Expert Systems With Applications 308 (2026) 131090

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Frequency-aware and lifting-based efficient transformer for person search

Qilin Shu [a,b,1], Qixian Zhang [a,b,1], Duoqian Miao [a,b,*], Qi Zhang [a,b], Hongyun Zhang [a,b], Cairong Zhao [a,b]

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

ARTICLE INFO

Keywords:
Person search
Learnable lifting scheme
High-pass filtering
Model efficiency
Parameter sharing

ABSTRACT

The person search task aims to locate a target person within a set of scene images. In recent years, transformer-based models in this field have made some progress. However, they still face two primary challenges: 1) the self-attention mechanism tends to suppress high-frequency components in the features, which severely impacts model performance; 2) the computational cost of transformers is relatively high. To address these issues, we propose a novel Frequency-Aware and Lifting-Based Efficient Transformer (FLET) method for person search. FLET is designed to enhance the discriminative feature extraction capabilities of transformers while reducing computational overhead and improving efficiency. Specifically, we develop a three-stage framework that progressively optimizes both detection and re-identification performance. Our model enhances the perception of high-frequency features by learning from augmented inputs. The augmented inputs are generated via High-Pass Filtering (HPF) and contain additional high-frequency components. Furthermore, we replace the self-attention layers in the transformer with a Learnable Lifting Block (LLB) to capture multiscale features. LLB not only lowers the computational complexity but also alleviates the suppression of high-frequency features and enhances the ability to exploit multiscale information. Extensive experiments demonstrate that FLET achieves state-of-the-art performance on both the CUHK-SYSU and PRW datasets.

1. Introduction

Person search (Li & Miao, 2021; Xiao et al., 2017; Zhang et al., 2024a, 2025, 2024b; Zheng et al., 2017) aims to detect and identify a query person from a large collection of scene images, typically captured in unconstrained environments. The task comprises two subtasks: pedestrian detection and person re-identification (ReID) (Dou et al., 2023). Pedestrian detection focuses on identifying all persons in the scene and generating bounding-box proposals. ReID is responsible for identifying the target individual from foregoing proposals.

Existing methods for person search can be categorized into two step and one-step approaches. The two-step methods (Chen et al., 2018; Dong et al., 2020b; Lan et al., 2018; Zheng et al., 2017) utilize two separate models to address two subtasks, and then combine the results. Typically, a detector is tasked with locating and cropping the person patches, which are then fed into a ReID model for identification. Corresponding to the simplicity in concept and architecture, these approaches exhibit significant disadvantages in computational efficiency and performance. Alternatively, one-step approaches architecture, these approaches exhibit significant (Dong et al., 2020a; Munjal et al., 2019; Xiao et al., 2017; Yan et al., 2021) perform both detection and ReID within a single model. Given the original scene image, the model can identify all potential persons at once. Without a doubt, compared to two-step methods, one-step approaches demonstrate superior performance and greater scalability.

In recent years, transformer-based methods have emerged and achieved state-of-the-art results in both person re-identification and person search, often surpassing CNN-based baselines (Ye et al., 2024). Compared to CNN-based approaches, transformer-based models offer unique advantages in terms of discriminative feature learning, global dependencies modeling and robustness to occlusion, pose variation, and scale changes (Chen et al., 2024; Hu et al., 2024; Ji et al., 2025; Liu et al., 2025). Nevertheless, the application of transformers faces several challenges, primarily in the following two aspects:

Limited ability to capture high-frequency components. As illustrated in Fig. 1, high-frequency information refers to regions of rapid pixel variation, such as edges and textures, that are crucial for preserving fine-grained details. Studies have shown that the self-attention

* Corresponding author.
E-mail addresses: 2331950@tongji.edu.cn (Q. Shu), zhangqx@tongji.edu.cn (Q. Zhang), dqmiao@tongji.edu.cn (D. Miao), zhangqi_cs@tongji.edu.cn (Q. Zhang), zhanghongyun@tongji.edu.cn (H. Zhang), zhaocairong@tongji.edu.cn (C. Zhao).
[1] The authors contributed equally to this work.

https://doi.org/10.1016/j.eswa.2026.131090
Received 13 September 2025; Received in revised form 13 December 2025; Accepted 1 January 2026
Available online 9 January 2026

Fig. 1. Illustration of low-frequency and high-frequency information in person images. In person search, high-frequency components—such as image details, edges, and textures—play a critical role in both pedestrian detection and re-identification. This motivates our introduction of high-pass filtering (HPF) and frequency-aware learning to enhance the model's sensitivity to such details.

mechanism in transformers tends to attenuate high-frequency signals, leading to performance degradation (Zhang et al., 2023).

High computational complexity. The quadratic computational cost of self-attention with respect to the input sequence length has prompted the development of techniques such as low-rank and sparse attention. However, in the person search domain, existing methods have yet to effectively address the issue.As a result, transformer-based models offer limited inference speed advantages compared to their CNN counterparts.

To address the limitations, we propose a novel transformer-based framework for person search, which called FLET. Inspired by COAT (Yu et al., 2022), our FLET adopts a three-stage design that progressively refines the detection and ReID performance. The first stage is dedicated to person detection, distinguishing persons from the background. In the second and third stages, ReID and frequency-aware proxy loss (FP Loss) are introduced to further enhance identity discrimination. In each stage, we incorporate a branch-filtering transformer, in which self-attention layer is replaced with a learnable lifting block.This design enables multiscale feature extraction, thereby improving discriminability while reducing computational cost.

Moreover, the transformers in the second and third stages introduce high-pass filtering(HPF) and a corresponding frequency-aware proxy loss. Tokens augmented with high-frequency components are guided by a proxy-based loss.This supervision boosts the model's sensitivity to fine-grained details, enhancing its ability to capture and utilize high-frequency information. Experiments on the CUHK-SYSU (Xiao et al., 2017) and PRW (Zheng et al., 2017) datasets confirm that our model achieves state-of-the-art performance.

Our main contributions can be summarized as follows.

1) We propose a novel transformer-based framework for person search that enhances the perception and utilization of high-frequency features while reducing computational complexity.
2) We replace standard self-attention layers with learnable lifting blocks, which effectively reduce complexity and improve multiscale feature representation.
3) We introduce high-pass filtering augmentation and a frequency-aware proxy loss. These components strengthen the model's ability to represent high-frequency features by bringing the tokens of the same identity and their enhanced counterparts closer while pushing apart those of different identities.
4) Extensive experiments on two benchmark datasets (CUHK-SYSU and PRW) demonstrate that FLET achieves 96.4% mAP and 97.3% top-1 accuracy on CUHK-SYSU, and 58.2% mAP and 91.2% top-1 accuracy on PRW, outperforming previous methods. Moreover, experiments also demonstrate that the computational efficiency of FLET substantially exceeds that of all existing methods.

The rest of this article is structured as follows. Section 2 surveys the relevant literature, Section 3 elaborates on the methodological framework, Section 4 outlines the experimental design and analyzes the findings, and Section 5 provides the conclusion.

2. Related work

2.1. Person search

Person search methods can be broadly categorized into two main paradigms based on their architectural design.

1) Two-step methods: These approaches divide the person search task into two independent subtasks: pedestrian detection and person re-identification (ReID), each handled by separately trained models. Zheng et al. (2017) were the first to combine different pedestrian detectors with various ReID models, pioneering this research direction. Chen et al. (2018) proposed the Mask-Guided Two-Stream CNN Model (MGTS) and highlighted the issue of target conflict. Lan et al. (2018) introduced Cross-Level Semantic Alignment (CLSA) in the ReID stage to address the multiscale feature alignment problem. Dong et al. (2020b) proposed the Instance-guided Proposal Network (IGPN), which incorporates query information into the detection module.

2) One-step methods: These methods integrate detection and ReID into a single end-to-end trainable framework. They generally offer lower computational cost and higher efficiency. Xiao et al. (2017) first introduced an end-to-end model based on Faster R-CNN. Dong et al. (2020a) designed a bidirectional interaction network to suppress redundant contextual information. Munjal et al. (2019) were the first to incorporate query information, proposing the Query-Guided End-to-End Person Search (QEEPS) framework. Yan et al. (2021) developed AlignPS, an anchor-free person search model that reduces computational overhead while addressing the challenges of multiscale and region misalignment in feature reconstruction. Li and Miao (2021) introduced SeqNet, a sequential framework using two Faster R-CNNs to handle detection and ReID in order. Jaffe and Zakhor (2023) further extended SeqNet with SeqNeXt by replacing Faster R-CNN with ConvNeXt and reducing similar background scenarios to lower the search space and computation. Zhang et al.(Zhang et al., 2024a,b) proposed AMPN and ASTD frameworks to tackle scale variation, occlusion, and pose changes, achieving state-of-the-art performance. Additionally, Zhang et al. (2025) introduced PS-DFSI, which fuses frequency and spatial information to enhance robustness in person search. Jiang et al. (2024) propose Scene-Adaptive Person Search (SEAS), which introduces bilateral modulation networks to suppress background and foreground noise.This design enables SEAS to achieve scene-invariant person representations and state-of-the-art performance.

2.2. Transformers in person search and ReID

Since Vision Transformers (ViT) demonstrated their potential in computer vision tasks, many subfields such as person search and person re-identification have begun to explore their application. He et al. (2021) were the first to adopt a pure transformer architecture for ReID. They use shuffled and rearranged patch embeddings to generate more discriminative and robust features. Wang et al. (2022a) proposed NFormer, which enhances robustness and discrimination by explicitly modeling relationships between all input images. Luo et al. (2020) combined spatial transformer networks (STN) with a ReID module in STNReID.This approach mitigates the misalignment between partial and holistic images that occurs during alignment and feature extraction. Li et al. (2021) introduced the PAT model, the first to apply transformer architecture to occluded person ReID, also leveraging weakly supervised learning. Cao et al. (2022) proposed PSTR, the first transformer-based one-step person search framework.They introduce a person search-specialized module with a discriminative re-id decoder and multiscale design to jointly optimize detection and re-identification. Yu et al. (2022) proposed COAT, a cascaded transformer framework for end-to-end person search.The model progressively balances detection

and re-identification via a three-stage coarse-to-fine design.It also introduces an occluded attention mechanism to enhance robustness against scale variation and occlusion. Kim et al. (2025) proposed PAD, which distills prototype-guided attention into Re-ID queries and employs multiple part prototypes with adaptive momentum to address misalignment and intra-class variation.

2.3. Improvements to self-attention

Although self-attention forms the core of transformer architectures, it incurs high computational cost and struggles to capture local textures and multiscale information. Numerous studies have proposed modifications to mitigate these issues. D'Ascoli et al. (2021) introduced Gated Positional Self-Attention (GPSA), which initializes as a convolutional layer and gradually learns attention through gating. Dai et al. (2021) proposed CoAtNet, which combining DW-Conv with multi-head attention, arranged in alternating layers. Liu et al. (2021) developed window-based self-attention by partitioning the image into non-overlapping local windows and cyclically shifting them to enable cross-window interaction. As a result, the computational complexity becomes linear with respect to the image size. Wang et al. (2022b) constructed a Pyramid Vision Transformer (PVT), which progressively reduces token length while employing Spatial-Reduction Attention (SRA) to cut computation. These designs enable PVT to support high-resolution dense prediction while maintaining global receptive fields. Spravil et al. (2024) introduced HyenaPixel, which extends the Hyena operator to bidirectional and 2D image modeling with extremely large convolutional kernels. This extension enables global context aggregation at sub-quadratic complexity.

Tolstikhin et al. (2021) proposed a pure MLP-based vision model (MLP-Mixer), which alternates between token-mixing and channel-mixing MLPs to aggregate spatial and channel-wise information. Under large-scale pretraining, MLP-Mixer reaches performance comparable to transformers. Rao et al. (2021) introduced GFNet, which replacing self-attention with global filtering using 2D FFT and learned frequency-domain filters, followed by an inverse FFT to return to the spatial domain. This approach shows strong potential as a transformer alternative on ImageNet and downstream tasks.

In contrast, our FLET retains the cascaded architecture of COAT and introduces High-Pass Filtering(HPF) to augment the input with additional high-frequency content. This enhancement improves the model's ability to perceive fine-grained features. Furthermore, we replace the traditional self-attention layers in transformers with learnable lifting block. This substitution preserves the global receptive field while significantly reducing computational cost. The hierarchical structure also provides additional advantages in capturing multiscale information effectively.

3. Methodology

3.1. Overall architecture

The overall architecture proposed in this article is based on COAT (Yu et al., 2022), a transformer-based multistage framework for person search. COAT exhibits excellent performance in handling occlusion and scale variation. Its effectiveness lies in its unique occlusion-aware attention mechanism. During training, its transformer exchanges tokens among all proposals within a batch to simulate real-world occlusions. This strategy enables the model to better adapt to occluded scenarios, learning more robust and discriminative features.

Furthermore, COAT adopts a three-stage cascade design inspired by Cascade R-CNN (Cai & Vasconcelos, 2018), allowing the model to refine detection and ReID results progressively from coarse to fine. This approach helps mitigate the conflict between detection and ReID tasks and enhances overall performance. Together, COAT serves as an ideal and promising baseline with ample room for further improvement.

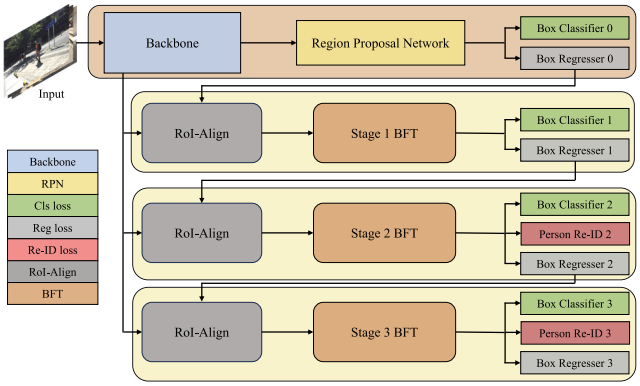


Fig. 2. Overall architecture of the proposed FLET framework. The model adopts a three-stage cascaded design, where each stage progressively refines detection and re-identification features. A backbone and RPN generate initial proposals, followed by Branch-Filtering Transformers (BFTs) and ReID heads that enhance multiscale and high-frequency feature representation. This coarse-to-fine structure improves both localization accuracy and identity discrimination.

In our model, we retain the cascade structure and token-exchange mechanism. To reduce computational complexity and enhance multiscale feature extraction, we replace the multi-head self-attention layers in the encoder with learnable lifting block. Additionally, we introduce a new auxiliary branch and a frequency-aware proxy loss during proposal processing to enhance the model's ability to represent high-frequency components.

As shown in Fig. 2, our framework first extracts features using ResNet-50 (He et al., 2016), then generates candidate proposals via a Region Proposal Network (RPN) (Ren et al., 2016). RoI-Align (Ren et al., 2016) is used to pool these proposals of varying sizes into a uniform size. Subsequently, a multistage cascade strategy is employed to progressively refine the representations and obtain more precise detection and ReID results. The first stage includes only classification and regression branches, while the second and third stages incorporate an additional ReID branch.

3.2. Branch-filtering transformer

One limitation of transformer architecture is that as the number of layers increases, high-frequency components tend to be gradually diluted (Zhang et al., 2023). The lack of fine-grained details significantly deteriorates model performance. To address the issue, we propose a frequency-aware proxy loss in the Branch-Filtering Transformer (BFT) to help the model better extract high-frequency features. The architecture of this component is depicted in Fig. 3. For feature maps input into the BFT, we perform high-pass filtering(HPF). Specifically, for a feature tensor $x \in \mathcal{R}^{h\times w\times c}$ we convert it into a single-channel tensor, and then apply the Fast Fourier Transform(FFT) to it to obtain $F(x)$. We first apply a Gaussian high-pass filter $F(x)$ to obtain $F_h(x)$, which retains only the high-frequency components. Next, we compute $F^{'}(x)$ according to the following formula for continued processing:

$$F^{'}(x) = (1-c)F(x) + cF_h(x) \tag{1}$$

where c is the enhancement coefficient. Next, we apply the inverse Fast Fourier Transform (IFFT) and reshaping to convert $F^{'}(x)$ back into the spatial domain. Fig. 4 briefly illustrates the entire process.

Both the original feature maps and the high-frequency enhanced version are sliced along the channel dimension. Each slice is processed by separate convolution layers with distinct parameters, followed by reshaping and reassembling into token maps. Due to these diverse convolutions, each token encodes information from different scales. Then, a token exchange occurs within the same batch to simulate occlusion scenarios and enhance robustness. The token exchange mechanism is

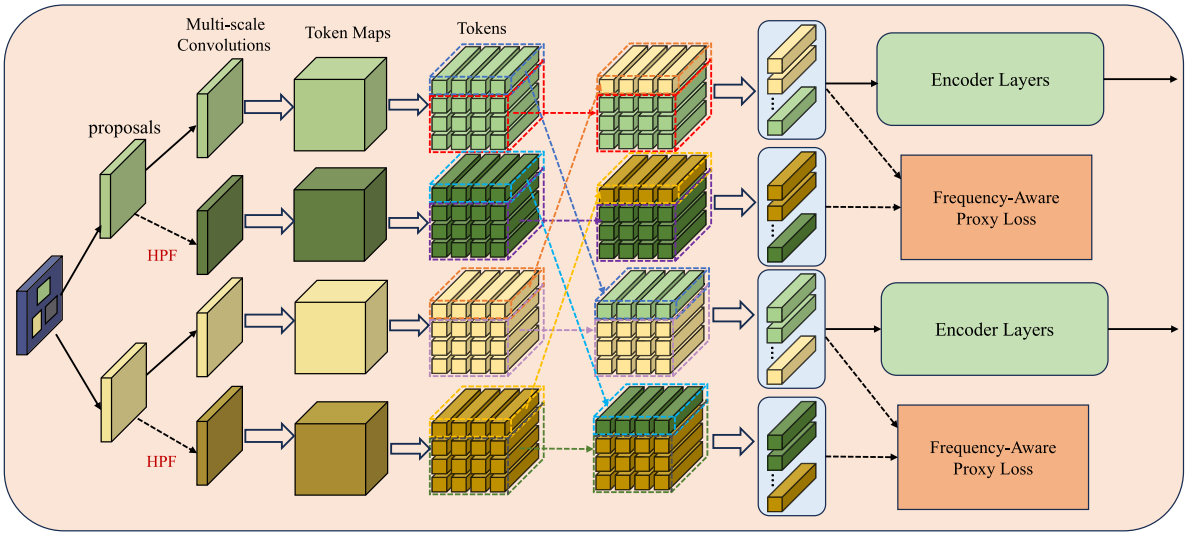


Fig. 3. Overview of the Branch-Filtering Transformer (BFT). The BFT module enhances high-frequency components through high-pass filtering (HPF). HPF generates original and enhanced token maps, and performs intra-batch token exchange to simulate occlusion. These enriched tokens are then fed into the encoder to improve fine-grained feature representation for person search. Note that all operations indicated by black dashed lines are enabled only during the second and third stages of training.

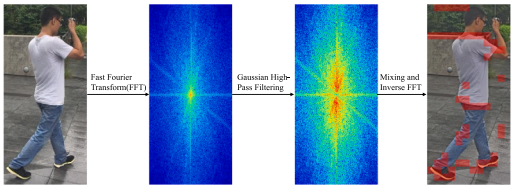


Fig. 4. Illustration of the high-pass filtering (HPF) process. The HPF branch injects additional high-frequency components by applying FFT, Gaussian high-pass filtering, and inverse FFT. This process produces enhanced tokens that strengthen the model's sensitivity to fine-grained details. The red blocks in the figure represent the high-frequency components in the reconstructed image.

inherited directly from COAT, and all token maps participate in this process. The exchange takes place within a randomly selected horizontal or vertical strip that is shared across the entire batch. For each token map, the tokens in this strip are replaced by tokens from another randomly sampled token map in the batch. The source of the tokens being replaced and the destination receiving its own tokens are generally not the same. Note that original token maps exchange only with other original token maps, while enhanced token maps exchange only among themselves.

Next, we emphasize frequency-aware proxy loss. For the sample x in a batch, if a matching identity V_y exists in V, the loss $\mathcal{L}_P$ is calculated and the corresponding entry is updated as:

$$v_y \leftarrow \lambda v_y + (1-\lambda)x \tag{2}$$

where λ is the momentum coefficient. If no matching identity is found in V, and the identity of x is known, add x to V. Otherwise, insert x into the FIFO queue Q.

To compute the loss $\mathcal{L}_P$, we define:

$$S(x,V,i) = \exp\left(\frac{\mathbf{x_k}}{||\mathbf{x_k}||_2}^{\top} \cdot \frac{V_{i,k}^h}{\left|V_{i,k}^h\right|_2}\right) \tag{3}$$

Let x_k be the token in feature map x ranked k-th by L_2 norm, and $V_{i,k}^h$ be the high-frequency enhanced token ranked k-th in the table or queue for the i-th feature map. Then, the frequency-aware proxy loss is computed as:

$$\mathcal{L}_P = -\sum_{k=1}^{len(x)} \log \frac{s(x,V,y)}{\sum_{i=1}^{L} S(x,V,i) + \sum_{j=1}^{U} S(x,Q,j)} \tag{4}$$

where L and U are the lengths of V and Q respectively, and y denotes the identity index of sample x in V.

All high-pass filtering operations are applied only in the second and third stages during training. Afterwards, the original tokens are passed into the encoder layers for further processing. The detailed structure is illustrated in Fig. 5(a).

3.3. Learnable lifting block

Existing transformer-based person search models often retain self-attention mechanisms (Cao et al., 2022; Yu et al., 2022), resulting in high computational complexity. When dealing with long inputs, the computational cost becomes prohibitively high. In contrast, lifting scheme has linear complexity, maintaining high efficiency regardless of input size. In addition, the work of Pranav Jeevan *et al.* also inspired us to explore multilevel decomposition and fusion (Jeevan & Sethi, 2022). Therefore, we replace the self-attention mechanism with a learnable lifting block. Besides computational efficiency, it also allows for multiscale feature extraction. Specifically, we do not stack multiple LLBs beyond the depth of the original encoder. Each multi-head self-attention layer is simply replaced by one LLB. This design ensures architectural consistency across stages and avoids introducing extra depth or parameters.

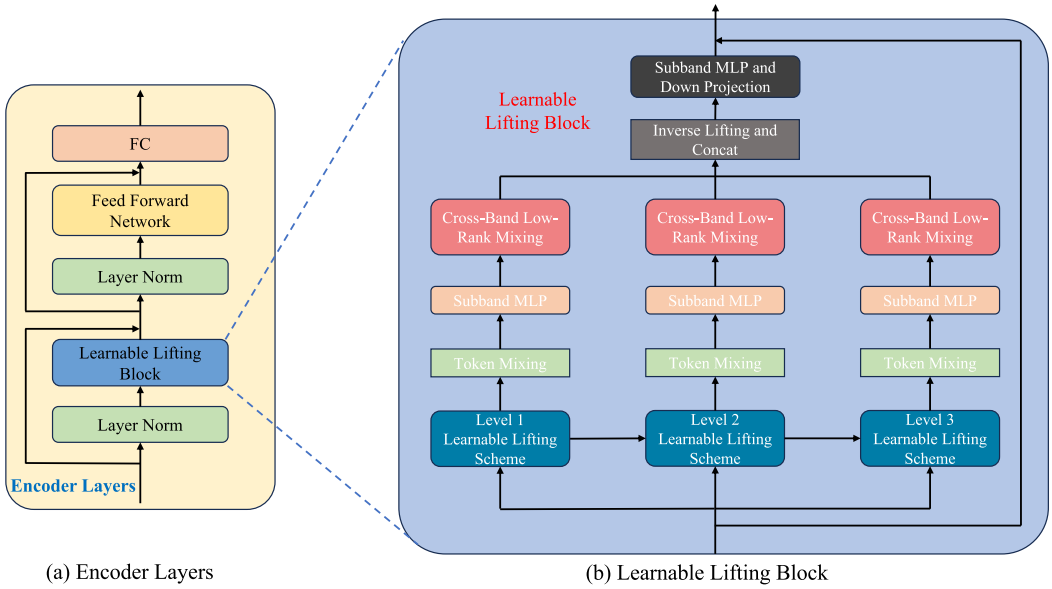
The forward process of the lifting scheme involves dividing the sequence into odd and even subsequences, denoted as o and e, based on their respective indices. The even sequence is used to predict the odd sequence, and the residual is denoted as d. The residual d is then used to update the even sequence, resulting in s, as expressed by the following equation:

$$\begin{cases} d = o - P(e) \\ s = e + U(d) \end{cases} \tag{5}$$

where P and U are operators. Similarly, the inverse process is as follows:

$$\begin{cases} e = s - U(d) \\ o = d + P(e) \end{cases} \tag{6}$$

Implementing P and U using small-scale neural networks constitute a learnable lifting scheme. In our learnable lifting scheme, the P and U operators are implemented with depthwise-separable convolutions and GEGLU gating. The 2D lifting transform is achieved by performing the lifting operation horizontally first and vertically afterward.



(a) Encoder Layers (b) Learnable Lifting Block

Fig. 5. (a) Architecure of the encoder layers. The encoder follows a standard transformer architecture, where Learnable Lifting Blocks replace the self-attention layers. Layer normalization, feed-forward modules, and residual connections remain unchanged. Architecture of the Learnable Lifting Block (LLB). The LLB replaces the multi-head self-attention layer with a multilevel lifting scheme, subband mixing, and cross-band low-rank fusion. This design reduces computational complexity while capturing multiscale features across different decomposition levels.

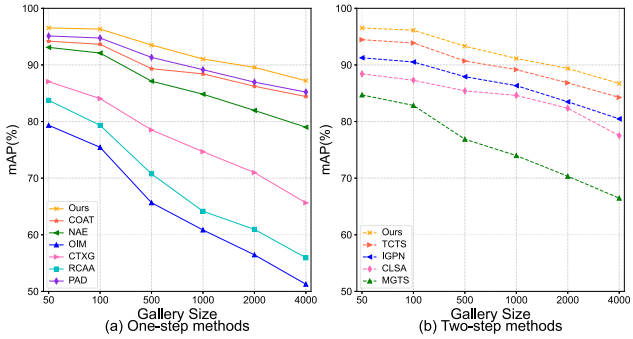


Fig. 6. Performance comparison with other methods on the CUHK-SYSU dataset under different gallery sizes. (a) Results of FLET compared with one-step person search methods. (b) Results of FLET compared with two-step methods.

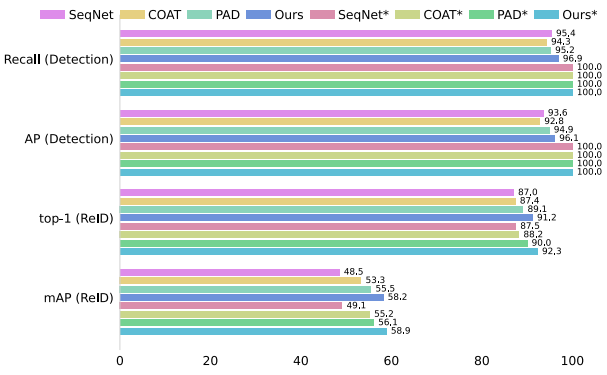


Fig. 7. Comparison of person search and detection scores on the PRW dataset with and without using ground-truth annotations. * indicates the use of ground-truth.

For a tensor $x \in \mathcal{R}^{h \times w \times c}$, as illustrated in Fig. 5(b), we first perform a Learnable Lifting Scheme to obtain four subbands:$x_{LL}, x_{LH}, x_{HL}, x_{HH} \in \mathcal{R}^{\frac{h}{2} \times \frac{w}{2} \times c}$,which is the level-1 decomposition. We then apply another Learnable Lifting Scheme to the low-frequency subband x_{LL} to obtain smaller subbands $y_{LL}, y_{LH}, y_{HL}, y_{HH} \in \mathcal{R}^{\frac{h}{4} \times \frac{w}{4} \times c}$ which forms the level-2 decomposition. This process continues until decomposition is no longer possible.

For the subbands obtained at each decomposition level, token mixing is first performed using a two-dimensional convolution. Each subband is then processed by an MLP module (i.e., Subband MLP). Subsequently, the subbands at each level are concatenated and subjected to cross-band low-rank mixing. Finally, all subbands are restored to their original size through an inverse lifting scheme and concatenated. The output is produced through Subband MLP, down-projection, and residual connections. It is subsequently forwarded to the following encoder layers, as illustrated in Fig. 5(b).

We now offer theoretical justification for the validity of replacing self-attention with the proposed Learnable Lifting Block. From the perspective of receptive fields, LLB can be viewed as a learnable realization of a 2D lifting scheme. Classical lifting-based wavelet transforms and multiresolution analysis show that repeated prediction-update steps progressively aggregate information over larger spatial supports. This process ultimately produces representations that encode the global structure of the signal (Sweldens, 1998). Recent studies have also shown that carefully designed global filters or large-kernel depthwise convolutions can serve as efficient alternatives to self-attention. These operators are capable of modeling long-range dependencies in vision tasks (Ding et al., 2022; Rao et al., 2021). Our LLB is based on the same underlying idea. Instead of computing pairwise dot-product attention, it aggregates global context through multi-scale lifting and repeated depthwise filtering. This yields a globally aware yet computationally efficient alternative to self-attention.

3.4. Loss function

During training, our model is supervised by a combination of four types of loss functions: detection loss $\mathcal{L}_{det}$, re-identification loss $\mathcal{L}_{OIM}$ and $\mathcal{L}_{ID}$, frequency-aware proxy loss $\mathcal{L}_{P}$, and reconstruction loss $\mathcal{L}_{rec}$.

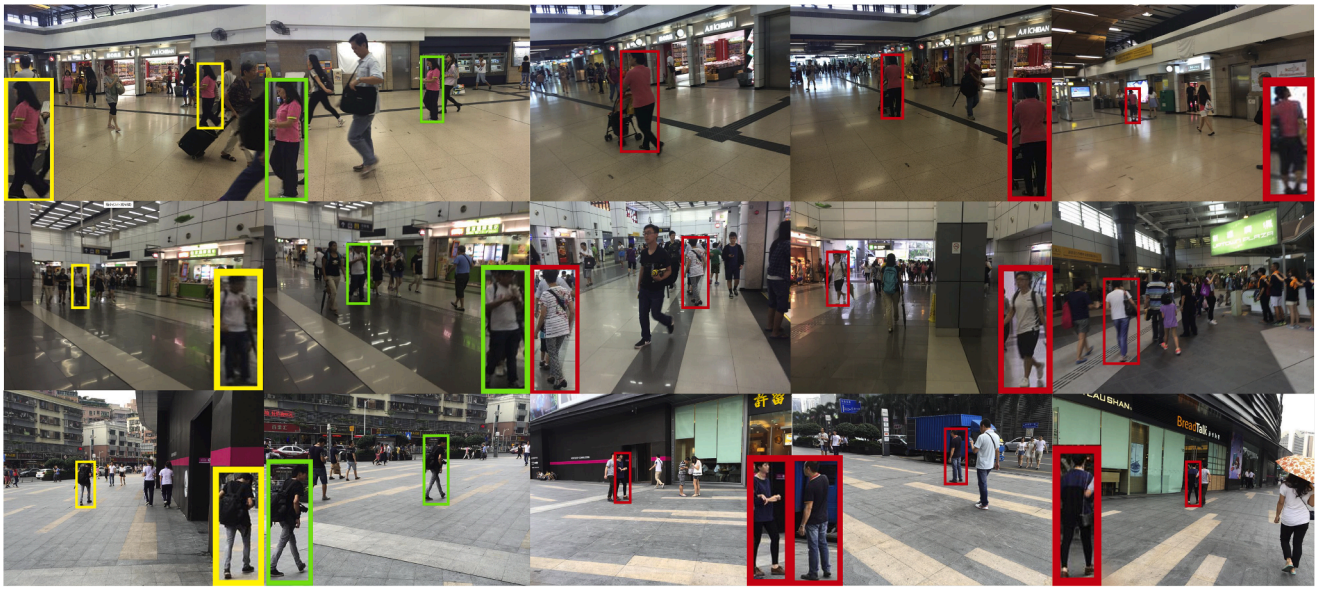


Fig. 8. Top-1 person search qualitative results of FLET, SeqNet, PSTR, and COAT on the CUHK-SYSU dataset. Yellow, green, and red boxes indicate the query, correct matches, and incorrect matches, respectively.

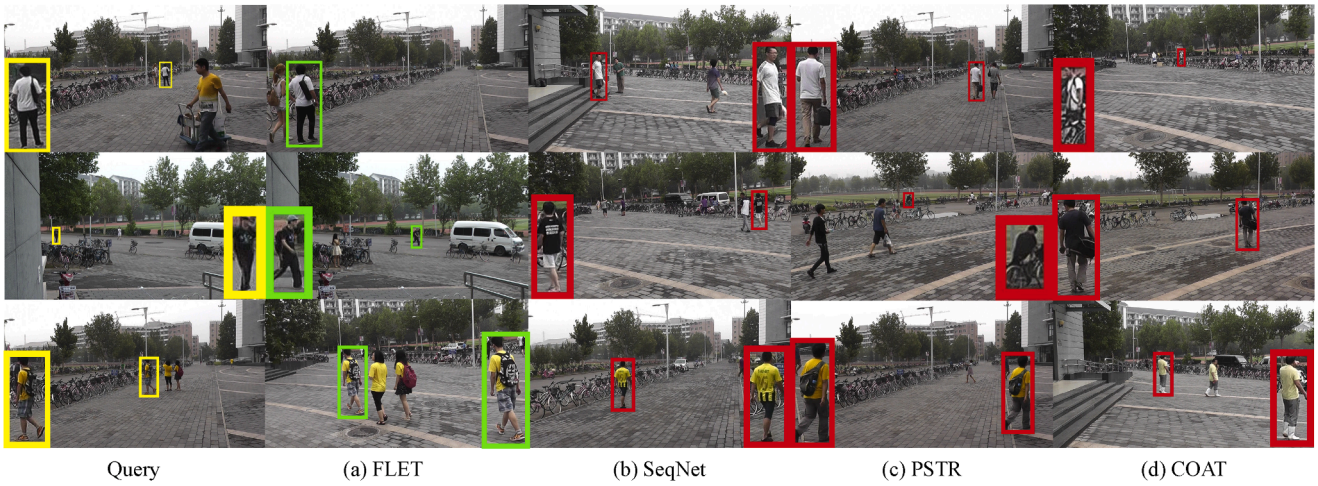


Fig. 9. Top-1 person search qualitative results of FLET, SeqNet, PSTR, and COAT on the PRW dataset. Yellow, green, and red boxes indicate the query, correct matches, and incorrect matches, respectively.

The detection loss $\mathcal{L}_{det}$ includes a cross-entropy loss $\mathcal{L}_{cls}$ for distinguishing persons from the background, and a Smooth-L1 loss $\mathcal{L}_{reg}$ for refining bounding box coordinates. Thus, the final detection loss is defined as:

$$\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{reg} \tag{7}$$

The OIM loss $\mathcal{L}_{OIM}$ maintains a lookup table (LUT) for labeled identities and a circular queue (CQ) for unlabeled ones It optimizes the embedding using cosine similarity. The identity loss $\mathcal{L}_{ID}$, also a cross-entropy loss, provides an additional supervision signal by predicting the identity class of persons.

As described above, the frequency-aware proxy loss $\mathcal{L}_P$ enhances the model's sensitivity to high-frequency features by leveraging enhanced inputs as auxiliary signals. It further computes a proxy-based loss on similarity to known or unknown identities.

$\mathcal{L}_{rec}$ is used to ensure that the lifting scheme is invertible, thereby preventing information loss. Whenever the forward process of the lifting scheme is applied to an input x, the inverse process may immediately follow with probability p. In such cases, it produces $\hat{x}$. Then we have:

$$\mathcal{L}_{rec} = \| x - \hat{x} \|_2^2 \tag{8}$$

Among the three losses, the detection loss and reconstruction loss are applied in all three stages, while the ReID and frequency-aware proxy losses are applied only in the second and third stages. The total loss function is defined as:

$$\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_{\text{det}}^t + \mathcal{L}_{rec}^t + \mathcal{I}(t > 1)(\lambda_{\text{OIM}}\mathcal{L}_{\text{OIM}}^t + \lambda_{\text{ID}}\mathcal{L}_{\text{ID}}^t + \lambda_{\text{P}}\mathcal{L}_{\text{P}}^t) \tag{9}$$

where λ_{OIM}, λ_{ID} and λ_P are weighting coefficients to balance the contributions of each loss.

For better understanding, we provide the complete training procedure of the FLET in Algorithm1.

4. Experiments

In this section, we first introduce the datasets and implementation settings. Then, we compare our FLET with various approaches. Ablation

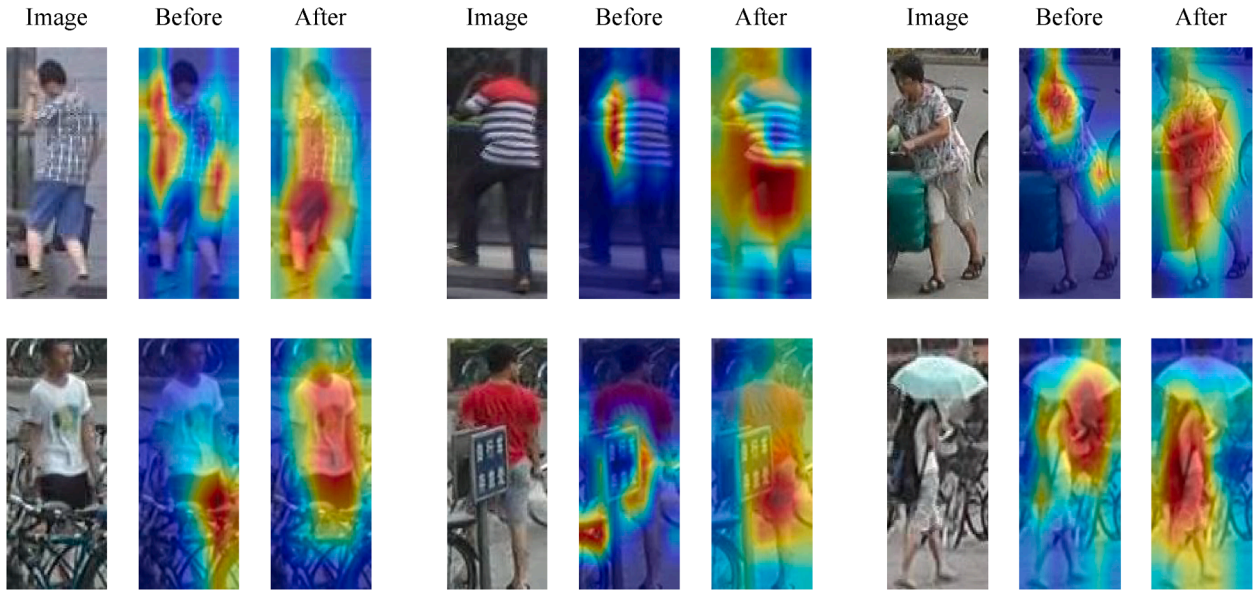


Fig. 10. Activation maps before and after applying the proposed high-pass filtering. Each group shows the original image, the baseline activation, and the enhanced activation.

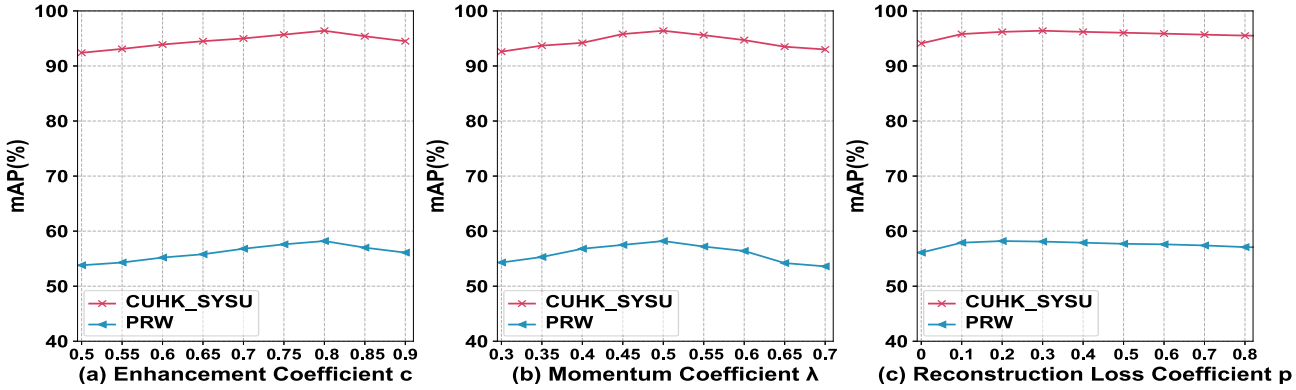


Fig. 11. Model performance under different hyperparameter values on CUHK-SYSU and PRW datasets. (a) Enhancement coefficient c in high-pass filtering. (b) Momentum coefficient λ in the frequency-aware proxy loss. (c) Reconstruction loss coefficient p.

studies are conducted to evaluate the contribution of each component. Finally, we present more performance analysis to further demonstrate the effectiveness of FLET.

4.1. Datasets and settings

4.1.1. Datasets

We conduct experiments on two widely used person search benchmark datasets: CUHK-SYSU and PRW.

CUHK-SYSU (Xiao et al., 2017) is a large-scale person search benchmark containing 18,184 scene images with 96,143 annotated bounding boxes covering 8432 identities. It includes images from both street photography and movie scenarios. The street images feature significant variations in viewpoint, lighting, resolution, and occlusion, while movie scenes offer even more diverse and challenging conditions.

PRW (Zheng et al., 2017) is a dataset collected with six synchronized cameras on a university campus. It focus on pedestrian detection and re-identification in outdoor scenarios. It includes 11,816 video frames, with 34,304 annotated bounding boxes covering 932 identities.

4.1.2. Implementation details

Our model is implemented using the PyTorch framework and trained on NVIDIA Tesla V100 GPUs. We use the SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. The model is trained for 20 epochs with a batch size of 3 and an initial learning rate of 0.003. A learning rate warm-up is applied during the first epoch, and the learning rate is decayed after the 15th epoch. During inference, Non-Maximum Suppression (NMS) is used to remove redundant bounding boxes, with NMS thresholds set to 0.4, 0.5, and 0.6 for the three stages, respectively.

4.2. Comparison with state-of-the-art

In this subsection, we evaluate FLET on both benchmark datasets and compare its performance with various methods.

4.2.1. Results on CUHK-SYSU

Table 1 presents the results on the CUHK-SYSU dataset. Our FLET achieves 96.4% mAP and 97.3% top-1 accuracy, outperforming most models. Our approach surpasses the baseline method COAT by +2.2% mAP and +2.6% top-1 accuracy. Compared with SOLIDER, which employs semantically controllable self-supervised learning, FLET also performs better.

These improvements demonstrate that our model has a stronger capability for perceiving and capturing high-frequency features.

To further validate robustness, we compare FLET with various approaches under different gallery sizes on the CUHK-SYSU test set. As

Algorithm 1 Training process of FLET.

Input: Training set S, total epochs E, iterations T
Output: Trained Model weight $\mathcal{W}$

```
Initialize the model weight W
for e = 1 to E do
  for each batch B sampled from I do
    Obtain person features F via Backbone and RoI-Align from B
    Calculate detection loss L_det and reconstruction loss L_rec
    for t = 1 to T do
      if t = 1 then
        Pass F through the Branch-Filtering Transformer and optimize
        Compute the detection loss L_det
      else
        Leverage bounding box regression generated in the prior stage
        Obtain augmented input by using high-pass filtering
        Generate token maps and perform exchanging
        Calculate frequency-aware proxy loss L_P by Eqs. (3) and (4)
        Pass the original tokens into encoder layers and compute detection loss L_det, identity loss L_ID and OIM loss L_OIM
      end if
      Calculate total loss L by Eq. (9)
      Back propagate to update W
    end for
  end for
end for
return trained model weight W
```

Table 1
Comparison with the state-of-the-art methods on CUHK-SYSU and PRW datasets. The bold entities denote the best performance.

Methods	Backbone	CUHKSYSU		PRW	
		mAP	top-1	mAP	top-1
Two-steps methods					
IDE (Zheng et al., 2017)	ResNet50	-	-	20.5	48.3
MGTS (Chen et al., 2018)	VGG16	83.0	83.7	32.6	72.1
CLSA (Lan et al., 2018)	ResNet50	87.2	88.5	38.7	65.0
RDLR (Han et al., 2019)	ResNet50	93.0	94.2	24.9	70.2
IGPN (Dong et al., 2020b)	ResNet50	90.3	91.4	47.2	87.0
TCTS (Wang et al., 2020)	ResNet50	93.9	95.1	46.8	87.5
One-step with CNNs					
OIM (Xiao et al., 2017)	ResNet50	75.5	78.7	21.3	49.4
RCAA (Chang et al., 2018)	ResNet50	79.3	81.3	-	-
CTXG (Yan et al., 2019)	ResNet50	84.1	86.5	33.4	73.6
NAE (Chen et al., 2020)	ResNet50	91.5	92.4	43.3	80.9
NAE+ (Chen et al., 2020)	ResNet50	92.1	92.9	44.0	81.1
AlignPS (Yan et al., 2021)	ResNet50-DCN	93.1	93.4	45.9	81.9
SeqNet (Li & Miao, 2021)	ResNet50	94.6	95.3	48.5	87.0
CANR+ (Zhao et al., 2022)	ResNet50	93.9	94.5	44.8	83.9
SPG (Song et al., 2023)	ResNet50	95.0	95.9	48.4	89.8
SeqNeXt+GFN (Jaffe & Zakhor, 2023)	ResNet50	95.2	95.6	50.9	91.2
DMRNet++ (Han et al., 2023)	ResNet50	94.5	95.7	52.1	87.0
SEAS (Jiang et al., 2024)	ResNet50	95.5	97.0	52.6	85.7
SeqNet-D+DDPS (Jia et al., 2025)	ResNet18	92.6	93.5	41.9	79.6
PS-DFSI (Zhang et al., 2025)	ResNet50	95.5	95.9	55.2	88.6
One-step with Transformers					
PSTR (Cao et al., 2022)	ResNet50	94.4	95.2	50.7	87.4
PSTR (Cao et al., 2022)	PVTv2-B2	95.2	96.2	56.5	89.7
SAT (Fiaz et al., 2023)	ResNet50	95.3	95.8	55.3	89.2
SOLIDER (Chen et al., 2023)	Swin-s	95.5	95.8	**59.8**	86.7
COAT (Yu et al., 2022)	ResNet50	94.2	94.7	53.3	87.4
ASTD (Zhang et al., 2024a)	ResNet50	95.8	96.2	55.7	90.2
PAD (Kim et al., 2025)	ResNet50	94.8	95.4	55.5	89.1
FLET (Ours)	ResNet50	**96.4**	**97.3**	58.2	**91.2**

Table 2
An ablation study on the PRW dataset to analyze the stage-wise impact of high-pass filtering and frequency-aware proxy loss.

Index	Stage 2	Stage 3	PRW	
			mAP	top-1
1			55.9	88.8
2	✓		57.3	90.1
3	✓	✓	**58.2**	**91.2**

shown in Fig. 6, performance of all methods drops as the gallery size increases due to the growing number of distractor persons. Nevertheless, FLET consistently outperforms all baselines, showing strong potential and robustness in large-scale search scenarios.

4.2.2. Results on PRW

Table 1 reports the performance on the PRW dataset. Compared with CUHK-SYSU, PRW is more challenging due to fewer training samples, a larger gallery, and many visually similar identities. Our FLET achieves 58.2% mAP and 91.2% top-1 accuracy. Compared to CNN-based methods such as AlignPS and SeqNet, our FLET outperforms them significantly. It also improves upon COAT by +4.9% mAP and 3.8% top-1 accuracy. In Fig. 7, we further compare FLET with SeqNet, COAT, and SAT with and without using ground-truth boxes on PRW. The results indicate that FLET not only achieves superior performance on person search but also benefits from ground-truth information. This demonstrates its strong ability to address the re-identification sub-task.

4.3. Ablation study

In this section, we conduct detailed ablation studies to evaluate the design choices of our model. We assess the effectiveness of each component and analyze the contribution of different stages. Additionally, we show that the proposed learnable lifting block effectively reduces inference time.

4.3.1. Analysis of the contribution of HPF and FP loss

To investigate the joint contribution of high-pass filtering and frequency-aware proxy loss, we analyze the effect of introducing them at different stages on the PRW dataset. As shown in Table 2, introducing them only at stage 2 leads to a 1.4% improvement in mAP and 1.3% in top-1 accuracy. Further introducing them at stage 3 yields an additional 0.9% mAP gain and 1.1% improvement in top-1. These results demonstrate the effectiveness of enhancing high-frequency components via HPF and FP Loss. The continued performance gains when applied in multiple stages further confirm the cascaded synergistic effect of them.

4.3.2. Analysis of the contribution of learnable lifting block

To assess the performance and efficiency of the LLB, we replace them with corresponding modules from various architectures, including CvT (Wu et al., 2021), Swin Transformer (Liu et al., 2021), PVT (Wang et al., 2021), MLP-Mixer (Tolstikhin et al., 2021), and GFNet(Rao et al., 2021). We evaluate these variants on the PRW dataset. The results show that our proposed learnable lifting block consistently outperforms all these alternatives in both accuracy and inference speed. This demonstrates its superior effectiveness. The detailed performance of different self-attention alternatives on the PRW dataset is presented in Table 3.

4.3.3. Analysis of the effectiveness of learnable lifting scheme

The Learnable Lifting Scheme (LLS) is not the only approach capable of performing multilevel decomposition, subband processing, and reconstruction. The Haar wavelet transform (Haar WT) (Mallat, 2008) and the Cohen-Daubechies-Feauveau Biorthogonal Wavelet Transform (Daubechies & Feauveau, 1992) can achieve similar functionality. For the latter, the CDF 5/3 and CDF 9/7 variants are the most widely

Table 3
Comparison of inference speed and accuracy on the PRW dataset among different self-attention alternatives.

Layer	Time (ms)	PRW	
		mAP	top-1
ViT(Vaswani et al., 2017)	89	55.7	87.6
CvT(Wu et al., 2021)	78	56.2	87.4
Swin Transformer (Liu et al., 2021)	75	57.0	89.7
PVT (Wang et al., 2021)	83	55.8	88.9
MLP-Mixer (Tolstikhin et al., 2021)	62	56.1	87.9
GFNet (Rao et al., 2021)	55	56.6	88.2
LLB (Ours)	**35**	**58.2**	**91.2**

Table 4
Comparison of inference speed and accuracy on the PRW dataset among different LLS alternatives.

Metthods	Parameterization	mAP	top-1	Time (ms)
Haar WT (Mallat, 2008)	fixed	55.4	86.3	31
CDF 5/3 (Daubechies & Feauveau, 1992)	fixed	55.9	87.0	33
CDF 9/7 (Daubechies & Feauveau, 1992)	fixed	55.7	86.9	33
LLS (Ours)	learnable	**58.2**	**91.2**	**35**

Table 5
Comparison of different learnable operators, with kernel size and gating function set to their optimal choices on a small-scale validation set.

Method	Kernel Size	Gating function	mAP	top-1	Time (ms)
Conv	3×3	-	56.3	89.2	31
Conv-Gated	3×3	Sigmoid	56.5	89.5	32
DWConv	3×3	-	56.9	90.2	33
DWConv-Gated	3×3	SiLU	57.3	91.0	35
SepConv	3×3	-	57.6	90.7	34
SepConv-Gated(Ours)	3×3	GEGLU	**58.2**	**91.2**	**35**

adopted. To evaluate whether LLS provides advantages in terms of performance and computational efficiency compared with these methods, we conducted experiments on the PRW dataset. The evaluation metrics include mAP, top-1 accuracy, and inference speed. For fairness, all inputs were resized to 900 × 1500, and inference were carried out on a Tesla V100 GPU. As shown in Table 4, LLS consistently outperforms the alternatives across all metrics. We attribute this to the fact that fixed decomposition operations are unable to adapt to diverse data distributions.

4.3.4. Comparison of learnable operators

In the previous section, we have demonstrated the advantages of learnable operators over fixed ones. In this part, we further investigate which type of learnable operator yields the best performance. Specifically, we compare convolution, depthwise convolution, depthwise separable convolution, and their gated variants. We refer to these modules as Conv, DWConv, SepConv, Conv-Gated, DWConv-Gated, and SepConv-Gated. The experiments are conducted on the PRW dataset, with all inputs resized to 900 × 1500, and inference performed on a Tesla V100 GPU. The kernel size and gating function were selected based on their optimal performance on a small-scale validation set. As shown in Table 5, SepConv-Gated achieves the best results across all metrics. We attribute the performance to the parameter efficiency and computational advantages of depthwise separable convolution (Park et al., 2024), as well as the training stability brought by GEGLU (Shazeer, 2020).

4.3.5. Analysis of parameters sharing

In this section, we investigate the issue of parameter sharing in the operators of the learnable lifting scheme. The operators P and U are involved in two types of computations. First, they operate between the forward and inverse processes, i.e., during decomposition and reconstruction. Second, they are applied along both the horizontal and vertical directions, which is necessary to extend the one-dimensional lifting scheme to two dimensions. Experiments are conducted on the PRW dataset, where all images are resized to 900 × 1500 and inference is performed on a Tesla V100 GPU. Parameter sharing between the forward and inverse processes and between the horizontal and vertical directions are denoted as Share on FI and Share on HV, respectively. As shown in Table 6, sharing parameters in both cases yields the best overall performance, while computational efficiency remains nearly identical. We hypothesize that this is due to the inherent reversibility of the learnable lifting scheme, which allows it to function effectively without requiring multiple sets of parameters. Adding redundant parameters could interfere with this reversibility.

Table 6
Ablation on parameter sharing in the learnable lifting scheme: Share on FI–shared between forward/inverse; share on HV–shared between horizontal/vertical.

Index	Share on FI	Share on HV	mAP	top-1	Time(ms)
0			57.4	90.7	**35**
1	✓		57.9	91.1	**35**
2		✓	57.7	90.8	**35**
3	✓	✓	**58.2**	**91.2**	**35**

Table 7
Performance comparison of TransReID (He et al., 2021) and ResT-ReID (Chen et al., 2022) with and without the proposed LLB on the Market-1501, and DukeMTMC-reID datasets.

Methods	Market-1501		DukeMTMC	
	mAP	Rank-1	mAP	Rank-1
TransReID	88.9	95.2	82.0	90.7
+LLB	89.5	95.5	82.5	91.1
ResT-ReID	95.3	88.2	90.0	80.6
+LLB	95.6	88.6	90.6	81.1

4.4. Generalization to transformer-based ReIDid frameworks

To further evaluate the generality of the proposed Learnable Lifting Block (LLB), we integrate it into two representative transformer-based ReID backbones: TransReID (He et al., 2021) and ResT-ReID (Chen et al., 2022). In TransReID, LLB replaces the multi-head self-attention (MHSA) in the transformer layer. In ResT-ReID, LLB substitutes its PW-MSA module. All other components remain unchanged to ensure a fair comparison.

We evaluate each backbone and its LLB-enhanced variant on Market-1501, and DukeMTMC-reID datasets, as shown in Table 7. Across all datasets and both architectures, introducing LLB consistently improves mAP and Rank-1 accuracy. These results confirm that LLB is not tied to the COAT-based person search framework and can effectively replace different transformer-style attention modules in standalone ReID models.

4.5. Qualitative analysis

We conduct qualitative analysis on the PRW dataset and compare FLET with SeqNet, PSTR, and COAT. As shown in Figs. 8 and 9, other methods are prone to misidentifications when persons have similar colors or are too small. This indicates insufficient modeling of high-frequency details, which limits their ability to extract discriminative features. In contrast, our model FLET, trained with high-frequency enhanced features, can more accurately detect subtle cues on persons and avoid such errors.

Table 8
Comparison of parameter count, FLOPs and inference time for person search methodss on the PRW dataset.

Methods	Params(M)	FLOPs(G)	Time(ms)	mAP
NAE+ (Chen et al., 2020)	33	575	95	44
SeqNet (Li & Miao, 2021)	48	550	88	48.5
AlignPS (Yan et al., 2021)	42	380	60	45.9
PSTR (Cao et al., 2022)	43	356	56	50.7
COAT (Yu et al., 2022)	37	473	89	53.3
PSDiff (Zhang et al., 2025)	31	390	72	53.5
ASTD (Zhang et al., 2024a)	43	348	96	55.7
FLET(Ours)	**28**	**316**	**35**	**58.2**

4.6. Visual analysis of high-frequency enhancement

To better illustrate how the proposed HPF branch and frequency-aware proxy loss improve feature representation, we visualize activation maps before and after enhancement, as shown in Fig. 10. Each group contains the original image, the baseline activation map, and the enhanced activation map obtained with our method.

The baseline activations mainly respond to coarse body regions and often focus on large smooth areas. In contrast, the enhanced activation maps highlight clearer edges, textures, and fine-grained patterns, especially along clothing contours, accessory details, and high-frequency boundaries. These changes indicate that our method strengthens the model's ability to perceive discriminative high-frequency cues that are easily overlooked by standard transformer features.

This visualization confirms that high-frequency enhancement contributes to more distinctive feature learning, thereby improving robustness in challenging person search scenarios.

4.7. Hyperparameter analysis

Enhancement Coefficient c: The coefficient c is used to generate augmented data in high-pass filtering. A larger c results in a higher proportion of high-frequency components in the augmented data, whereas a smaller c leads to a lower proportion. As shown in Fig. 11(a), the method achieves its best performance when $c = 0.8$.

Momentum Coefficient λ: The coefficient λ is used for updating the Table V during the computation of FP Loss. It balances the importance between new and old information. Smaller λ values make the model more sensitive and adaptive, while larger values lead to more stable and noise-resistant behavior. As shown in Fig. 11(b), $\lambda = 0.5$ yield optimal results.

Reconstruction Loss Coefficient p: The coefficient p denotes the proportion of the input that is used to compute the loss $\mathcal{L}_{rec}$. As illustrated in Fig. 11(c), the model achieves its best performance when $p = 0.2$ or $p = 0.3$. As p increases further, performance gradually declines, while setting $p = 0$ leads to a sharp drop. We hypothesize that when $p > 0.3$, the constraint imposed by $\mathcal{L}_{rec}$ compromises the reversibility of the learnable lifting scheme. In contrast, when $p = 0$, the absence of such a constraint reduces training stability.

4.8. Efficiency comparison

To demonstrate the efficiency of FLET, we report the inference time (in milliseconds) on a Tesla V100 GPU. For fair comparison, all input images are resized to 900 × 1500. As shown in Table 8, Our method achieves the best performance in terms of parameter count, FLOPs, inference time, and mAP. This improvement is mainly attributed to replacing the self-attention mechanism with the proposed learnable lifting block.

5. Conclusion

We propose a transformer-based end-to-end person search model. The model improves the perception and extraction of high-frequency and multiscale information and simultaneously reduces computational cost and increases inference speed. Specifically, we introduce a frequency-aware proxy loss, enabling the model to learn stronger sensitivity to high-frequency components from specially augmented inputs. Furthermore, we replace the self-attention layers in the transformer encoder with learnable lifting block. As a result, the model retains a global receptive field, requires substantially less computation, and achieves better multiscale feature extraction. Extensive experiments on CUHK-SYSU and PRW datasets demonstrate that our FLET achieves state-of-the-art performance and exhibits high computational efficiency. These findings highlight its effectiveness and suitability for real-world person search applications.

Limitations and Future Work: Although FLET achieves state-of-the-art performance with high computational efficiency, there remains considerable room for improvement. (1) This work does not extensively examine the rationale of components within the learnable lifting block beyond the learnable lifting scheme itself. Potential enhancements may lie in more efficient fusion of subbands information or in eliminating unnecessary steps to further accelerate inference. (2) While we have discussed issues related to scale, texture, and edges, aspects such as pose and viewpoint remain unexplored. Future work may incorporate pose-guided alignment modules or view-invariant feature learning to address these limitations.

CRediT authorship contribution statement

Qilin Shu: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing; **Qixian Zhang:** Conceptualization, Investigation, Writing – review & editing; **Duoqian Miao:** Supervision, Funding acquisition; **Qi Zhang:** Supervision, Funding acquisition; **Hongyun Zhang:** Supervision, Funding acquisition; **Cairong Zhao:** Supervision, Funding acquisition.

Data availability

The code is still in an experimental phase and lacks the necessary optimization and stability for public release. It has not been fully tested or documented, and is tailored to internal workflows.

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.

Acknowledgement

The work is supported by the National Natural Science Foundation of China (Grant Nos. 62576251, 62376198, 62576247), the National Key Research and Development Program (Grant No. 2022YFB3104700).

References

Cai, Z., & Vasconcelos, N. (2018). Cascade R-CNN: Delving into high quality object detection. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit. (cvpr)* (pp. 6154–6162). https://openaccess.thecvf.com/content_cvpr_2018/{html}/Cai_Cascade_R-CNN_Delving_CVPR_2018_paper.{html}.

Cao, J., Pang, Y., Anwer, R. M., Cholakkal, H., Xie, J., Shah, M., & Khan, F. S. (2022). PSTR: End-to-end one-step person search with transformers. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit. (CVPR)* (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}.

Chang, X., Huang, P.-Y., Liang, X., Yang, Y., & Hauptmann, A. G. (2018). RCAA: Relational context-aware agents for person search. In *Proc. eur. conf. comput. vis. (ECCV)* (pp. 84–100). https://openaccess.thecvf.com/content_ECCV_2018/{html}/Xiaojun_Chang_RCAA_Relational_Context-Aware_ECCV_2018_paper.{html}.

Chen, D., Zhang, S., Ouyang, W., Yang, J., & Tai, Y. (2018). Person search via a mask-guided two-stream cnn model. In *Proc. eur. conf. comput. vis. (ECCV)* (pp. 734–750). https://openaccess.thecvf.com/content_ECCV_2018/papers/Di_Chen_Person_Search_via_ECCV_2018_paper.pdf.

Chen, D., Zhang, S., Yang, J., & Schiele, B. (2020). Norm-aware embedding for efficient person search. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit. (CVPR)* (pp. 12615–12624). https://openaccess.thecvf.com/content_CVPR_2020/{html}/Chen_Norm-Aware_Embedding_for_Efficient_Person_Search_CVPR_2020_paper.{html}.

Chen, W., Xu, X., Jia, J., Luo, H., Wang, Y., & Sun, X. (2023). Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit. (CVPR)* (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}.

Chen, Y., Li, Z., & Song, A. (2024). Multi-query person search with transformers. In *Advances in knowledge discovery and data mining: 28th pacific-asia conference on knowledge discovery and data mining, PAKDD 2024, taipei, taiwan, may 7-10, 2024, proceedings, part IV* (p. 116-128). Berlin, Heidelberg: Springer-Verlag. https://dl.acm.org/doi/abs/10.1007/978-981-97-2238-9_9.

Chen, Y., Xia, S., Zhao, J., Zhou, Y., Niu, Q., Yao, R., Zhu, D., & Liu, D. (2022). RestreID: Transformer block-based residual learning for person re-identification. *Pattern Recognition Letters*, *157*, 90–96. https://www.sciencedirect.com/science/article/pii/S016786552200085X.

Dai, Z., Liu, H., Le, Q. V., & Tan, M. (2021). Coatnet: Marrying convolution and attention for all data sizes. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems (neurIPS)* (pp. 3965–3977). Curran Associates, Inc. (*vol. 34*). https://proceedings.neurips.cc/paper/2021/hash/20568692db622456cc42a2e853ca21f8-Abstract.html.

D'Ascoli, S., Touvron, H., Leavitt, M. L., Morcos, A. S., Biroli, G., & Sagun, L. (2021). Convit: Improving vision transformers with soft convolutional inductive biases. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning (ICML 2021)* (pp. 2286–2296). PMLR (*vol. 139*). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v139/d-ascoli21a.html.

Daubechies, I., & Feauveau, J.-C. (1992). Biorthogonal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, *45*(5), 485–560. https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160450502.

Ding, X., Zhang, X., Han, J., & Ding, G. (2022). Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 11963–11975). https://openaccess.thecvf.com/content/CVPR2022/papers/Ding_Scaling_Up_Your_Kernels_to_31x31_Revisiting_Large_Kernel_Design_CVPR_2022_paper.pdf.

Dong, W., Zhang, Z., Song, C., & Tan, T. (2020a). Bi-directional interaction network for person search. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit. (CVPR)* (pp. 2839–2848). https://openaccess.thecvf.com/content_CVPR_2020/{html}/Dong_Bi-Directional_Interaction_Network_for_Person_Search_CVPR_2020_paper.{html}.

Dong, W., Zhang, Z., Song, C., & Tan, T. (2020b). Instance guided proposal network for person search. In *Proc. IEEE/CVF conf. comput vis. pattern recognit. (CVPR)* (pp. 2585–2594). https://openaccess.thecvf.com/content_CVPR_2020/{html}/Dong_Instance_Guided_Proposal_Network_for_Person_Search_CVPR_2020_paper.{html}.

Dou, S., Zhao, C., Jiang, X., Zhang, S., Zheng, W.-S., & Zuo, W. (2023). Human co-parsing guided alignment for occluded person re-identification. *IEEE Transactions on Image Processing*, *32*, 458–470. https://doi.org/10.1109/TIP.2022.3229639.

Fiaz, M., Cholakkal, H., Anwer, R. M., & Khan, F. S. (2023). SAT: Scale-augmented transformer for person search. In *Proc. IEEE winter conf. appl. comput. vis. (WACV)* (pp. 4820–4829). https://openaccess.thecvf.com/content/WACV2023/{html}/Fiaz_SAT_Scale-Augmented_Transformer_for_Person_Search_WACV_2023_paper.{html}.

Han, C., Ye, J., Zhong, Y., Tan, X., Zhang, C., Gao, C., & Sang, N. (2019). Re-ID driven localization refinement for person search. In *Proc. IEEE/CVF int. conf. comput. vis. (ICCV)* (pp. 9814–9823). https://openaccess.thecvf.com/content_ICCV_2019/papers/Han_Re-ID_Driven_Localization_Refinement_for_Person_Search_ICCV_2019_paper.pdf.

Han, C., Zheng, Z., Su, K., Yu, D., Yuan, Z., Gao, C., Sang, N., & Yang, Y. (2023). DMRNet++: Learning discriminative features with decoupled networks and enriched pairs for one-step person search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(6), 7319–7337. https://doi.org/10.1109/TPAMI.2022.3221079.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proc. IEEE conf. comput. vis. pattern recognit. (CVPR)* (pp. 770–778). https://openaccess.thecvf.com/content_cvpr_2016/{html}/He_Deep_Residual_Learning_CVPR_2016_paper.{html}.

He, S., Luo, H., Wang, P., Wang, F., Li, H., & Jiang, W. (2021). Transreid: Transformer-based object re-identification. In *Proc. IEEE/CVF int. conf. comput. vis. (ICCV)* (pp. 15013–15022). https://openaccess.thecvf.com/content/ICCV2021/papers/He_TransReID_Transformer-Based_Object_Re-Identification_ICCV_2021_paper.pdf.

Hu, B., Wang, X., & Liu, W. (2024). Personvit: Large-scale self-supervised vision transformer for person re-identification. https://arxiv.org/abs/2408.05398v2.

Jaffe, L., & Zakhor, A. (2023). Gallery filter network for person search. In *Proc. IEEE winter conf. appl. comput. vis.* (pp. 1684–1693). https://openaccess.thecvf.com/content/WACV2023/papers/Jaffe_Gallery_Filter_Network_for_Person_Search_WACV_2023_paper.pdf.

Jeevan, P., & Sethi, A. (2022). Wavemix: Resource-efficient token mixing for images. arXiv:2203.03689.

Ji, Z., Cheng, D., & Feng, K. (2025). Exploring stronger transformer representation learning for occluded person re-identification. *Multimedia System*, *31*(5). https://dl.acm.org/doi/abs/10.1007/s00530-025-01986-0.

Jia, Y., Quan, R., Chen, H., Liu, J., Yan, Y., Bai, S., & Qin, J. (2025). Disaggregation distillation for person search. *IEEE Transactions on Multimedia*, *27*, 158–170. https://ieeexplore.ieee.org/abstract/document/10817642.

Jiang, Y., Wang, H., Peng, J., Fu, X., & Wang, Y. (2024). Scene-adaptive person search via bilateral modulations. In *Proc. int. joint conf. artif. intell. (IJCAI)*. https://arxiv.org/pdf/2405.02834.

Kim, H., Lee, J., & Sohn, K. (2025). Prototype-guided attention distillation for discriminative person search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(1), 99–115. https://doi.org/10.1109/TPAMI.2024.3461778.

Lan, X., Zhu, X., & Gong, S. (2018). Person search by multi-scale matching. In *Proc. eur. conf. comput. vis. (ECCV)* (pp. 536–552). https://openaccess.thecvf.com/content_ECCV_2018/{html}/Xu_Lan_Person_Search_by_ECCV_2018_paper.{html}.

Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., & Wu, F. (2021). Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 2898–2907). https://openaccess.thecvf.com/content/CVPR2021/{html}/Li_Diverse_Part_Discovery_Occluded_Person_Re-Identification_With_Part-Aware_Transformer_CVPR_2021_paper.{html}.

Li, Z., & Miao, D. (2021). Sequential end-to-end network for efficient person search. In *Proc. AAAI conf. artif. intell. (AAAI)* (pp. 2011–2019). (*vol. 35*). https://doi.org/10.1609/aaai.v35i3.16297.

Liu, G., Xu, K., Zhu, J., Ge, Y., & Chen, X. (2025). A local-global transformer-based model for person re-identification. *PloS One*, *20*(11), e0335848. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0335848.

Liu, Z., Lin, Y., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF int. conf. comput. vis. (ICCV)* (pp. 10012–10022). https://openaccess.thecvf.com/content/ICCV2021/{html}/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.

Luo, H., Jiang, W., Fan, X., & Zhang, C. (2020). Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. *IEEE Transactions on Multimedia*, *22*(11), 2905–2913. https://ieeexplore.ieee.org/abstract/document/8955948.

Mallat, S. (2008). A wavelet tour of signal processing: The sparse way. https://dl.acm.org/doi/abs/10.5555/1525499.

Munjal, B., Amin, S., Tombari, F., & Galasso, F. (2019). Query-guided end-to-end person search. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit. (CVPR)* (pp. 811–820). https://openaccess.thecvf.com/content_CVPR_2019/{html}/Munjal_Query-Guided_End-To-End_Person_Search_CVPR_2019_paper.{html}.

Park, C., Park, M., Moon, H., Yoon, M. K., Go, S., Kim, S., & Ro, W. W. (2024). Deprune: Depth-wise separable convolution pruning for maximizing gpu parallelism. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in neural information processing systems (neurIPS)* (pp. 106906–106923). Curran Associates, Inc. (*vol. 37*). https://proceedings.neurips.cc/paper_files/paper/2024/file/c16a99558b0b4f6b10966ca9bdb98ade-Paper-Conference.pdf.

Rao, Y., Zhao, W., Zhu, Z., Lu, J., & Zhou, J. (2021). Global filter networks for image classification. In *Proc. adv. neural inf. process. syst. (neurIPS)* (pp. 980–993). (*vol. 34*). https://proceedings.neurips.cc/paper/2021/hash/07e87c2f4fc7f7c96116d8e2a92790f5-Abstract.html.

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.

Shazeer, N. (2020). Glu variants improve transformer. arXiv:2002.05202.

Song, Z., Zhao, C., Hu, G., & Miao, D. (2023). Learning scene-pedestrian graph for end-to-end person search. *IEEE Transactions on Industrial Informatics*, *20*(2), 2979–2990. https://doi.org/10.1109/TII.2023.3298473.

Spravil, J., Houben, S., & Behnke, S. (2024). Hyenapixel: Global image context with convolutions. In *Proceedings of the 27th European conference on artificial intelligence (ECAI 2024)* (pp. 521 – 528). https://arxiv.org/pdf/2402.19305.

Sweldens, W. (1998). The lifting scheme: A construction of second generation wavelets. *SIAM Journal on Mathematical Analysis*, *29*(2), 511–546. https://cm-bell-labs.github.io/who/wim/papers/lift2.pdf.

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J. et al. (2021). Mlp-mixer: An all-mlp architecture for vision. *Advances Neural Information Process System*, *34*, 24261–24272. https://proceedings.neurips.cc/paper/2021/hash/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Abstract.{html}.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Process Advances Neural Information Process System (NeurIPS)*, *30*, 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.{html}.

Wang, C., Ma, B., & Chen, X. (2020). TCTS: A task-consistent two-stage framework for person search. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit. (CVPR)* (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.

Wang, H., Shen, J., Liu, Y., Gao, Y., & Gavves, E. (2022a). Nformer: Robust person re-identification with neighbor transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 7297–7307). https://openaccess.thecvf.com/content/CVPR2022/{html}/Wang_NFormer_Robust_Person_Re-Identification_With_Neighbor_Transformer_CVPR_2022_paper.{html}.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 568–578). https://openaccess.thecvf.com/content/ICCV2021/{html}/Wang_Pyramid_Vision_Transformer_A_Versatile_Backbone_for_Dense_Prediction_Without_ICCV_2021_paper.{html}.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2022b). PVT v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, *8*(3), 415–424. https://link.springer.com/article/10.1007/s41095-022-0274-8.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). CVT: Introducing convolutions to vision transformers. In *Ieee/cvf int. conf. comput. vis. (iccv)* (pp. 22–31). https://openaccess.thecvf.com/content/ICCV2021/{html}/Wu_CvT_Introducing_Convolutions_to_Vision_Transformers_ICCV_2021_paper.{html}.

Xiao, T., Li, S., Wang, B., Lin, L., & Wang, X. (2017). Joint detection and identification feature learning for person search. In *Proc. IEEE conf. comput. vis. pattern recognit. (CVPR)* (pp. 3415–3424). https://openaccess.thecvf.com/content_cvpr_2017/papers/Xiao_Joint_Detection_and_CVPR_2017_paper.pdf.

Yan, Y., Li, J., Qin, J., Bai, S., Liao, S., Liu, L., Zhu, F., & Shao, L. (2021). Anchor-free person search. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit. (CVPR)* (pp. 7690–7699). https://openaccess.thecvf.com/content/CVPR2021/{html}/Yan_Anchor-Free_Person_Search_CVPR_2021_paper.{html}?ref=https://githubhelp.com.

Yan, Y., Zhang, Q., Ni, B., Zhang, W., Xu, M., & Yang, X. (2019). Learning context graph for person search. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit. (CVPR)* (pp. 2158–2167). https://openaccess.thecvf.com/content_CVPR_2019/{html}/Yan_Learning_Context_Graph_for_Person_Search_CVPR_2019_paper.{html}.

Ye, M., Chen, S., Li, C., Zheng, W.-S., Crandall, D., & Du, B. (2024). Transformer for object re-identification: A survey. *International Journal of Computer Vision*, *133*(5), 2410-2440. https://dl.acm.org/doi/abs/10.1007/s11263-024-02284-4.

Yu, R., Du, D., Davila, D., Funk, C., Hoogs, A., & Clipp, B. (2022). Cascade transformers for end-to-end person search. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit. (CVPR)* (pp. 7267–7276). https://openaccess.thecvf.com/content/CVPR2022/{html}/Yu_Cascade_Transformers_for_End-to-End_Person_Search_CVPR_2022_paper.{html}.

Zhang, G., Zhang, Y., Zhang, T., Li, B., & Pu, S. (2023). PHA: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit. (CVPR)* (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}.

Zhang, Q., Miao, D., Zhang, Q., Wang, C., Li, Y., Zhang, H., & Zhao, C. (2024a). Learning adaptive shift and task decoupling for discriminative one-step person search. *Knowledge-Based Systems*, *304*, 112483. https://doi.org/10.1016/j.knosys.2024.112483.

Zhang, Q., Miao, D., Zhang, Q., Zhao, C., Zhang, H., Sun, Y., & Wang, R. (2025). Dynamic frequency selection and spatial interaction fusion for robust person search. *Information Fusion*, *124*, 103314. https://doi.org/10.1016/j.inffus.2025.103314.

Zhang, Q., Wu, J., Miao, D., Zhao, C., & Zhang, Q. (2024b). Attentive multi-granularity perception network for person search. *Information Science*, *681*, 121191. https://doi.org/10.1016/j.ins.2024.121191.

Zhao, C., Chen, Z., Dou, S., Qu, Z., Yao, J., Wu, J., & Miao, D. (2022). Context-aware feature learning for noise robust person search. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(10), 7047–7060. https://doi.org/10.1109/TCSVT.2022.3179441.

Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., & Tian, Q. (2017). Person re-identification in the wild. In *Proc. IEEE conf. comput. vis. pattern recognit. (CVPR)* (pp. 1367–1376). https://openaccess.thecvf.com/content_cvpr_2017/papers/Zheng_Person_Re-Identification_in_CVPR_2017_paper.pdf.