatenated to produce the output feature F_{out}. SFIM effectively fuses spatial and frequency information, improving the model's capacity to capture fine-grained details.

introduce a triplet loss with a hyperparameter Φ to control the margin between cosine similarities of different and same pedestrians.

In each min-batch, $B \in \mathbb{R}^{D\times B}$, given a feature representation $x \in \mathbb{R}^D$ for a labeled identity I, we sample a batch from the combined set of B and LUT, $B \cup V$. The batch includes P feature vectors with the same label I, denoted as $X_1 \in \mathbb{R}^{D\times P}$, and N feature vectors with different labels, denoted as $X_0 \in \mathbb{R}^{D\times N}$. The triplet loss is expressed as:

$$L_\Phi = \max\{\Phi - [\min(X_1^\top x) - \max(X_0^\top x)], 0\} \tag{19}$$

where $X_1^\top x$ and $X_0^\top x$ represent the cosine similarities between the feature representation and the corresponding sets. Finally, the L_{TOIM} loss introduces a weighting parameter λ to combine the two terms.

$$L_{\text{TOIM}} = L_{\text{OIM}} + \lambda L_\Phi \tag{20}$$

4. Experiments

In this section, we evaluate our method on two widely used person search datasets. We first describe the datasets, evaluation metrics, and implementation details, followed by comparisons with state-of-the-art models. Ablation studies are performed to analyze the contribution of each module. Finally, additional experiments and qualitative analysis are provided to further validate the effectiveness of our approach.

4.1. Datasets and settings

4.1.1. CUHK-SYSU

CUHK-SYSU [3] is a large-scale person search dataset with 18,184 scene images from street surveillance and movie screenshots. It contains 96,143 annotated bounding boxes and 8432 pedestrian IDs, split into two non-overlapping subsets. The training set includes 11,206 images, 55,272 bounding boxes, and 5532 pedestrians, while the test set comprises 6978 images, 40,871 bounding boxes, and 2900 pedestrian IDs. During testing, galleries range from 50 to 4000 images to assess model scalability. Unless specified, results are reported with a gallery size of 100, otherwise.

4.1.2. PRW

PRW [2] was collected on a university campus using six static cameras, featuring diverse viewpoints and notable scale variations. It contains 11,816 frames and 34,304 annotated bounding boxes for 932 pedestrian identities. The training set includes 5704 images with 18,048 bounding boxes and 482 identities, while the test set consists of 6112 images with 2057 query pedestrians across 450 identities. Unlike CUHK-SYSU, PRW uses the entire gallery set as the search space during testing. Following [18], we use 20% of the training set as a validation set for hyperparameter tuning.

4.1.3. Evaluation metrics

Following [3,4], we use standard person search metrics. A predicted box is a match if its IoU with the ground truth exceeds 0.5. For pedestrian detection, we report recall and Average Precision (AP). For person ReID, we evaluate mean Average Precision (mAP) and top-1 accuracy.

4.1.4. Implementation details

All experiments are performed on a single NVIDIA Tesla V100 GPU using PyTorch. SeqNet [4] serves as the baseline, with detection components removed in the ReID network. ResNet50 [55], ConvNeXt-B [56], and Swin-S [52] pretrained on ImageNet are used as backbones. We set the batch size to 5 for CUHK-SYSU and 8 for PRW, with automatic mixed precision (AMP) enabled for efficiency. The model is trained for 20 epochs using the Adam optimizer, with an initial learning rate of 0.0001, increased in the first epoch and reduced by a factor of 10 at epochs 8 and 14. In DFSM, filter sizes are 3×3 and 5×5, with $m = 2$ and $g = 8$. In L_{TOIM}, $\Phi = 0.2$, and λ is 0.6 for CUHK-SYSU and 0.8 for PRW. During testing, Non-Maximum Suppression (NMS) with a 0.5 threshold removes redundant bounding boxes. Detailed hyperparameter analysis is in Section 4.5.

4.2. Comparison with state-of-the-art methods

In this section, we compare PS-DFSI with seven two-step methods and 19 one-step methods. To assess the scalability of different backbones, we display results using the latest convolutional network ConvNext-B [56] with SeqNeXt [14] and the Transformer-based PVT [63] with PSTR [39]. For baseline comparisons, we use the official implementations where available.

4.2.1. Comparison on CUHK-SYSU dataset

In Table 1, we report the mAP and top-1 metrics on the CUHK-SYSU dataset with a gallery size of 100. Among ResNet50-based models, PS-DFSI achieves the second-highest mAP (95.5%) and the fourth-highest top-1 (95.9%), comparable to SEAS [30], and surpasses the best two-step model, TCTS [16], which separates detection and ReID into two

Table 1
A comparison of mAP and top-1 accuracy with state-of-the-art methods on the CUHK-SYSU and PRW datasets, with the optimal and suboptimal results in each group marked in **bold** and underline, respectively.

Methods	Ref	Backbone	CUHK-SYSU		PRW	
			mAP	top-1	mAP	top-1
Two-step methods						
IDE [2]	CVPR17	ResNet50	–	–	20.5	48.3
MGTS [24]	ECCV18	VGG16	83.0	83.7	32.6	72.1
CLSA [23]	ECCV18	ResNet50	87.2	88.5	38.7	65.0
RDLR [15]	ICCV19	ResNet50	93.0	94.2	42.9	70.2
IGPN [57]	CVPR20	ResNet50	90.3	91.4	47.2	87.0
TCTS [16]	CVPR20	ResNet50	93.9	95.1	46.8	87.5
OR [58]	TIP21	ResNet50	92.3	93.8	52.3	71.5
One-step with CNNs						
OIM [3]	CVPR17	ResNet50	75.5	78.7	21.3	49.4
RCAA [59]	ECCV18	ResNet50	79.3	81.3	–	–
HOIM [17]	AAAI20	ResNet50	89.7	90.8	39.8	80.4
APNet [60]	CVPR20	ResNet50	88.9	89.3	41.9	81.4
NAE+ [27]	CVPR20	ResNet50	92.1	92.9	44.0	81.1
AlignPS+ [35]	CVPR21	ResNet50	94.0	94.5	46.1	82.1
SeqNet [4]	AAAI21	ResNet50	94.8	95.7	47.6	87.6
OIMNet++ [61]	ECCV22	ResNet50	93.1	93.9	46.8	83.9
DMRNet++ [29]	TPAMI23	ResNet50	94.5	95.7	52.1	87.0
SeqNeXt [14]	WACV23	ConvNeXt-B	96.1	<u>96.5</u>	57.6	89.5
PAD [18]	TPAMI24	ConvNeXt-B	95.9	96.4	58.6	**89.9**
SEAS [30]	IJCAI24	ResNet50	<u>96.2</u>	96.1	52.0	85.7
Tian *et al.* [32]	AAAI24	ResNet50	95.4	96.0	54.5	87.6
One-step with Transformers						
PSTR [39]	CVPR22	ResNet50	93.5	95.0	49.5	87.8
PSTR [39]	CVPR22	PVTv2-B2	95.2	96.2	56.5	89.7
COAT [40]	CVPR22	ResNet50	94.2	94.7	53.3	87.4
SAT [41]	WACV23	ResNet50	95.3	96.0	55.0	89.2
SOLIDER [62]	CVPR23	Swin-S	95.5	95.8	**59.8**	86.7
Yang *et al.* [31]	TCSVT24	ResNet50	94.9	95.2	58.3	<u>89.7</u>
PS-DFSI(Ours)	–	ResNet50	95.5	95.9	55.2	88.6
PS-DFSI(Ours)	–	ConvNeXt-B	96.0	<u>96.5</u>	58.4	89.2
PS-DFSI(Ours)	–	Swin-S	**96.4**	**96.8**	<u>59.5</u>	**89.9**

stages, using a detector to generate bounding boxes and a ReID model to refine them.

Comparisons with Transformer-based methods (e.g., PSTR, COAT, and SOLIDER [62]) further demonstrate the effectiveness of frequency-domain learning. Using Swin-S [52] as the backbone, PS-DFSI achieves the best mAP (96.4%) and top-1 (96.8%), significantly outperforming COAT and PSTR, highlighting its scalability with Transformer backbones.

We further evaluate the models across different gallery sizes (ranging from 50 to 4000) to assess their robustness in larger-scale scenarios, as shown in Fig. 4. As the gallery size increases, mAP scores of all methods decrease, highlighting the challenges of instance recognition in broader search spaces. PS-DFSI consistently outperforms other methods across all gallery sizes, surpassing SAT [41] and COAT [40] with ResNet50 and PSTR [39] with PVT. This superior outcome stems from the efficient fusion of frequency and spatial domain features, enhancing adaptability to complex backgrounds and occlusions. Even with a larger search space, PS-DFSI maintains robust performance.

4.2.2. Comparison on PRW dataset

Compared to the CUHK-SYSU dataset, the PRW dataset features a larger gallery size, with many identities having similar appearances, as well as more variations in scale, complex environments, and occlusions. These factors make distinguishing pedestrians significantly more challenging. Notably, on the ResNet50 backbone, our PS-DFSI achieves competitive performance on PRW, with a second-highest mAP of 55.2% and a third-highest top-1 accuracy of 88.6%.

Among the one-step state-of-the-art methods, PS-DFSI outperforms AlignPS, which considers feature misalignment, and SeqNet, which uses a two-stage refinement approach. Compared to PSTR, which recently introduced the more powerful DETR detector, our method achieves a 5.7% improvement in mAP. For the top-1 metric, our approach performs comparably to SeqNeXt [14] and COAT. Similar to the CUHK-SYSU dataset, when Swin-S is used as the backbone network, PS-DFSI achieves a mAP of 59.5% and a top-1 score of 89.9%, significantly outperforming COAT and SeqNeXt. However, compared to SOLIDER [62], which also uses the Swin-S backbone, PS-DFSI's mAP is slightly lower. This can be attributed to SOLIDER's use of a semantic controller that dynamically adjusts the balance between semantic and appearance information, allowing it to better adapt to tasks that require more semantic details. In contrast, PS-DFSI focuses on robust feature learning for the ReID task, which may limit its flexibility in tasks requiring more semantic adaptation. Nevertheless, PS-DFSI still delivers strong performance in person search.

Table 2 presents the benchmark results under the multi-view gallery setting in the PRW dataset. In this setting, PS-DFSI achieves the best performance, with a mAP of 52.2% and a top-1 score of 77.4%, surpassing our baseline SeqNet and achieving competitive performance with recent strong methods such as SAT and PAD. This setting simulates a real-world scenario where multiple images of the same person are captured from different camera viewpoints. The experiment aims to evaluate the robustness and adaptability of the PS-DFSI method in handling viewpoint changes. By comparing with other methods in this multi-view scenario, the experimental results demonstrate that PS-DFSI is more effective at capturing invariant features across viewpoints, leading to a significant improvement in performance.

4.3. Qualitative performance

4.3.1. CUHK-SYSU

Fig. 5 shows the top-1 search results on the CUHK-SYSU dataset, comparing PS-DFSI with state-of-the-art models like SeqNet [4], PAD [18], and SAT [41]. The results highlight the impact of occlusion (first row) and cluttered backgrounds (second and third rows) on

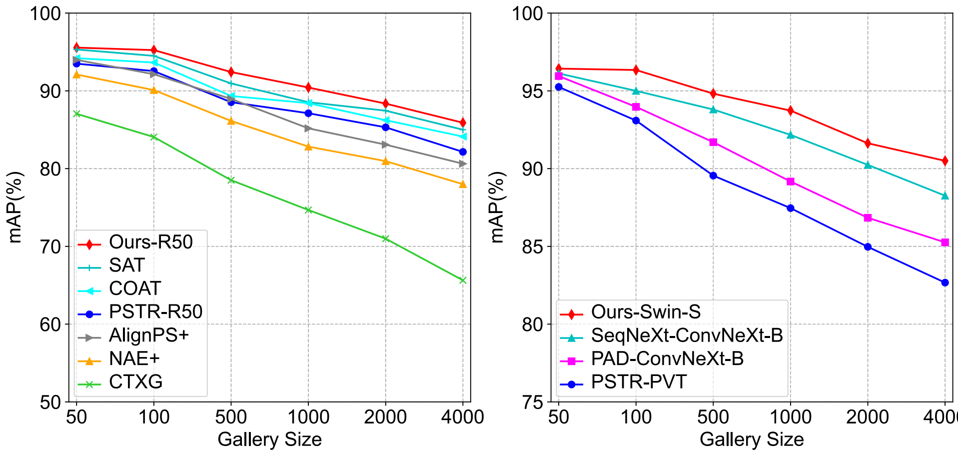

Fig. 4. Impact of gallery sizes on mAP across different configurations in the CUHK-SYSU dataset. We evaluate state-of-the-art methods using ResNet50 as the backbone (left) and the latest ConvNeXt and PVT backbones (right).

Table 2
Quantitative results evaluated on the PRW dataset with a multi-view gallery. All experiments are conducted using the ResNet50 backbone.

Methods	Backbone	PRW	
		mAP	top-1
HOIM [17]	ResNet50	36.5	65.0
APNet [60]	ResNet50	38.7	66.7
NAE+ [27]	ResNet50	40.0	67.5
SeqNet [4]	ResNet50	43.6	68.5
COAT [40]	ResNet50	50.9	75.1
SAT [41]	ResNet50	52.1	75.4
PAD [18]	ResNet50	52.1	77.3
PS-DFSI(Ours)	ResNet50	52.2	77.4

search performance. PS-DFSI consistently outperforms other methods in detecting and identifying the query person. This demonstrates that PS-DFSI effectively handles occlusions and background interference through the fusion of frequency-domain and spatial-domain features, producing more discriminative embeddings. Additionally, PS-DFSI excels at fine-grained feature capture, as seen in the fourth row, where it correctly identifies the white stripes on the shirt hem, while other methods misidentify a person in a black shirt as the target.

4.3.2. PRW

Fig. 6 shows the rank-3 search results on the PRW dataset for PAD [18] and PS-DFSI. While PAD performs well in top-1 ranking, it struggles with consistent detection due to appearance similarities (first row) and viewpoint variations (second and third rows). In contrast, PS-DFSI, through efficient fusion of frequency-domain and spatial-domain features, captures finer details from different regions, addressing intra-person variations such as appearance and pose changes. Despite some failures (e.g., fourth row with minimal differences and partial occlusion), PS-DFSI consistently captures key details (e.g., briefcase in the third row), while PAD misidentifies it as an umbrella. Although the retrieved person is not the correct identity, the presence of the briefcase indicates that PS-DFSI focuses more accurately on fine-grained visual cues. This demonstrates the method's strength in capturing discriminative features and maintaining robustness under challenging conditions.

4.4. Ablation study

In this section, we perform ablation experiments to assess the contribution of each component and validate the design choices of our method. To ensure a fair comparison while optimizing computational resources and time, we select ResNet50 as the backbone for the subsequent experiments.

4.4.1. Analysis of different components

We conduct ablation experiments on the PRW dataset to assess the contributions of each proposed component. Table 3 shows the incremental performance improvements achieved by adding components to the baseline model, which is derived from SeqNet by removing detection-related parts, initially achieving 48.6% mAP and 87.6% top-1 accuracy. Specifically, the DFSM, comprising the Frequency Decoupling Operation (FDO) and Frequency Selection Modulator (FSM), shows significant improvements. Rows 2–4 show that introducing FDO improves mAP by 2.0%, demonstrating its effectiveness in decoupling features into high-frequency and low-frequency components. Furthermore, further adding FSM increases mAP to 52.3% and top-1 accuracy to 87.9%, highlighting the significant role of FSM in modulating key frequency components. In addition, the SFIM, consisting of the Multi-scale Mapping Operation (MMO) and Cross-feature Attention Interaction (CAI), further boosts performance. Rows 5–7 show that adding MMO increases mAP to 53.7% (a 1.4% gain) and top-1 accuracy to 88.0%, demonstrating MMO's ability to capture multi-scale local information. Adding CAI boosts mAP to 54.1% and top-1 accuracy to 88.2%, confirming the critical contribution of CAI in cross-feature interaction and fusion of global and local features. Finally, introducing the triplet loss term L_{Φ} in L_{TOIM} improves performance by 0.7% in mAP and 0.3% in top-1 accuracy. This proves that L_{TOIM} enhances the network's discriminative ability by improving intra-person consistency and inter-person separability.

4.4.2. Analysis of different decoupling methods

To validate the effectiveness of the decoupling strategy in DFSM and to further analyze the application and limitations of different methods in person search, we compare the performance of fixed and learnable alternatives. As shown in Table 4, standard convolution (Conv) operates in the spatial domain and struggles with frequency separation, resulting in lower performance. This limitation becomes more evident in scenarios with background clutter and occlusions. Gaussian filters provide basic frequency decomposition but lack adaptive optimization, limiting performance (mAP 54.2%). Wavelet transform improves results via multi-scale information (mAP 54.9%, top-1 88.5%), yet struggles with viewpoint variation. Dynamic convolution offers better flexibility and slightly higher accuracy (mAP 55.4%), but its high parameter count (2.53M) increases computational cost, making it less suitable for resource-limited settings. In contrast, our FDO employs learnable low-pass and high-pass filters for efficient frequency decomposition, achieving the best trade-off in performance with fewer parameters (2.34M).

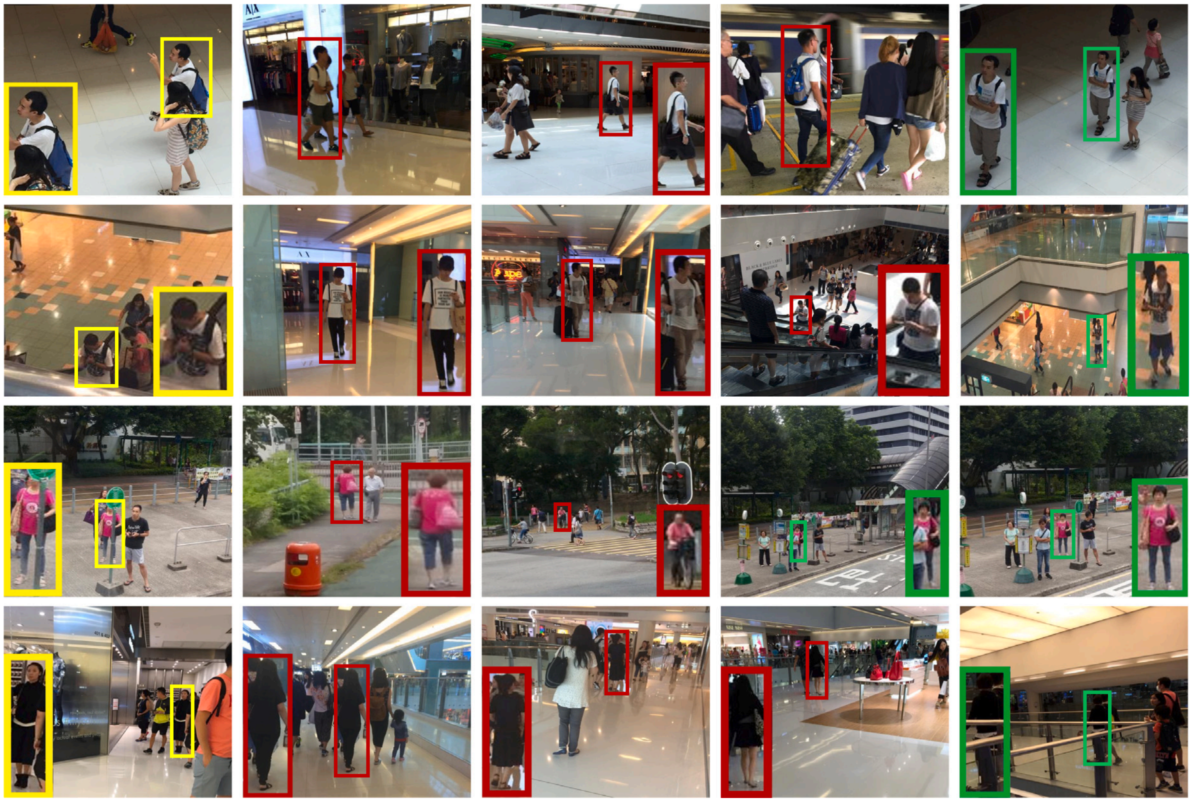

Fig. 5. Qualitative results on the CUHK-SYSU dataset. We visualize the top-1 matching results for the state-of-the-art methods SeqNet, PAD, SAT, and our method for each query image. Red boxes indicate failed matches, while green boxes indicate successful matches. For comparison, we provide zoomed-in views of some detected instances.

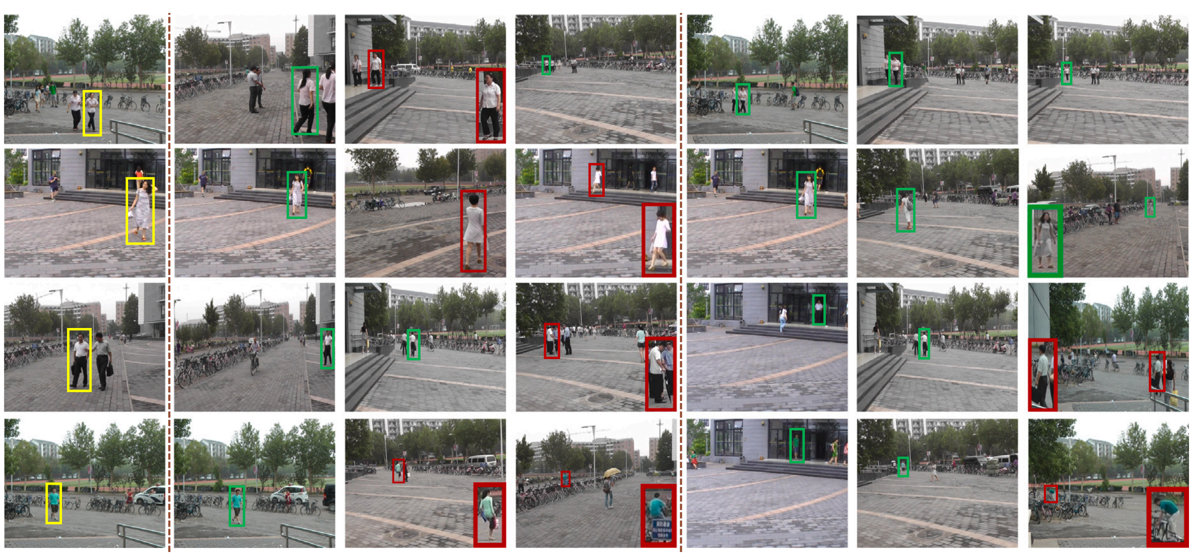

Fig. 6. Qualitative results on the PRW dataset. For each query image, we visualize the rank-3 search results for PAD and our method. Red boxes indicate failed matches, while green boxes indicate correct matches.

4.4.3. Analysis of different scales

To further investigate the impact of multi-scale mapping in SFIM, we examine how different scales affect network performance. As shown in Table 5, the first three rows represent feature extraction at a single scale, while the middle three rows and the last row use two and three scales for feature extraction, respectively. The experimental results indicate that using three scales achieves the best performance on both the CUHK-SYSU and PRW datasets. This suggests that by modeling the human visual system to extract essential information from diverse perspectives, the network can capture rich and comprehensive pedestrian features.

4.4.4. Analysis of different fusion methods

To further investigate the effectiveness of the Cross-feature Attention Interaction (CAI) module in SFIM, we replace CAI with several common feature fusion strategies (such as FPN, ASPP, HRNet, and

Table 3
Ablation analysis demonstrating the impact of each added component on the PRW dataset.

Baseline	DFSM		SFIM		L_{TOIM}		ReID		Detection	
	FDO	FSM	MMO	CAI	w/o L_{Φ}	w/ L_{Φ}	mAP	top-1	Recall	AP
✓							48.6	87.6	95.4	92.9
✓	✓						50.6	87.4	95.7	93.2
✓		✓					51.4	87.7	95.3	93.1
✓	✓	✓					52.3	87.9	95.8	93.6
✓	✓	✓	✓				53.7	88.0	96.2	93.6
✓	✓	✓		✓			53.5	87.8	96.4	93.5
✓	✓	✓	✓	✓			54.1	88.2	96.6	93.7
✓	✓	✓	✓	✓	✓		54.5	88.3	96.8	94.1
✓	✓	✓	✓	✓		✓	**55.2**	**88.6**	**97.1**	**94.6**

Table 4
Analysis of different decoupling methods on the PRW dataset.

Decoupling methods		#Params(M)	PRW	
			mAP	top-1
Fixed	Conv	2.88	53.8	87.6
	Gaussian [64]	**2.34**	54.2	88.2
	Wavelet [19]	2.36	54.9	88.5
learnable	Dynamic Conv [65]	2.53	55.4	88.4
	FDO(Ours)	**2.34**	**55.2**	**88.6**

Table 5
Ablation study on scale combinations for SFIM multiscale mapping on the CUHK-SYSU and PRW datasets. Bold indicates the best performance.

Scale combinations	CUHK-SYSU		PRW	
	mAP	top-1	mAP	top-1
{3×1, 1×3}	94.4	95.2	53.8	88.0
1 {5×1, 1×5}	94.9	95.2	54.3	88.3
{7×1, 1×7}	94.8	95.3	54.2	88.4
{3×1, 1×3},{5×1, 1×5}	95.2	95.5	54.6	88.3
1 {3×1, 1×3},{7×1, 1×7}	95.3	95.7	54.8	**88.7**
{5×1, 1×5},{7×1, 1×7}	95.1	95.7	54.9	88.5
3 {3×1, 1×3},{5×1, 1×5},{7×1, 1×7}	**95.5**	**95.9**	**55.2**	88.6

Table 6
Ablation study on feature fusion methods for CAI in SFIM on CUHK-SYSU and PRW datasets.

Fusion methods	CUHK-SYSU		PRW	
	mAP	top-1	mAP	top-1
FPN [48]	94.4	94.9	54.2	87.4
ASPP [49]	95.2	95.7	54.8	88.2
HRNet [50]	95.4	95.6	54.4	87.8
SE module [66]	94.2	94.4	53.5	86.7
CAI(Ours)	**95.5**	**95.9**	**55.2**	**88.6**

SE) to examine their impact on network performance. As shown in Table 6, although traditional methods improve performance to some extent (for example, HRNet achieves an mAP of 54.4% on the PRW dataset), using CAI leads to a performance improvement to 55.2%. This performance difference further validates the importance of CAI. CAI effectively integrates global information from the frequency domain and local detail features from the spatial domain through a cross-domain attention mechanism, enhancing the model's ability to perceive details.

4.4.5. Analysis of F_{Fre} and F_{Spa} in SFIM

To assess the contributions of F_{Fre} and F_{Spa} in SFIM, we conduct an ablation study by using each feature individually. As shown in Table 7, the performance of using only F_{Fre} (54.6% mAP, 88.3% top-1) slightly exceeds that of using only F_{Spa}, indicating that frequency-domain features contribute more effectively to fine-grained discrimination. Notably, combining both yields the best performance, confirming the complementary strengths of the two branches and validating the effectiveness of our fusion design.

Table 7
Analysis of the contributions of F_{Fre} and F_{Spa} in SFIM on the PRW dataset.

Methods	PRW	
	mAP	top-1
Only F_{Fre}	54.6	88.3
Only F_{Spa}	54.1	87.9
Concat(F_{Fre}, F_{Spa})	55.2	88.6

Table 8
Performance comparison under different RoI-Align resolutions on PRW.

Resolution	DFSM	SFIM	PRW	
			mAP	top-1
14 × 14			50.6	87.2
14 × 14	✓		53.0	87.9
14 × 14		✓	52.8	87.5
14 × 14	✓	✓	54.5	88.4
24 × 12			51.2	87.4
24 × 12	✓	✓	55.2	88.6

4.4.6. Analysis of different RoI-align resolution

We conduct ablation experiments under both the standard 14 × 14 and our adopted 24 × 12 RoI-Align resolutions. As shown in Table 8, even with the 14 × 14 setting, adding DFSM or SFIM individually yields 2.4%/2.2% mAP and 0.7%/0.3% top-1 gains. Using both together leads to 3.9% mAP and 1.2% top-1 improvement. In contrast, increasing resolution alone provides only 0.6% mAP and 0.2% top-1 gain. At 24 × 12, integrating our modules achieves a significantly higher boost of 4.0% mAP and 1.2% top-1, indicating that the improvement mainly comes from our proposed modules, not resolution alone.

4.5. Hyperparameter analysis

4.5.1. Number of groups g

We evaluate the effect of g on frequency-domain decomposition and selection, focusing on its impact on mAP and top-1 accuracy. As g increases, finer grouping enables filters to capture more diverse frequency information, generally enhancing performance. However, an excessively large g may cause over-decomposition, leading to higher computational cost, and degrading performance (As shown in Fig. 7). The optimal $g = 8$ balances granularity and performance, achieving the best feature representation.

4.5.2. Number of branches m

We analyze the impact of the multi-branch parameter m on feature capture at different scales, aiming to balance fine-grained perception and computational complexity. As shown in Fig. 8, when $m = 2$, both mAP and top-1 accuracy stabilize at optimal levels on the CUHK-SYSU and PRW datasets. While increasing m (e.g., $m = 3, 4$) may provide slight performance gains, it also significantly raises computational cost. Thus, $m = 2$ is selected as the optimal configuration, balancing performance and computational efficiency.

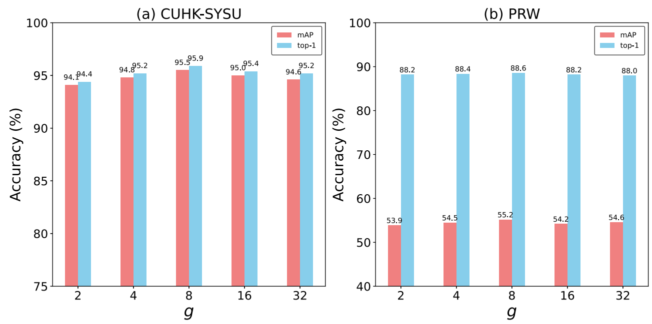

Fig. 7. Impact of g on CUHK-SYSU and PRW datasets.

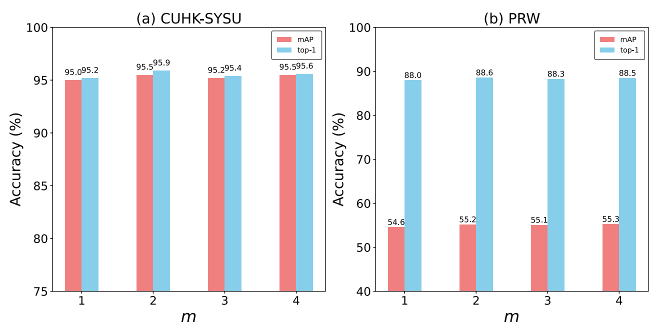

Fig. 8. Impact of m on CUHK-SYSU and PRW datasets.

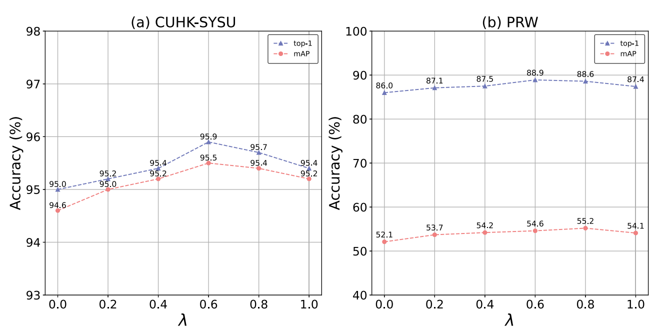

Fig. 9. Impact of λ on CUHK-SYSU and PRW datasets.

4.5.3. Loss weight λ

We evaluate the impact of the weight parameter $\lambda \in [0, 1]$ in L_{TOIM} (Eq. (20)) on model performance. As shown in Fig. 9, the optimal values for λ are 0.6 for CUHK-SYSU and 0.8 for PRW. A higher λ on PRW enhances L_{Φ}'s effect, improving intra-person consistency for cross-camera matching. This demonstrates that L_{TOIM}, with λ, dynamically balances inter-class discriminability and intra-person consistency, boosting feature learning.

4.5.4. Hyperparameter Φ

We evaluate the impact of the hyperparameter Φ on L_{Φ}. As shown in Fig. 10, the optimal value of Φ is 0.2 for CUHK-SYSU and PRW. This indicates that an appropriate setting of Φ can effectively control the minimum cosine similarity margin between pedestrians, thereby enhancing inter-person discriminability while ensuring the compactness of intra-person features. Consequently, this significantly improves the overall performance of the model.

4.6. Qualitative analysis of DFSM

In the DFSM module, we visualize the spectral features of different groups (as shown in Fig. 11). The results show that the DFSM module adaptively selects low-frequency or high-frequency signals, demonstrating its dynamic selection capability and enhancing the diversity of frequency-domain features. The filters in different groups exhibit distinct and targeted characteristics, focusing on specific frequency components, which build richer and more refined frequency representations and validate the crucial role of the DFSM module in capturing fine-grained features. In addition, we visualize the spatial-domain representations of the convolutional kernels from different groups to further reveal their capability in extracting local features. The spatial-domain results show that the responses of kernels across groups exhibit clear differences on the image, which helps to understand their roles in capturing local details and contours.

We analyze the impact of DFSM on the feature map (as shown in Fig. 12). DFSM dynamically selects frequency domains, decomposing features into low-frequency (structural) and high-frequency (local detail) components. Through modulation and fusion, the final feature map enhances detail in blurred regions, such as textures and edges, with the heatmap showing DFSM's focus on key pedestrian features, clearly distinguishing them from the background.

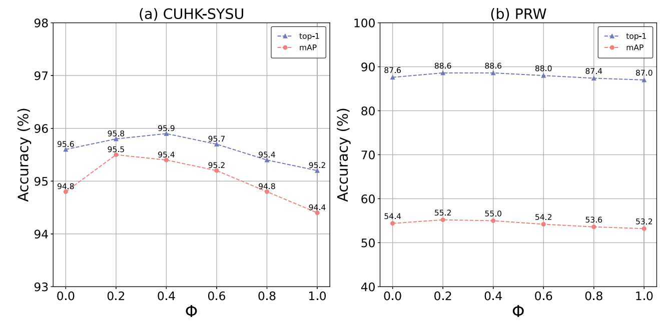

Fig. 10. Impact of Φ on CUHK-SYSU and PRW datasets.

4.7. Effectiveness of L_{TOIM}

To validate the improvement in feature discriminability and robustness brought by L_{TOIM}, we use t-SNE to compare the feature distributions with and without L_{TOIM}. Six identities are selected from the PRW dataset for analysis. As shown in Fig. 13(a), features of the same class are scattered, and the inter-class boundaries are blurred with significant overlapping regions. In contrast, Fig. 13(b) shows that features of the same class are well-clustered, and inter-class features are clearly separated, significantly enhancing feature separability. This demonstrates that L_{TOIM} effectively optimizes intra-person consistency and inter-person discriminability, thereby improving feature representation for person search tasks.

4.8. Effectiveness in cross-dataset scenarios

To validate the generalization ability of PS-DFSI, we evaluate its performance in cross-dataset scenarios, where the model is trained on a source dataset (e.g., PRW) and tested on a target dataset (e.g., CUHK-SYSU) without any fine-tuning. As shown in Table 9, PS-DFSI outperforms state-of-the-art methods in both cross-dataset scenarios. This excellent performance can be attributed to the effective integration of frequency-domain and spatial-domain features, particularly through its dynamic frequency selection and spatial interaction mechanisms, which enhance robustness in cross-domain tasks.

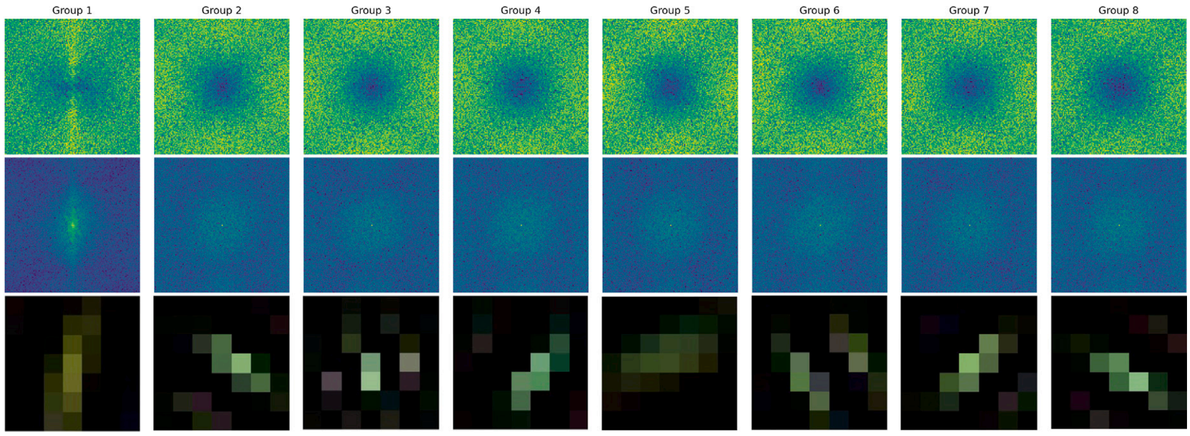

Fig. 11. Visualization of the discrete Fourier transform results from different groups generated by DFSM. Top: High frequency. Middle: Low frequency. Bottom: Spatial-domain convolutional kernels.

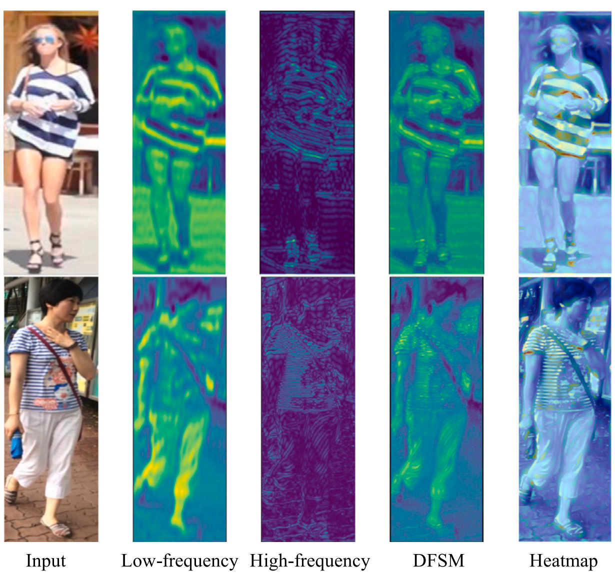

Fig. 12. Feature changes and heatmap analysis of the DFSM module. It illustrates the processing of input images in low-frequency, high-frequency, and fused features, highlighting its ability to enhance key details in pedestrian regions and its effectiveness in background separation.

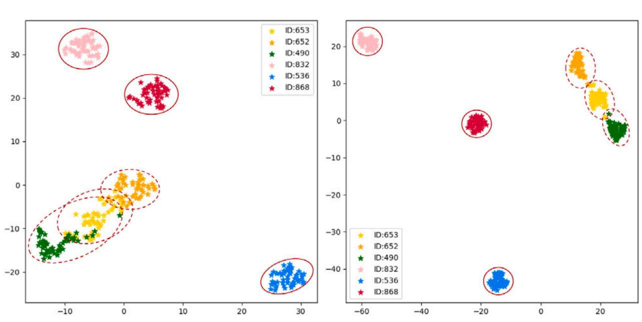

Fig. 13. Impact of L_{TOIM} loss on feature distribution. t-SNE visualization comparing feature distributions with and without L_{TOIM} loss, using six identities from PRW. The right panel highlights improved intra-person compactness and inter-person separability.

Table 9
Performance comparison on cross-dataset scenario. "PRW → CUHK-SYSU" means that the model is trained on PRW dataset while tested on CUHK-SYSU.

Methods	PRW → CUHK-SYSU		CUHK-SYSU → PRW	
	mAP	top-1	mAP	top-1
OIM [3]	49.2	54.8	20.4	42.2
SeqNet [4]	50.6	55.6	25.6	71.8
DMRNet [29]	52.1	57.5	27.4	76.5
COAT [40]	53.2	58.8	28.5	76.5
PAD [18]	53.7	58.5	29.2	76.8
PS-DFSI(Ours)	**55.2**	**60.2**	**30.4**	**77.5**

Table 10
Comparison of computational complexity.

Methods	#Params (M)	FLOPs (G)	Time (ms)
NAE+ [27]	**33**	575	98
AlignPS [35]	42	**380**	**61**
SeqNet [4]	48	550	86
COAT [40]	37	473	102
PS-DFSI(Ours)	35	452	95

4.9. Computational complexity

In Table 10, we compare the computational complexity of PS-DFSI with other end-to-end networks, reporting the number of parameters, FLOPs, and inference time (ms). All tests use 1500 × 900 input images on the same Tesla V100 GPU for fairness. While PS-DFSI has slightly higher computational cost and inference time than the anchor-free detector AlignPS [35], its complexity is much lower than SeqNet and COAT [40], which rely on sequential processing for detection and re-identification. Notably, PS-DFSI's efficient design with dynamic frequency-domain selection and spatial interaction fusion achieves superior performance while maintaining low computational cost, balancing efficiency and performance.

4.10. Effectiveness of DFSM and SFIM

To further validate the generalization and effectiveness of the proposed DFSM and SFIM modules, we evaluate them on two representative ReID datasets: Occluded-DukeMTMC and Market-1501. We integrate DFSM and SFIM into the mainstream TransReID framework and conduct controlled comparisons under four settings. As shown in Table 11, both modules individually improve the mAP and Rank-1 accuracy compared to the baseline, demonstrating their standalone effectiveness. More importantly, the best performance is achieved when DFSM and SFIM are used together, suggesting that the two modules are complementary in feature enhancement and can jointly benefit the ReID task.

Table 11
Effectiveness of DFSM and SFIM on the ReID task.

Methods	Occluded-DukeMTMC		Market-1501	
	mAP	Rank-1	mAP	Rank-1
TransReID [67]	59.2	66.4	88.9	95.2
+DFSM	61.2	70.5	89.5	95.4
+SFIM	60.7	69.7	89.3	95.4
+DFSM+SFIM	62.3	71.9	89.8	95.6

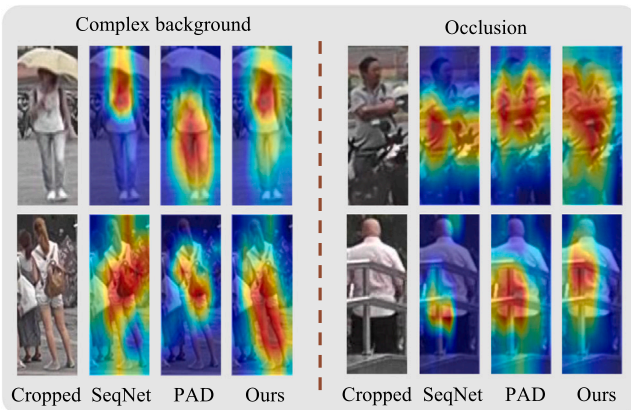

Fig. 14. Qualitative comparison of our frequency-enhanced model with SeqNet and PAD in challenging scenarios.

This experiment confirms the applicability and generalizability of our modules across different vision tasks.

4.11. Visualization under challenging scenarios

To further validate the effectiveness of frequency modeling in complex scenarios, we select samples with background clutter and occlusion from the PRW dataset and conduct a visual comparison between our frequency-enhanced model and two representative methods. As shown in Fig. 14, our method focuses more accurately on the key human regions in occluded areas and suppresses irrelevant background responses. In contrast, the baseline methods are more easily affected by background clutter or fail to precisely localize the target under occlusion. These results demonstrate that incorporating frequency-aware features helps improve the model's discriminative ability in complex scenes.

4.12. High-frequency attention ratio

We further analyze the trend of the normalized attention ratio between high-frequency and low-frequency components ($W^{high}/(W^{high}+W^{low})$) in the frequency selection module throughout the training process, as illustrated in Fig. 15. The results show that the ratio fluctuates significantly during the early training stage and shifts toward low-frequency dominance in the middle stage. In the later phase, however, the ratio gradually increases and stabilizes at a higher level, indicating that the model progressively learns to enhance high-frequency responses for more effective feature representation. This trend suggests that although the model dynamically balances high- and low-frequency information during training, it ultimately allocates more attention to high-frequency features, which helps capture fine-grained semantic details more effectively.

5. Conclusion

In this paper, we propose a novel dynamic frequency-domain selection and spatial interaction fusion network (PS-DFSI) to address the lack of robustness in person search under complex backgrounds and occlusions. PS-DFSI combines the DFSM and the SFIM to enhance feature representation and model robustness through the joint processing of frequency and spatial domains. Specifically, DFSM decomposes features into high-frequency and low-frequency components to extract critical information and suppress background noise, while SFIM enhances multi-scale perception and achieves efficient fusion of frequency and spatial domains. Experiments demonstrate that PS-DFSI achieves significant performance improvements on the CUHK-SYSU and PRW datasets, particularly excelling in cross-dataset testing.

Limitations and Future Work: Although PS-DFSI performs exceptionally well in cross-dataset experiments, its generalization ability can still be improved in scenarios with large camera viewpoint variations or severe occlusions. Moreover, the joint processing of frequency and spatial domain features may limit its applicability to real-time tasks. Future research could focus on optimizing module design and reducing computational complexity to further enhance efficiency and applicability.

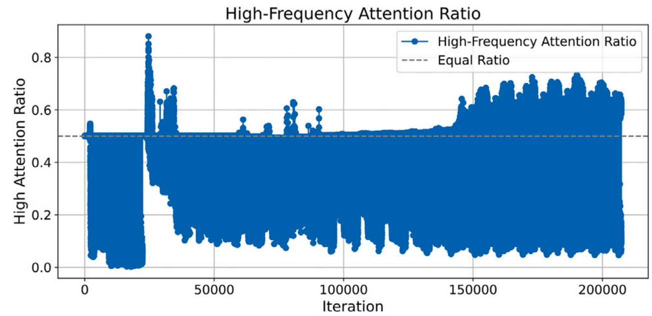

Fig. 15. Ratio of high-frequency to total attention weights ($W^{high}/(W^{high}+W^{low})$) over training iterations.

CRediT authorship contribution statement

Qixian Zhang: Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Duoqian Miao:** Supervision, Resources, Funding acquisition. **Qi Zhang:** Writing – original draft, Methodology, Conceptualization. **Cairong Zhao:** Visualization, Supervision. **Hongyun Zhang:** Supervision, Funding acquisition. **Ye Sun:** Visualization, Formal analysis. **Ruizhi Wang:** Visualization, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (Grant Nos. 62376198, 62006172, 62076182, 62163016), the National Key Research and Development Program, China (Grant No. 2022YFB3104700), the Jiangxi Double Thousand Plan, China (No. jxsq2019102088), the Jiangxi Provincial Natural Science Fund, China.

Data availability

Data will be made available on request.

References

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

[2] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, 2017, pp. 1367–1376, https://openaccess.thecvf.com/content_cvpr_2017/papers/Zheng_Person_Re-Identification_in_CVPR_2017_paper.pdf.

[3] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, 2017, pp. 3415–3424, https://openaccess.thecvf.com/content_cvpr_2017/papers/Xiao_Joint_Detection_and_CVPR_2017_paper.pdf.

[4] Z. Li, D. Miao, Sequential end-to-end network for efficient person search, in: Proc. AAAI Conf. Artif. Intell., vol. 35, (3) 2021, pp. 2011–2019, http://dx.doi.org/10.1609/aaai.v35i3.16297.

[5] Q. Zhang, J. Wu, D. Miao, C. Zhao, Q. Zhang, Attentive multi-granularity perception network for person search, Inform. Sci. 681 (2024) 121191, http://dx.doi.org/10.1016/j.ins.2024.121191.

[6] Q. Zhang, D. Miao, Q. Zhang, C. Wang, Y. Li, H. Zhang, C. Zhao, Learning adaptive shift and task decoupling for discriminative one-step person search, Knowl.-Based Syst. 304 (2024) 112483, http://dx.doi.org/10.1016/j.knosys.2024.112483.

[7] S. Park, H. Kim, Y.M. Ro, Integrating language-derived appearance elements with visual cues in pedestrian detection, IEEE Trans. Circuits Syst. Video Technol. 34 (9) (2024) 7975–7985, http://dx.doi.org/10.1109/TCSVT.2024.3383914.

[8] X. Tan, X. Gong, Y. Xiang, CLIP-based camera-agnostic feature learning for intra-camera supervised person re-identification, IEEE Trans. Circuits Syst. Video Technol. 34 (2024) http://dx.doi.org/10.1109/TCSVT.2024.3522178, Early Access.

[9] M. Yu, Y. Ge, Z. Chen, R. You, L. Zhu, M. Lin, Z. Xu, No escape: Towards suggestive clues guidance for cross-modality person re-identification, Inf. Fusion 122 (2025) 103185, http://dx.doi.org/10.1016/j.inffus.2025.103185.

[10] Q. Min, F. Luo, W. Ding, Bidirectional domain transfer knowledge distillation for catastrophic forgetting in federated learning with heterogeneous data, Knowl.-Based Syst. 311 (2025) 113008, http://dx.doi.org/10.1016/j.knosys.2025.113008.

[11] X. Gong, X. Tan, Y. Xiang, Contrastive mean teacher for intra-camera supervised person re-identification, IEEE Trans. Circuits Syst. Video Technol. 34 (10) (2024) 9786–9797, http://dx.doi.org/10.1109/TCSVT.2024.3402533.

[12] T. Zhao, Y. Zhang, D. Miao, Granular correlation-based label-specific feature augmentation for multi-label classification, Inform. Sci. 689 (2025) 121473, http://dx.doi.org/10.1016/j.ins.2024.121473.

[13] S. Huang, Y. Wang, H. Luo, CCSUMSP: A cross-subject Chinese speech decoding framework with unified topology and multi-modal semantic pre-training, Inf. Fusion. 119 (2025) 103022, http://dx.doi.org/10.1016/j.inffus.2025.103022.

[14] L. Jaffe, A. Zakhor, Gallery filter network for person search, in: Proc. IEEE Winter Conf. Appl. Comput. Vis., WACV, 2023, pp. 1684–1693, https://openaccess.thecvf.com/content/WACV2023/papers/Jaffe_Gallery_Filter_Network_for_Person_Search_WACV_2023_paper.pdf.

[15] C. Han, J. Ye, Y. Zhong, X. Tan, C. Zhang, C. Gao, N. Sang, Re-id driven localization refinement for person search, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV, 2019, pp. 9814–9823, https://openaccess.thecvf.com/content_ICCV_2019/papers/Han_Re-ID_Driven_Localization_Refinement_for_Person_Search_ICCV_2019_paper.pdf.

[16] C. Wang, B. Ma, X. Chen, TCTS: A task-consistent two-stage framework for person search, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2020, pp. 11952–11961, https://openaccess.thecvf.com/content_CVPR_2020/papers/Wang_TCTS_A_Task-Consistent_Two-Stage_Framework_for_Person_Search_CVPR_2020_paper.pdf.

[17] D. Chen, S. Zhang, W. Ouyang, J. Yang, B. Schiele, Hierarchical online instance matching for person search, in: Proc. AAAI Conf. Artif. Intell., vol. 34, (7) 2020, pp. 10518–10525, http://dx.doi.org/10.1609/aaai.v34i07.6623.

[18] H. Kim, J. Lee, K. Sohn, Prototype-guided attention distillation for discriminative person search, IEEE Trans. Pattern Anal. Mach. Intell. 47 (1) (2025) 99–115, http://dx.doi.org/10.1109/TPAMI.2024.3461778.

[19] G. Zhang, Y. Zhang, T. Zhang, B. Li, S. Pu, PHA: Patch-wise high-frequency augmentation for transformer-based person re-identification, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, CVPR, 2023, pp. 14133–14142, https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_PHA_Patch-Wise_High-Frequency_Augmentation_for_Transformer-Based_Person_Re-Identification_CVPR_2023_paper.html.

[20] C. Li, S. Chen, M. Ye, Adaptive high-frequency transformer for diverse wildlife re-identification, in: Proc. Eur. Conf. Comput. Vis., ECCV, Springer, 2025, pp. 296–313, https://link.springer.com/chapter/10.1007/978-3-031-72784-9_17.

[21] K. Yi, Q. Zhang, W. Fan, S. Wang, P. Wang, N. An, L. Cao, Frequency-domain MLPs are more effective learners in time series forecasting, in: Proc. Adv. Neural Inf. Process. Syst., vol. 36, 2024, https://proceedings.neurips.cc/paper_files/paper/2023/hash/f1d16af76939f476b5f040fd1398c0a3-Abstract-Conference.html.

[22] Y. Yang, G. Yuan, J. Li, SFFNet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation, IEEE Trans. Geosci. Remote Sens. 62 (2024) http://dx.doi.org/10.1109/TGRS.2024.3427370, Art. no. 3000617.

[23] X. Lan, X. Zhu, S. Gong, Person search by multi-scale matching, in: Proc. Eur. Conf. Comput. Vis., ECCV, 2018, pp. 536–552, https://openaccess.thecvf.com/content_ECCV_2018/html/Xu_Lan_Person_Search_by_ECCV_2018_paper.html.

[24] D. Chen, S. Zhang, W. Ouyang, J. Yang, Y. Tai, Person search via a mask-guided two-stream cnn model, in: Proc. Eur. Conf. Comput. Vis., ECCV, 2018, pp. 734–750, https://openaccess.thecvf.com/content_ECCV_2018/papers/Di_Chen_Person_Search_via_ECCV_2018_paper.pdf.

[25] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2016) 1137–1149, http://dx.doi.org/10.1109/TPAMI.2016.2577031.

[26] W. Dong, Z. Zhang, C. Song, T. Tan, Bi-directional interaction network for person search, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2020, pp. 2839–2848, https://openaccess.thecvf.com/content_CVPR_2020/html/Dong_Bi-Directional_Interaction_Network_for_Person_Search_CVPR_2020_paper.html.

[27] D. Chen, S. Zhang, J. Yang, B. Schiele, Norm-aware embedding for efficient person search, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, CVPR, 2020, pp. 12615–12624, https://openaccess.thecvf.com/content_CVPR_2020/html/Chen_Norm-Aware_Embedding_for_Efficient_Person_Search_CVPR_2020_paper.html.

[28] B. Munjal, S. Amin, F. Tombari, F. Galasso, Query-guided end-to-end person search, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2019, pp. 811–820, https://openaccess.thecvf.com/content_CVPR_2019/html/Munjal_Query-Guided_End-To-End_Person_Search_CVPR_2019_paper.html.

[29] C. Han, Z. Zheng, K. Su, D. Yu, Z. Yuan, C. Gao, N. Sang, Y. Yang, DMRNet++: Learning discriminative features with decoupled networks and enriched pairs for one-step person search, IEEE Trans. Pattern Anal. Mach. Intell. 45 (6) (2023) 7319–7337, http://dx.doi.org/10.1109/TPAMI.2022.3221079.

[30] Y. Jiang, H. Wang, J. Peng, X. Fu, Y. Wang, Scene-adaptive person search via bilateral modulations, in: Proc. Int. Joint Conf. Artif. Intell., 2024, https://arxiv.org/pdf/2405.02834.

[31] X. Yang, M. Tian, N. Wang, X. Gao, Unleashing the feature hierachy potential: An efficient tri-hybrid person search model, IEEE Trans. Circuits Syst. Video Technol. 34 (11) (2024) 11551–11563, http://dx.doi.org/10.1109/TCSVT.2024.3424261.

[32] Y. Tian, D. Chen, Y. Liu, J. Yang, S. Zhang, Divide and conquer: Hybrid pre-training for person search, in: Proc. AAAI Conf. Artif. Intell., vol. 38, (6) 2024, pp. 5224–5232, http://dx.doi.org/10.1609/aaai.v38i6.28329.

[33] Y. Jia, R. Quan, Y. Feng, H. Chen, J. Qin, Doubly contrastive learning for source-free domain adaptive person search, in: Proc. AAAI Conf. Artif. Intell., vol. 39, (4) 2025, pp. 3949–3957, http://dx.doi.org/10.1609/aaai.v39i4.32413.

[34] H. Zhu, X. Yang, N. Wang, Optimizing label assignment for weakly supervised person search, in: Proc. AAAI Conf. Artif. Intell., vol. 39, (10) 2025, pp. 10941–10949, http://dx.doi.org/10.1609/aaai.v39i10.33189.

[35] Y. Yan, J. Li, J. Qin, S. Bai, S. Liao, L. Liu, F. Zhu, L. Shao, Anchor-free person search, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, CVPR, 2021, pp. 7690–7699, https://openaccess.thecvf.com/content/CVPR2021/html/Yan_Anchor-Free_Person_Search_CVPR_2021_paper.html?ref=https://githubhelp.com.

[36] Z. Tian, C. Shen, H. Chen, T. He, FCOS: Fully convolutional one-stage object detection, in: Proc. IEEE/CVF Int. Conf. Comput. Vis, ICCV, 2019, pp. 9627–9636, https://openaccess.thecvf.com/content_ICCV_2019/html/Tian_FCOS_Fully_Convolutional_One-Stage_Object_Detection_ICCV_2019_paper.html.

[37] A. Dosovitskiy, L. Beyer, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: Proc. Int. Conf. Learn. Representations, 2021, https://arxiv.org/pdf/2010.11929/1000.

[38] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Proc. Eur. Conf. Comput. Vis., ECCV, Springer, 2020, pp. 213–229, https://link.springer.com/chapter/10.1007/978-3-030-58452-8_13.

[39] J. Cao, Y. Pang, R.M. Anwer, H. Cholakkal, J. Xie, M. Shah, F.S. Khan, PSTR: End-to-end one-step person search with transformers, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, CVPR, 2022, pp. 9458–9467, https://openaccess.thecvf.com/content/CVPR2022/html/Cao_PSTR_End-to-End_One-Step_Person_Search_With_Transformers_CVPR_2022_paper.html.

[40] R. Yu, D. Du, D. Davila, C. Funk, A. Hoogs, B. Clipp, Cascade transformers for end-to-end person search, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2022, pp. 7267–7276, https://openaccess.thecvf.com/content/CVPR2022/html/Yu_Cascade_Transformers_for_End-to-End_Person_Search_CVPR_2022_paper.html.

[41] M. Fiaz, H. Cholakkal, R.M. Anwer, F.S. Khan, SAT: scale-augmented transformer for person search, in: Proc. IEEE Winter Conf. Appl. Comput. Vis., WACV, 2023, pp. 4820–4829, https://openaccess.thecvf.com/content/WACV2023/html/Fiaz_SAT_Scale-Augmented_Transformer_for_Person_Search_WACV_2023_paper.html.

[42] Z. Cai, N. Vasconcelos, Cascade R-CNN: Delving into high quality object detection, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6154–6162, https://openaccess.thecvf.com/content_cvpr_2018/html/Cai_Cascade_R-CNN_Delving_CVPR_2018_paper.html.

[43] E. Oyallon, E. Belilovsky, S. Zagoruyko, Scaling the scattering transform: Deep hybrid networks, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV, 2017, pp. 5618–5627, https://openaccess.thecvf.com/content_iccv_2017/html/Oyallon_Scaling_the_Scattering_ICCV_2017_paper.html.

[44] Y. Rao, W. Zhao, Z. Zhu, J. Lu, J. Zhou, Global filter networks for image classification, in: Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 980–993, https://proceedings.neurips.cc/paper/2021/hash/07e87c2f4fc7f7c96116d8e2a92790f5-Abstract.html.

[45] A. Lao, Q. Zhang, C. Shi, L. Cao, K. Yi, L. Hu, D. Miao, Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector, in: Proc. AAAI Conf. Artif. Intell., vol. 38, (16) 2024, pp. 18426–18434, http://dx.doi.org/10.1609/aaai.v38i16.29803.

[46] T. Qin, J. Zhu, Z. Li, X. Hu, A.M. Mostafa, Path detectability verification for time-dependent systems with application to flexible manufacturing systems, Inform. Sci. 689 (2025) 121404, http://dx.doi.org/10.1016/j.ins.2024.121404.

[47] Y. Zhang, T. Zhao, D. Miao, Y. Yao, Three-way multi-label classification: A review, a framework, and new challenges, Appl. Soft Comput. 171 (2025) 112757, http://dx.doi.org/10.1016/j.asoc.2025.112757.

[48] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, 2017, pp. 2117–2125, https://openaccess.thecvf.com/content_cvpr_2017/html/Lin_Feature_Pyramid_Networks_CVPR_2017_paper.html.

[49] L.-C. Chen, Y. Zhu, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proc. Eur. Conf. Comput. Vis., ECCV, 2018, pp. 801–818, https://openaccess.thecvf.com/content_ECCV_2018/html/Liang-Chieh_Chen_Encoder-Decoder_with_Atrous_ECCV_2018_paper.html.

[50] J. Wang, K. Sun, Y. Mu, M. Tan, X. Wang, Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2020) 3349–3364, http://dx.doi.org/10.1109/TPAMI.2020.2983686.

[51] Q. Wang, W. Zuo, ECA-Net: Efficient channel attention for deep convolutional neural networks, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2020, pp. 11534–11542, https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_ECA-Net_Efficient_Channel_Attention_for_Deep_Convolutional_Neural_Networks_CVPR_2020_paper.html.

[52] Z. Liu, Y. Lin, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV, 2021, pp. 10012–10022, https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.

[53] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proc. IEEE Int. Conf. Comput. Vis., ICCV, 2017, pp. 2961–2969, https://openaccess.thecvf.com/content_iccv_2017/html/He_Mask_R-CNN_ICCV_2017_paper.html.

[54] S.U. Pillai, S. Cha, The perron-frobenius theorem: some of its applications, IEEE Signal Process. Magaz 22 (2005) 62–75, http://dx.doi.org/10.1109/MSP.2005.1406483.

[55] K. Yuan, D. Miao, W. Pedrycz, H. Zhang, L. Hu, Multigranularity data analysis with zentropy uncertainty measure for efficient and robust feature selection, IEEE T. Cybern. 55 (2) (2025) 740–752, http://dx.doi.org/10.1109/TCYB.2024.3499952.

[56] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2022, pp. 11976–11986, https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html.

[57] W. Dong, Z. Zhang, C. Song, T. Tan, Instance guided proposal network for person search, in: Proc. IEEE/CVF Conf. Comput Vis. Pattern Recognit., CVPR, 2020, pp. 2585–2594, https://openaccess.thecvf.com/content_CVPR_2020/html/Dong_Instance_Guided_Proposal_Network_for_Person_Search_CVPR_2020_paper.html.

[58] H. Yao, C. Xu, Joint person objectness and repulsion for person search, IEEE Trans. Image Process. 30 (2021) 685–696, http://dx.doi.org/10.1109/TIP.2020.3038347.

[59] X. Chang, P.-Y. Huang, X. Liang, Y. Yang, A.G. Hauptmann, RCAA: Relational context-aware agents for person search, in: Proc. Eur. Conf. Comput. Vis., ECCV, 2018, pp. 84–100, https://openaccess.thecvf.com/content_ECCV_2018/html/Xiaojun_Chang_RCAA_Relational_Context-Aware_ECCV_2018_paper.html.

[60] Y. Zhong, X. Wang, S. Zhang, Robust partial matching for person search in the wild, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, CVPR, 2020, pp. 6827–6835, https://openaccess.thecvf.com/content_CVPR_2020/html/Zhong_Robust_Partial_Matching_for_Person_Search_in_the_Wild_CVPR_2020_paper.html.

[61] S. Lee, Y. Oh, J. Lee, B. Ham, Oimnet++: Prototypical normalization and localization-aware learning for person search, in: Proc. Eur. Conf. Comput. Vis., ECCV, 2022, pp. 621–637, https://link.springer.com/chapter/10.1007/978-3-031-20080-9_36.

[62] W. Chen, X. Xu, J. Jia, H. Luo, Y. Wang, X. Sun, Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2023, pp. 15050–15061, https://openaccess.thecvf.com/content/CVPR2023/html/Chen_Beyond_Appearance_A_Semantic_Controllable_Self-Supervised_Learning_Framework_for_Human-Centric_CVPR_2023_paper.html.

[63] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pvt v2: Improved baselines with pyramid vision transformer, Comput. Vis. Media 8 (3) (2022) 415–424, https://link.springer.com/article/10.1007/s41095-022-0274-8.

[64] Q. Zhang, X. Zhu, Y. Liu, Iris recognition based on adaptive optimization log-gabor filter and rbf neural network, in: Proc. Chin. Conf. Biometr. Recognit., CCBR, 2019, pp. 312–320, http://dx.doi.org/10.1007/978-3-030-31456-9_35.

[65] Y. Chen, X. Dai, L. Yuan, Z. Liu, Dynamic convolution: Attention over convolution kernels, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2020, pp. 11030–11039, https://openaccess.thecvf.com/content_CVPR_2020/papers/Chen_Dynamic_Convolution_Attention_Over_Convolution_Kernels_CVPR_2020_paper.pdf.

[66] K. Yuan, D. Miao, W. Pedrycz, Ze-HFS: Zentropy-based uncertainty measure for heterogeneous feature selection and knowledge discovery, IEEE Trans. Knowl. Data En. 36 (11) (2024) 7326–7339, http://dx.doi.org/10.1109/TKDE.2024.3419215.

[67] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, Transreid: Transformer-based object re-identification, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV, 2021, pp. 15013–15022, https://openaccess.thecvf.com/content/ICCV2021/papers/He_TransReID_Transformer-Based_Object_Re-Identification_ICCV_2021_paper.pdf.