Information Sciences 681 (2024) 121191

Contents lists available at ScienceDirect

Information Sciences

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

Attentive multi-granularity perception network for person search

Qixian Zhang [a], Jun Wu [b], Duoqian Miao [a,*], Cairong Zhao [a], Qi Zhang [a]

[a] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
[b] School of Computer Science, Fudan University, Shanghai 200438, China

ARTICLE INFO

Keywords:
Person search
Person re-identification
Multi-granularity
Feature mixer

ABSTRACT

Person search is an extremely challenging task that seeks to identify individuals through joint person detection and person re-identification from uncropped real scene images. Previous studies primarily focus on learning rich features to enhance identification. However, arbitrary feature enhancement strategies may introduce unwanted background noise. Moreover, different scenarios usually exhibit varying pedestrian appearances or even intricate occlusions, leading to inconsistent/incomplete pedestrian features in different images. In this paper, we introduce a novel Attentive Multi-granularity Perception (AMP) module seamlessly integrated into our AMPN network. This module specifically addresses appearance variations and occlusions within a person's Region of Interest (RoI). The AMP module harnesses discriminative relationship features from various local regions, significantly enhancing identification accuracy. It comprises two principal components: the Pedestrian Perception Enhancement (PPE) block and the Background Interference Suppressor (BIS). The PPE block introduces a Spatial-wise Feature Mixer and a Channel-wise Feature Mixer, which effectively capture and refine discriminative relation features. Simultaneously, the BIS operates in parallel with the PPE block, enriching the discriminative relation features and enhancing the distinctiveness between the foreground and background. Our AMP module is plug-and-play and can integrate with other person search models. Extensive experiments validate our model's merits, achieving state-of-the-art performance on CUHK-SYSU and a 4.8% mAP gain over SeqNet on PRW at a desirable speed. Our code is accessible at https://github.com/zqx951102/AMPN.

1. Introduction

Person search [1–4] is a promising yet extremely challenging task that involves fine-grained recognition and retrieval to locate and identify specific pedestrians from uncropped real-world scene images. It primarily consists of two subtasks: 1) *person detection* [5], which entails localizing bounding boxes around all pedestrians in the scene images, and 2) *person re-identification (ReID)* [6,7], which involves matching the cropped gallery person images with the query person images obtained from detection. Person search is closely aligned with real-world scenarios, thus facing challenges posed by various real-world factors, including occlusions or background clutters, pose/viewpoint variations, and scale variations.

Previous efforts in the field of person search can be broadly classified into two types: *two-step methods* [3,8–12] and *end-to-end methods* [4,13–15]. Two-step approaches train the detection and ReID networks independently and typically in a sequential manner. Initially, pedestrians are located using an off-the-shelf detector, and then their cropped images are fed into a ReID network for

* Corresponding author.
E-mail address: dqmiao@tongji.edu.cn (D. Miao).

https://doi.org/10.1016/j.ins.2024.121191
Received 22 October 2023; Received in revised form 28 June 2024; Accepted 11 July 2024

Available online 23 July 2024

Fig. 1. Illustrations of major challenges in person search, e.g., occlusion, pose variation, and scale variation. Bounding boxes sharing the same color indicate persons belonging to the same identity. For clarity, smaller-scale individuals are zoomed in and displayed in the bottom right corner.

Table 1
Main abbreviations of this paper.

Abbreviation	Meaning	Abbreviation	Meaning
AMP	Multi-granularity Perception module	MFE	Multi-granularity Feature Enhancer
PPE	Pedestrian Perception Enhancement block	SFM	Spatial-wise Feature Mixer
BIS	Background Interference Suppressor block	CFM	Channel-wise Feature Mixer

Table 1 lists the main abbreviations used throughout the paper.

identification. Although these methods yield promising results, they are computationally expensive. In contrast, end-to-end methods [1,16–20] adopt a unified network to address person detection and ReID simultaneously in an efficient multitask approach. Such methods [1,4,15,19] employ Faster R-CNN [21] as the underlying detection framework and introduce an additional prediction head branch to generate ReID features. However, as illustrated in Fig. 1, the aforementioned methods still face several challenges.

1.1. Motivation

- Occlusions and appearance changes significantly challenge the accurate identification and tracking of individuals. These issues often arise when pedestrians are obscured by background objects or other people, resulting in incomplete visual data that confuses recognition algorithms. Graph-based methods [22,23] utilize topological information to reconstruct obscured parts and infer pedestrian structures, improving scene understanding but increasing computational complexity. Alternatively, some techniques [24,25] shuffle and shift tokens to simulate occlusion effects. While these methods train models to predict pedestrian appearances under various occlusion patterns and potentially enhance performance, their simulation-based nature may lead to critical information loss, especially when handling complex occlusions not well represented in training data.
- Significant pose and scale variations among pedestrians further complicate ReID recognition. Changes in an individual's posture can result in varied appearances, while scale variations may lead the same person to occupy varying amounts of pixel space across scenes. These variations pose substantial challenges in maintaining consistency in feature extraction and recognition. Feature pyramids [9,26] and deformable convolutions [16] are designed to enhance the adaptability of feature learning to address these issues. Although these techniques improve the ReID system's ability to recognize features across different poses and scales, they also risk incorporating unnecessary background features into the feature representation. Such indiscriminate feature fusion might introduce unwanted background interference, thus reducing the model's discriminative capabilities.

To address the aforementioned challenges, we argue that an effective person search framework should embody two fundamental characteristics. *First*, the framework must be robust to appearance variations of query pedestrians, enabling accurate identification and matching target individuals even under varying conditions of poses, scales, and occlusions. This requires the development of a highly adaptable module capable of generating discriminative features for existing person search algorithms, tailored to the dynamic nature of real-world environments and the diverse appearances within RoI. *Second*, in real-world applications, imprecise bounding box positioning may inadvertently capture unwanted background information. Therefore, the framework must also prioritize enhancing foreground-background distinguishability to reduce the noise from irrelevant background elements. This involves implementing an effective mechanism within the framework focusing more precisely on relevant features, separating crucial pedestrian data from

noisy backgrounds, and ensuring accuracy and reliability in cluttered and dynamic settings. By overcoming these core challenges, the proposed framework aims to significantly enhance the accuracy and reliability of person search systems, rendering them more effective in complex urban scenarios where multiple factors interfere with the detection and recognition processes.

1.2. Innovation

Motivated by the above analysis, we tailor an end-to-end framework named AMPN to tackle occlusions and pose/scale variations in person search. This framework leverages a plug-and-play *Attentive Multi-granularity Perception* (AMP) module to capture and distinguish the crucial features while enhancing the distinction between foreground and background. Specifically, the module consists of a *Pedestrian Perception Enhancement* (PPE) block and a *Background Interference Suppressor* (BIS). The PPE block utilizes sequential spatial and channel attention methods to mitigate the effects of occlusion. Spatial attention targets crucial pedestrian regions by adapting activation responses, emphasizing key spatial cues. Concurrently, channel attention adjusts the importance of different channels to enhance the discriminability of pedestrian features, which is crucial for distinguishing individuals in complex visual scenes. Additionally, the PPE block incorporates a Multi-granularity Feature Enhancer that captures fine-grained pedestrian information within each RoI, enhancing feature robustness against appearance variations. Thus, by leveraging its advanced attention mechanisms, the PPE block focuses on the visible parts of pedestrians, ensuring accurate identification even in cases of heavy occlusion. Meanwhile, the BIS block is designed to minimize the interference of non-pedestrian elements in images. It utilizes 3D attention weights to regulate features from RoI-Align pooling, effectively filtering out irrelevant background information and enhancing the distinction between the foreground and background. This is particularly important in crowded public spaces, as the BIS module significantly enhances the model's reliability by reducing background noise, ensuring that pedestrian features are not obscured by background clutter. Extensive experiments show the state-of-the-art performance of our AMPN with an mAP of 95.2% on CUHK-SYSU and 52.4% on PRW dataset, validating its merits in addressing occlusion and pose/scale variations.

1.3. Contribution

- We propose an end-to-end **A**ttentive **M**ulti-granularity **P**erception **N**etwork (AMPN) to tackle challenging issues in person search, such as occlusions and pose/scale variations.
- To capture discriminative relationship features, we introduce a novel Attentive Multi-granularity Perception (AMP) module, which comprises a Pedestrian Perception Enhancement block and a Background Interference Suppressor. The PPE block aims to capture and enhance pedestrian features within local regions, mitigating the effects of occlusions and scale/pose variations, while the BIS block focuses on suppressing background interference to enhance foreground-background distinguishability.
- The AMP module serves as a plug-and-play component that can be seamlessly integrated with other person search algorithms at a low computational cost, thereby improving overall performance.
- Extensive experimental evaluations on two challenging standard datasets demonstrate the efficacy of our method in addressing occlusions and pose/scale variations in person search.

1.4. Organization

The rest of this paper is structured as follows: Section 2 offers a brief survey of the existing related work. Section 3 provides a detailed explanation of our methodology. In Section 4, extensive experiments and analyses are performed on CUHK-SYSU and PRW datasets. Finally, Section 5 presents the conclusion and future work.

2. Related work

2.1. Person search

Person search task involves locating and retrieving specific individuals in uncropped real scene images and encompasses two subtasks: person detection and ReID. This task was initially introduced by Xu et al. [2] and continues to garner significant attention in the research community. Existing methods are roughly categorized into two types based on their training steps.

Two-step methods: Zheng et al. [3] first propose a two-step approach for this task and explore different compositions of detectors and ReID models. Lan et al. [9] first discover the issue of diverse pedestrian target sizes. Chen et al. [8] indicate the inherent task optimization issue and get rich feature representations using a two-stream model. Han et al. [10] propose a Region of Interest transformation module to improve the quality of pedestrian bounding box outputs in the detection model. Dong et al. [27] introduce a novel detection network that learns similarity scores between proposals and queries to effectively reduce the number of proposals. Wang et al. [11] utilize an identity-guided query detector to achieve a higher query recall rate and address the consistency issue in the framework. Although two-step methods achieve remarkable performance, these approaches are both time-consuming and resource-intensive.

End-to-end methods: Most end-to-end person search methods are built upon the Faster R-CNN [21] detector. These approaches typically extend Faster R-CNN by incorporating an additional ReID branch, enabling simultaneous handling of the detection and ReID tasks. Among them, Xiao et al. [4] first present an end-to-end framework for this task and introduce an Online Instance Matching (OIM) loss. Xiao et al. [13] employ the center loss to enhance the intra-class compactness of learned representations, while Yan et

al. [28] develop a graph-based network to efficiently utilize complementary cues. Chen et al. [15] decoup the shared embedding of the two subtasks into radial norms and angles. Beyond the Faster R-CNN backbone, Yan et al. [16] first propose the anchor-free model in person search and resolve the misalignment problem across multiple levels (i.e., scale, region, and task). Yu et al. [24] introduce cascaded Transformers to generate discriminative fine-grained person representations. Li and Miao [1] propose a dual-headed network that shares the stem features for detection and ReID, sequentially addressing these two tasks. Jaffe et al. [19] reduce the size of the search gallery by reducing similar scenes, saving computational resources. Fiaz et al. [25] propose a scale-aware network that aggregates the scale information within RoI. Han et al. [26] propose an enhanced decoupling and memory-reinforced network to obtain more discriminative features. Song et al. [29] propose to improve the quality of person bounding boxes by considering interactions between persons and scenes.

Although these algorithms are highly efficient, their performance deteriorates in situations with heavy occlusions or noisy backgrounds. This issue arises because their design does not adequately consider the discriminative relational features between local regions and complex backgrounds. Consequently, these algorithms often fail to recognize key pedestrian features, thereby decreasing performance. To address this issue, we design a novel AMP module, which aims to obtain discriminative relational features among different local regions within the RoI. Simultaneously, it effectively suppresses background interference.

2.2. Person re-identification

Person re-identification seeks to match a query person with an extensive set of cropped pedestrian images. Initially, the focus was primarily on handcrafted features that possessed strong discriminative power. With the rise of deep learning [30–32,50], Sun et al. [33] apply a universal Part-based Convolutional Baseline (PCB), which evenly partitions images into six parts to extract finer-grained features. While effective, this method may have limited scalability in complex scenarios. Wang et al. [34] design a three-branch feature extraction network known as the Multiple Granularity Network (MGN), which segments global features to obtain features at different granularities, enhancing feature richness but increasing model complexity. Graph-based methods [22,23] employ topological information modeling to tackle occlusions, which is effective in theory but often computationally expensive. Zhou et al. [35] develop a lightweight Omni-Scale network that extracts and fuses features at various scales, improving efficiency but possibly struggling with extreme scale variations.

While these methods have improved performance, their model structures tend to be relatively complex. In contrast, our approach generates rich features using convolutional layers of varying sizes in both spatial and channel aspects. This strategy allows us to comprehensively capture details and semantic information in pedestrian images, without the need for additional branches or predefined fixed region partitions.

2.3. Attention mechanism

Attention mechanisms simulate human attention by allocating weights to focus on task-relevant information, thereby enhancing model performance. ECA [36] employs channel-wise convolution to adaptively learn dependencies, which is highly effective but may overlook spatial correlations. Woo et al. [37] utilize sequential channel and spatial attention structures to highlight important information in both dimensions, though sometimes at the expense of increased computational complexity. Hou et al. [38] address both spatial and channel relationships and long-range dependencies, enhancing global context awareness but potentially adding latency. Yan et al. [39] introduce a diversity regularization to improve the expressiveness of attention modules, which helps to prevent overfitting to dominant features yet could dilute the attention's focus. Additionally, variants like Multi-head Attention [40,41] allow handling of long sequences and multimodal inputs more effectively, though they may introduce complexity that complicates model training and inference.

However, these mechanisms often fail to capture subtle yet crucial features, leading to performance declines in specific tasks. To overcome this, our approach incorporates a sequential structure of spatial and channel attention modules and introduces Gaussian modulation functions. This innovation enhances subtle yet essential features in person search tasks, improving performance in occluded scenarios.

To quickly understand the work related to this paper, Table 2 provides a brief summary of the related work.

3. Methodology

In this section, we first introduce the problem formulation in Section 3.1. Then, the framework overview of AMPN is described in Section 3.2. Next, we introduce our proposed Attentive Multi-granularity Perception (AMP) in detail in Section 3.3. Lastly, we describe the training and inference stages in Section 3.4. As a preparation, the main variables of this paper are shown in Table 3.

3.1. Problem formulation

Given a query person q in the query image Q and a set of gallery images $I = \{I_1, I_2, \ldots, I_N\}$, the objective of person search is to detect a set of pedestrian bounding boxes B within I, and subsequently find the best matching pedestrian for q in B as the output. During training, the network model $\mathcal{F}(\cdot)$ utilizes labeled datasets $D = \{(x_i, y_i)\}_{i=1}^{N}$ to perform feature learning from images with occlusions and pose variations, resulting in discriminative feature maps $f_i = \mathcal{F}(x_i)$, where x_i represents the training images and y_i represents their corresponding labels. In the testing phase, for a query person q in the query image Q, the trained network uses $\mathcal{F}(\cdot)$

Table 2
A brief summary of the related work.

Typical work	Description	Evaluation
Person Search (Two-step methods):		
[3], [9], [8]	Enhance feature extraction mechanisms to derive richer, finer-grained features.	Enhance the model's ability to handle multi-scale information but significantly increase computational complexity.
[10], [27], [11]	Utilize query information to bolster the capabilities of detection networks.	Generate high-quality bounding boxes, but efficiency decreases with multiple instances.
Person Search (End-to-end methods):		
[4], [13]	Employ loss functions to enhance intra-class compactness and feature robustness.	Improve feature discriminability, but add computational burden during training.
[15], [1], [26]	Focus on optimizing model structure and parameter tuning to enhance performance.	Improve detection precision and ReID accuracy, but training complexity is high and prone to overfitting.
[24], [25]	Introduce Transformer structures to generate distinctive fine-grained person representations.	Achieve better accuracy, but the Transformer structure requires higher computational complexity.
[16]	Use anchor-free detection models for person search to boost efficiency and performance.	Simplify the model design and increase speed, but sacrifice some detection accuracy.
[28], [19], [29]	Utilize contextual cues to improve model precision and understand scene context.	Improve model accuracy, but increase the computational burden and rely on precise parameter adjustment.
Person Re-Identification:		
[33], [34], [35]	Extract features of different levels and scales to augment the model's recognition capabilities.	Adapt to different scenes and variations, but the complex model structure makes the optimization process more challenging.
[22], [23]	Utilize topological information to handle occlusions.	Boost model accuracy, but the computational expense is high.
Attention Mechanism:		
[36]	Use channel-level convolution to adaptively learn dependencies.	Improve model accuracy, but ignore spatial correlations.
[38], [37]	Address spatial and channel relationships and long-range dependencies.	Enhance global contextual information, but overlook subtle yet essential features.
[39]	Introduce diversity regularization to improve the performance.	Prevent overfitting, but it might divert the focus of attention.
[40], [41]	Employ multi-head attention mechanisms to more effectively handle long sequences and multimodal inputs.	Improve model accuracy, but model training and inference become complex.

Table 3
Main variables of this paper.

Variable	Meaning	Variable	Meaning
I_i	The gallery images	M_c	The channel attention weights
f_q	The feature of the query person q	$\mathcal{V}_A$	The Gaussian distribution
f_{ij}	The feature of the j-th pedestrian in the i-th image	K	The output of the PPE block
F	The feature of RoI-Align pooled	O	The output of the BIS block
M_s	The channel attention weights	H	The output of the AMP module

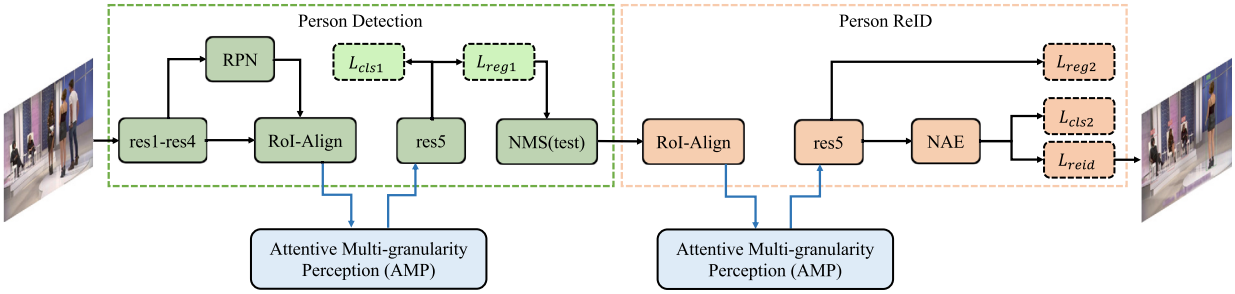

Fig. 2. An overall of the proposed AMPN for end-to-end person search, comprising two main branches: the Person Detection branch and the Person ReID branch. The detection branch is responsible for initially predicting the bounding box positions of the scene images. In contrast, the ReID branch refines these positions and employs Norm-Aware Embedding (NAE) to decouple the shared features between detection and ReID. Our main focus is on introducing a novel AMP module, which is independently integrated into both the detection and ReID branches.

to obtain the feature vector f_q. For each gallery image I_i, a detector outlines bounding boxes $B_{ij} = \{B_{i1}, B_{i2}, \ldots, B_{iM}\}$, from which feature vectors f_{ij} are then extracted. The network computes the similarity between f_q and each f_{ij}, subsequently identifying the bounding box with the highest similarity score to determine the best match for the query person.

3.2. Framework overview

Our proposed framework consists of two branches as depicted in Fig. 2: the *person detection branch* and the *person ReID branch*. In the detection branch, we utilize the widely used Faster R-CNN, which employs the ResNet-50 backbone network (res1-res4) to capture coarse-grained stem feature representations from the scene image. These stem features are then used by a Region Proposal

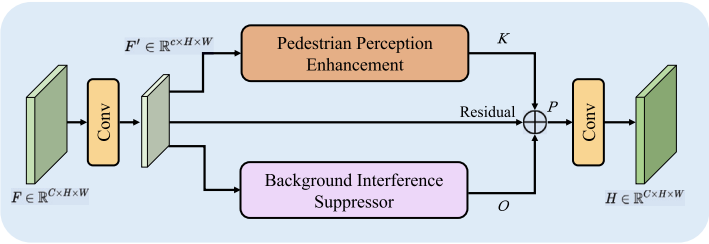

Fig. 3. The structure of the AMP module consists of two key components: Pedestrian Perception Enhancement (PPE) and Background Interference Suppressor (BIS). The PPE module enhances the recognition of pedestrian features under occlusion and scale variations by capturing discriminative relationship features. The BIS suppresses background noise and improves the distinguishability between pedestrian features and background clutter. Combining these components enables the AMP module to more accurately retrieve pedestrians in complex environments. After processing by the AMP module, the size and channel count of the input features remain unchanged.

Network (RPN) to generate region proposals. Each proposal is pooled into a unified size of 1024×14×14 regions using RoI-Align [42]. Following the RoI-Align operation, the feature maps are augmented by a novel AMP module before being fed into the res5 layer. The enhanced feature maps are subsequently processed by the res5 layer to further extract fine-grained features for bounding box regression and classification. In the ReID branch, the predicted bounding boxes from the detection branch are used as input, and RoI-Align pooling is performed on these bounding boxes again. The pooled features obtained are then employed for ReID tasks. We introduce the NAE [15] to decouple the shared features of person detection and ReID in polar coordinates.

The blue area in Fig. 2 represents our proposed AMP module. Specifically, we incorporate the AMP module separately between the RoI-Align and res5 layers of both the detection and ReID branches, without sharing the parameters between these two branches. Next, we provide a detailed explanation of the proposed AMP module.

3.3. Attentive multi-granularity perception

The structure and workflow of the AMP module are depicted in Fig. 3. The input features F are first dimensionally reduced through a point-wise (1×1) convolutional layer to decrease the channel dimension to $c = C/4$, enhancing computational efficiency, resulting in F'. Subsequently, F' is processed by two main branches: the Pedestrian Perception Enhancement (PPE) and the Background Interference Suppressor (BIS). The PPE branch aims to capture more detailed pedestrian features within various local regions to alleviate the impact of occlusions and scale/pose variations. While the BIS branch aims to suppress background noise, thus enhancing the distinction between foreground and background features. The outputs of these branches are then fused along with the original features F' through a residual connection. The combined features P are further refined by a convolution layer before producing the final output H, which maintains the same resolution as the input features.

3.3.1. Pedestrian perception enhancement

As mentioned earlier, it is advantageous to learn these perceptual-enhanced pedestrian features explicitly due to the variations in pedestrian pose/viewpoint/scale and occlusions within the RoI region. This approach enhances the model's generalization ability and operates without supervision. To achieve this, we develop a Multi-granularity Feature Enhancer (MFE) module, which utilizes convolutional operations of different sizes within the same feature map. This design allows for better control over the richness of information and adaptability to pedestrians of various scales. Smaller convolution kernels effectively capture detailed information for small-scale pedestrians, while larger kernels are more suitable for capturing coarse-grained information for large-scale pedestrians. Therefore, a PPE block integrates a Spatial-wise Feature Mixer and a Channel-wise Feature Mixer, as depicted in Fig. 4(a).

Spatial-wise Feature Mixer: When using MFE to learn enriched feature information in RoI sub-regions, there may exist cases where foreground regions are entangled with the background, which affects the performance of person re-identification and bounding box prediction. To distinguish the interfering background at the spatial feature level, we introduce spatial attention before the MFE to focus on the foreground information within the RoI area.

As shown in Fig. 4(b), for spatial attention, we first generate $F^s_{max} \in \mathbb{R}^{1\times H\times W}$ and $F^s_{avg} \in \mathbb{R}^{1\times H\times W}$ by performing MaxPool and AvgPool along the channel axis. These features are then concatenated and further convolved. Inspired by [43], we identify that the most sensitive features correspond to distinctive regions, while the minor features represent important but easily overlooked regions, and the insensitive features represent background elements. Therefore, we employ a modulation function to enhance the minor features and suppress the most sensitive and insensitive features. The operations described are represented as follows:

$$M_s = G\left(f^{7\times7}\left(\left[F^s_{max}; F^s_{avg}\right]\right)\right) \tag{1}$$

where M_s denotes the spatial attention weights, capturing high activation values reflecting the regions that are frequently neglected. $f^{7\times7}$ denotes a 7x7 convolution, and [] signifies concatenation along the channel dimension. The modulation function G redistributes feature map activations to highlight spatially important but easily overlooked features.

$$\mathcal{V}_A = G\left(\mathcal{V}_{A_f}\right) \tag{2}$$

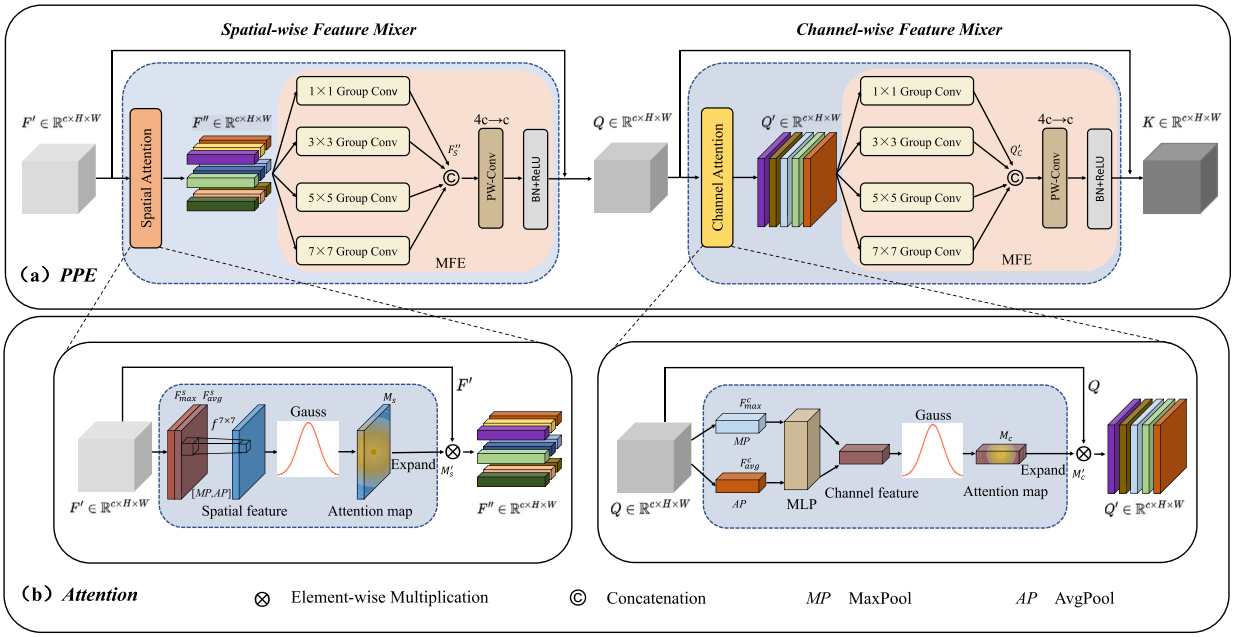

Fig. 4. Illustration of the Pedestrian Perception Enhancement (PPE) block. (a) Depicts the entire PPE pipeline, comprising two components: the Spatial-wise Feature Mixer and the Channel-wise Feature Mixer. Additionally, the MFE module utilizes a series of group convolution operations of different sizes to generate discriminative multi-granularity relational features, thereby alleviating the impact of appearance variations. (b) Represents the spatial and channel attention modules designed in our approach, each incorporating a Gaussian modulation function to modulate the activation maps of features in a sequential spatial-channel manner, thereby enhancing subtle yet vital features in occluded scenes.

where G represents the Gaussian function that maps all the activation values to a Gaussian distribution ($\mathcal{V}_A$). The parameters of "mean" and "std" are calculated by $\mathcal{V}_{A_f}$:

$$\mu = \frac{1}{M}\sum_{i=1}^{M}\left(\mathcal{V}_{A_f}^{i}\right), \quad \sigma = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\mathcal{V}_{A_f}^{i} - \mu\right)^2} \tag{3}$$

$$G(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{4}$$

where μ and σ respectively denote the mean and standard deviation of the activation map. Then, we follow the settings of μ and σ to project the activation values.

We perform element-wise multiplication between $F' \in \mathbb{R}^{c\times H\times W}$ and the spatial attention map to generate the reassigned feature, defined as:

$$F'' = F' \otimes M_s' \tag{5}$$

where $M_s' \in \mathbb{R}^{c\times H\times W}$ represents the feature map of spatial attention broadcasted along the channel dimension of M_s. The symbol $\otimes$ signifies element-wise multiplication, and F'' represents the output from the spatial attention.

These spatial attention-focused features are designed to distinguish background areas irrelevant to the foreground in the spatial domain. Subsequently, they are fed into a multi-granularity feature enhancer that employs four different-sized convolution operations $\{1\times1, 3\times3, 5\times5, 7\times7\}$. To ensure that the features generated by these convolutions have consistent size, we set the padding size to 0, 1, 2, and 3, respectively, according to the formula $p = (k-1)/2$. Group convolution is utilized to reduce parameters, with the number of groups set equal to the input feature channels. Finally, the four features are then combined in the channel dimension, as represented by the following operations:

$$F_S'' = \left[f_g^{1\times1}(F''); f_g^{3\times3}(F''); f_g^{5\times5}(F''); f_g^{7\times7}(F'')\right] \tag{6}$$

$$Q = ReLU\left(BN\left(PW\left(F_S''\right)\right)\right) \tag{7}$$

where $f_g^{1\times1}$, $f_g^{3\times3}$, $f_g^{5\times5}$, $f_g^{7\times7}$ represent group convolution operations with different kernel sizes, combined in the channel dimension to result in the feature vector F_S''. PW refers to a pointwise convolution used to reduce the dimensionality of F_S'' to c.

Overall, our Spatial-wise Feature Mixer block is designed to achieve precise and rich spatial feature fusion, generating feature Q.

Channel-wise Feature Mixer: As shown in Fig. 4 (b), to further prioritize channel features relevant to person detection and ReID, a channel attention layer is incorporated before the channel feature mixer. Specifically, channel-wise MaxPool and AvgPool are performed on the feature map Q, generating two feature descriptors: $F^c_{max} \in \mathbb{R}^{c\times1\times1}$ and $F^c_{avg} \in \mathbb{R}^{c\times1\times1}$. Each descriptor is separately passed through a MLP. Subsequently, we use the Gaussian function G to reassign feature distributions, emphasizing important but easily overlooked features in the channel dimension. The above operations are represented as follows:

$$M_c = G\left(W_1\left(W_0\left(F^c_{max}\right)\right) + W_1\left(W_0\left(F^c_{avg}\right)\right)\right) \tag{8}$$

where M_c represents the channel attention weights, W_0 and W_1 are the weights of the MLP, and G is the Gaussian function as mentioned earlier.

Element-wise multiplication between the input feature Q and the channel attention map results in the reweighted feature, defined as:

$$Q' = Q \otimes M'_c \tag{9}$$

where $M_c{}' \in \mathbb{R}^{c\times H\times W}$ represents the channel attention map broadcasted along the spatial dimensions of M_c, and Q' denotes the resultant of the channel attention layer.

These channel attention-focused features aim to emphasize the relevant channels required before channel feature mixing. Our channel feature mixer utilizes another shared MFE for the global mixing of channel information. The implementation process is described as follows:

$$Q'_C = \left[f_g^{1\times1}(Q'); f_g^{3\times3}(Q'); f_g^{5\times5}(Q'); f_g^{7\times7}(Q')\right] \tag{10}$$

$$K = ReLU\left(BN\left(PW\left(Q'_C\right)\right)\right) \tag{11}$$

where Q'_C is the feature vector obtained after concatenating along the channel dimension. BN stands for Batch Normalization, and $ReLU$ represents a non-linear activation function. $K \in \mathbb{R}^{c\times H\times W}$ represents the final output of the PPE module (as shown in Fig. 4(a)).

3.3.2. Background interference suppressor

To further enhance the performance of the pooled features in foreground/background discrimination, we apply a Background Interference Suppressor in the AMP module to simultaneously handle features in both spatial and channel aspects. Unlike traditional attention mechanisms, which treat spatial and channel attention separately, our approach considers them jointly, reflecting the interdependent nature of these mechanisms in human visual processing. Inspired by [44], we use an energy function to enhance the 3D RoI-Align pooled features. The implementation of our Background Interference Suppressor is defined as:

$$O = Sigmoid\left(\frac{1}{E}\right) \otimes F' \tag{12}$$

where E represents the obtained minimum energy function and O represents the output of the BIS.

As illustrated in Fig. 3, the features after the BIS are combined with the output of the PPE module, generating rich features P for person search tasks. The feature P is subsequently convolved and projected back to C dimensions ($H \in \mathbb{R}^{C\times H\times W}$) before being fed as input for the res5 layer.

3.4. Training and inference

During the *training* phase, we conduct end-to-end training for our AMPN model, addressing both person detection and person ReID tasks. For each branch, a threshold of 0.5 is used to select positive and negative samples for training. The model is supervised using five different loss functions combined linearly:

Regression loss (L_{reg1} and L_{reg2}): We use the Smooth-L1 loss function for bounding box regression, following the Faster R-CNN framework.

$$L_{reg} = \frac{1}{N_p}\sum_{i=1}^{N_p} L_{smooth-L1}\left(t_i, t_i^*\right) \tag{13}$$

where N_p represents the count of positive samples, t_i represents the computed regression for the i-th positive sample, t_i^* signifies the ground-truth regression, and $L_{smooth-L1}$ represents the Smooth L1 loss.

Classification loss (L_{cls1} and L_{cls2}): We utilize a binary cross-entropy loss to evaluate the ability of bounding box classification.

$$L_{cls1} = \frac{1}{N}\sum_{i=1}^{N} L_{CE}\left(p_i, p_i^*\right) \tag{14}$$

where N represents the sample count, p_i represents the estimated classification probability for the i-th sample, p_i^* represents the ground truth label, and L_{CE} represents the cross-entropy loss function.

$$f_{nae} = NAE\left(f_{reid}\right) \tag{15}$$

$$L_{cls2} = \frac{1}{N}\sum_{i=1}^{N} p_i^* L_{CE}\left(f_{\text{nae}}, p_i^*\right) \tag{16}$$

where f_{reid} represents the extracted 256-dim ReID feature and f_{nae} represents the feature after NAE [15] decoupling.

ReID loss: L_{reid}. Similar to conventional person search models, we use the classical non-parametric OIM [4] loss as the ReID loss. It maintains a lookup table $V \in \mathbb{R}^{D\times L} = \{v_1, \ldots, v_L\}$ and a circular queue $U \in \mathbb{R}^{D\times Q} = \{u_1, \ldots, u_L\}$ to retain the features of recent mini-batches of labeled and unlabeled identities. We can rapidly calculate the cosine similarity between the mini-batch samples and the LUT/CQ to facilitate feature learning.

$$L_{reid} = OIM\left(f_{nae}\right) \tag{17}$$

The overall loss function L_{sum} is defined as follows, with λ_1 set to 10 and others to 1 following [1].

$$L_{\text{sum}} = \lambda_1 L_{\text{reg1}} + \lambda_2 L_{\text{cls1}} + \lambda_3 L_{\text{reg2}} + \lambda_4 L_{\text{cls2}} + \lambda_5 L_{\text{reid}} \tag{18}$$

For clarity, we show the detailed training process of the proposed AMPN in Algorithm 1.

Algorithm 1 Training process of AMPN.

Input: Training set I, total epochs $epochs$, batch size b
Output: Trained Model weight $\mathbb{W}$

```
Initialize the model weight 𝕎
for e = 1 to epochs do
    for each batch b sampled from I do
        Extract pedestrian features F through Backbone (res1-res4) and RoI-Align from b
        Generate region proposals from a Region Proposal Network (RPN)
        for i = 1, 2 do
            Input F_i into AMP (get F_i')
            Scale F_i' and implement PPE according to Eqs. (1)-(11) (get K_i)
            Scale F_i' and implement BIS according to Eq. (12) (get O_i)
            Concatenate K_i, O_i, and F_i' along the channel dimension to form P_i and then scale to H_i
            Compute the regression loss L_reg_i and classification loss L_cls_i according to Eqs. (13)-(16)
        end for
        Compute ReID loss L_reid according to Eq. (17)
        Calculate the total loss L_sum to supervise the training process according to Eq. (18)
        Back propagate to update 𝕎
    end for
end for
return trained model weight 𝕎
```

During the *inference* phase, we initially use the provided bounding box to obtain the ReID features for a specific query person. Subsequently, we process the gallery images through our AMPN model to extract the predicted bounding boxes and their corresponding ReID features from the ReID branch. Finally, we use the cosine similarity between the ReID features to match the query person with any detected individuals in the gallery. It is worth noting that the ReID branch only utilizes the top 128 predicted bounding boxes retained by the detection branch through Non-Maximum Suppression (NMS).

4. Experiments

4.1. Datasets and evaluation protocols

CUHK-SYSU [4]: CUHK-SYSU is a large-scale person search dataset that includes 18,184 scene images and 96,143 annotated bounding boxes. The images are collected from two sources, real street/city scenes, and movie/TV snapshots. The training set includes 55,272 pedestrians, 11,206 frames, and 5,532 identities, while the testing set contains 40,871 pedestrians, 6,978 frames, and 2,900 identities. During the testing phase, we evaluate the search performance across a range of predefined gallery sizes from 50 to 4,000. Unless specified otherwise, our reported results are based on a gallery size of 100.

PRW [3]: PRW presents a more challenging person search dataset, comprising 11,816 video frames captured by six cameras placed at various locations in Tsinghua University. It contains 932 labeled pedestrians and 34,304 manually annotated bounding boxes. The annotations are categorized into labeled and unlabeled identities. The training set consists of 5,704 frames and 482 identities, while the testing set includes 2,057 query persons across 6,112 frames.

Evaluation protocol: We adhere to standard evaluation metrics to evaluate performance. For person detection, we use Recall and Average Precision (AP). For person ReID, we use the mean Average Precision (mAP) and top-1 score (top-1). Higher values of these metrics indicate superior model performance.

4.2. Implementation details

We adopt the SeqNet [1] as the baseline for our proposed method, which includes a ResNet-50 backbone, a Faster R-CNN detection head, and an OIM ReID head. For optimization, we use stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay

Table 4
Main hyper-parameters and computational environment of this paper.

Main hyper-parameters		Software and hardware	
Epochs (e)	15	Software Platform	PyTorch 1.7
Learning Rate (lr)	0.003	Python Version	3.8
NMS Threshold (u)	0.4	Hardware	NVIDIA Tesla V100 GPU
Batch Size (b)	3	Memory	32 GB

Table 5
Comparison with the state-of-the-art methods on CUHK-SYSU and PRW datasets. Our models are presented in italics. The bold entities denote the best performance achieved by two-stage and end-to-end methods, respectively.

Method		Ref	Backbone	CUHK-SYSU		PRW	
				mAP	top-1	mAP	top-1
Two-step	DPM [3]	CVPR17	ResNet50	-	-	20.5	48.3
	MGTS [8]	ECCV18	VGG16	83.0	83.7	32.6	72.1
	CLSA [9]	ECCV18	ResNet50	87.2	88.5	38.7	65.0
	RDLR [10]	ICCV19	ResNet50	93.0	94.2	42.9	70.2
	IGPN [27]	CVPR20	ResNet50	90.3	91.4	47.2	87.0
	TCTS [11]	CVPR20	ResNet50	**93.9**	**95.1**	46.8	**87.5**
	OR [12]	TIP21	ResNet50	92.3	93.8	**52.3**	71.5
End-to-end	OIM [4]	CVPR17	ResNet50	75.5	78.7	21.3	49.4
	RCAA [46]	ECCV18	ResNet50	79.3	81.3	-	-
	IAN [13]	PR19	ResNet50	76.3	80.1	23.0	61.9
	CTXG [28]	CVPR19	ResNet50	84.1	86.5	33.4	73.6
	HOIM [47]	AAAI20	ResNet50	89.7	90.8	39.8	80.4
	NAE [15]	CVPR20	ResNet50	91.5	92.4	43.3	80.9
	AlignPS+ [16]	CVPR21	ResNet50	94.0	94.5	46.1	82.1
	AGWF [17]	ICCV21	ResNet50	93.3	94.2	53.3	87.1
	CANR [18]	TCSVT22	ResNet50	92.4	93.2	43.4	83.8
	PSTR [45]	CVPR22	ResNet50	93.5	95.0	49.5	87.8
	COAT [24]	CVPR22	ResNet50	94.8	95.2	54.0	89.1
	GLCNet [20]	ICASSP23	ResNet50	94.3	94.9	45.7	87.7
	DMRNet++ [26]	TPAMI23	ResNet50	94.4	95.5	51.0	86.8
	SPG [29]	TII24	ResNet50	95.0	95.9	48.4	89.8
	SAT [25]	WACV23	ResNet50	95.3	96.0	55.0	89.2
	SeqNeXt+GFN [19]	WACV23	ResNet50	94.7	95.3	51.3	**90.6**
	SeqNet [1]	AAAI21	ResNet50	94.8	95.7	47.6	87.6
	AMPN(ours)	-	ResNet50	95.2	95.9	52.4	88.2
	AMPN(ours)	-	SE-ResNet50	95.8	96.1	53.6	88.2
	AMPN(ours)	-	Swin-S	**96.1**	**96.5**	**56.7**	89.5
	NAE+AMP(ours)	-	ResNet50	93.6(↑2.1)	94.0(↑1.6)	46.4(↑3.1)	81.7(↑0.8)
	SeqNet+AMP(ours)	-	ResNet50	95.2(↑0.4)	95.9(↑0.2)	52.4(↑4.8)	88.2(↑0.6)
	COAT+AMP(ours)	-	ResNet50	95.3(↑0.5)	96.1(↑0.9)	55.2(↑1.2)	89.7(↑0.6)
	GLCNet+AMP(ours)	-	ResNet50	95.1(↑0.8)	95.9(↑1.0)	50.6(↑4.9)	88.7(↑1.0)

of 5×10^{-4}. We train for 15 epochs on both datasets. *During training*, we apply data augmentation with only RandomHorizontalFlip at a probability of 0.5. The batch size is set to 3, and the input size is resized to 900×1500. We initialize the learning rate at 3×10^{-3}, with a warm-up period during the first epoch, followed by a reduction to 3×10^{-4} at the 8th epoch. *During the inference phase*, we utilize NMS with a threshold of 0.4 to eliminate overlapping bounding boxes. Our implementation is based on PyTorch 1.7, and runs under Python 3.8. All experiments are conducted on a single NVIDIA Tesla V100 GPU with 32 GB of memory. A summary of the main hyper-parameters and computational environment is shown in Table 4.

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

4.3.1. Results on CUHK-SYSU

The left column of Table 5 presents the comparison results of our method with other methods on CUHK-SYSU. Compared to the best-performing two-stage model, TCTS [11], our AMPN outperforms it by 1.3% in mAP. Among the end-to-end methods, AlignPS+ [16] adopts a multi-granularity anchor-free representation, COAT [24] uses a three-stage cascaded transformer, and SeqNet employs a two-stage refinement. Our AMPN with a ResNet50 surpasses these methods by 1.2%, 0.4%, and 0.4% in mAP, respectively. Notably, CUHK-SYSU, compared to PRW, provides fewer images per scene, and the proportion of complex scenes is not as high as in PRW. Consequently, the limited diversity in the data restricts AMPN's ability to enhance performance in complex scenarios.

4.3.2. Results on PRW

The performance of our method on PRW is detailed in the right column of Table 5. Compared to the existing two-step approaches, our method surpasses the best-performing OR [12] and TCTS [11] and achieves a 52.4% mAP and an 88.2% top-1 score. In terms of end-to-end methods, AlignPS+ and SeqNet, achieve mAP scores of 46.1% and 47.6%, respectively. Although AlignPS+ utilizes a stronger object detector, FCOS, compared to Faster R-CNN, it still exhibits poor performance. In comparison, our AMPN outperforms

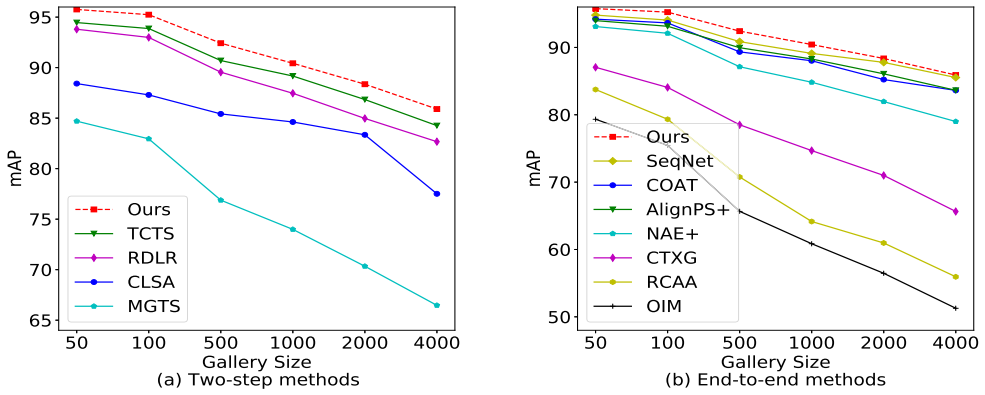

Fig. 5. Comparison with the existing two-step (a) and end-to-end approaches (b) on CUHK-SYSU with varying gallery sizes. The results are represented using dashed lines for our method and solid lines for the other methods.

SeqNet by **4.8**% in mAP and 0.6% in the top-1 score. Additionally, other end-to-end methods such as COAT [24], PSTR [45], and SeqNeXt [19] demonstrate excellent performance, with mAP scores exceeding 50% and top-1 scores above 86%.

4.3.3. Performance with different backbone

When we adopt a more powerful backbone (*SE-ResNet50, Swin-S*), the performance is further improved. On the PRW dataset, AMPN with Swin-S backbone achieves an mAP of 56.7% and an 89.5% top-1 score, surpassing SeqNet by **9**% in mAP. Although its top-1 score is slightly below that of SeqNext+GFN [19] (90.6%), it is important to note that SeqNext requires more complex GFN operations to filter scene images, potentially impacting its effectiveness. In addition, our AMPN outperforms all other methods on CUHK-SYSU, achieving an mAP of 96.1% and a top-1 score of 96.5%. This establishes a new benchmark for the person search task.

4.3.4. Generic of the AMP module

We further explore the compatibility and potential impact of our AMP method with other end-to-end frameworks. As shown in Table 5, our approach significantly improves the capability of NAE, COAT, GLCNet, and SeqNet. Notably, SeqNet and GLCNet demonstrate significant improvements on the PRW dataset, with respective mAP increases of 4.8% and 4.9%. These results confirm that our AMP module is not only versatile but also effective in enhancing the performance of various person search models.

4.3.5. Evaluation under different gallery sizes

We additionally assess the scalability of our method on CUHK-SYSU dataset with varying gallery sizes ranging from 50 to 4000. As depicted in Fig. 5, the mAP of all algorithms gradually declines as the gallery size increases, indicating that it becomes more difficult to find the individual under a larger search scope. However, our AMPN exhibits minimal performance degradation and outperforms existing models, achieving significant advantages across different gallery sizes. This indicates that our method demonstrates scalability and robustness, making it suitable for datasets with larger search scenarios.

4.3.6. Complexity analysis

Analyzing the complexity of our AMP module is crucial. The module incorporates multi-scale convolutions from MFE and a sequential structure of spatial and channel attention modules. Theoretically, the complexity of AMP is $O(b \times ((173C + 100) \times H \times W + 8C^2 \times H \times W + 2 \times \frac{C^2}{r}))$. Here, C represents the input channel size (i.e., 256 dimensions), b denotes the batch size, r is the downsampling rate, and H and W represent the width and height of the feature map, respectively. In our approach, the feature map size is 14×14, b is 3, and r is 16. Additional details about our actual computational complexity (MACs) and inference time (ms) are presented in Table 14.

4.4. Ablation study

4.4.1. Effectiveness of core components

We conduct a series of tests on the PRW dataset to assess the impact of each component in the designed AMP. Using SeqNet as our baseline, the results are displayed in Table 6. Data from rows 2 and 3 indicate that enabling the BIS alone improves ReID performance by 0.5%, confirming its effectiveness in reducing background noise. Furthermore, utilizing the PPE block, which includes the SFM and CFM, boosts ReID performance by 4.1%. This underscores the PPE module's efficiency in integrating multi-granularity information across spatial and channel domains, significantly improving the extraction of discriminative features.

Subsequently, we conduct a detailed ablation study on the PPE module. Data from rows 4 to 7 show that using SA, CA, and MFE independently enhances performance. Specifically, CA significantly boosts ReID performance, SA notably improves detection capabilities, and MFE proves crucial in the CFM. Data from rows 8 and 9 reveal that introducing the SFM and CFM separately positively impacts ReID performance, with both modules contributing similarly. The results in the final row show that when all core

Table 6
Comparison of the effects of different components in our method. In the table, "BIS" denotes the Background Interference Suppressor, "SFM" denotes the Spatial-wise Feature Mixer, "CFM" denotes the Channel-wise Feature Mixer, "SA" denotes the Spatial Attention, "CA" denotes the Channel Attention, and "MFE" denotes the Multi-granularity Feature Enhancer.

Baseline	BIS	SFM		CFM		ReID		Detection	
		SA	MFE	CA	MFE	mAP	top-1	Recall	AP
✓	✗	✗	✗	✗	✗	47.6	87.6	96.3	93.1
✓	✓	✗	✗	✗	✗	48.1	86.2	95.2	93.1
✓	✗	✓	✓	✓	✓	51.7	87.9	96.5	93.8
✓	✓	✓	✗	✗	✗	49.1	86.8	95.7	93.5
✓	✓	✗	✓	✗	✗	50.2	87.2	95.9	93.4
✓	✓	✗	✗	✓	✗	49.5	86.7	95.3	93.3
✓	✓	✗	✗	✗	✓	50.6	87.4	95.6	93.4
✓	✓	✓	✓	✗	✗	51.1	87.8	96.5	93.8
✓	✓	✗	✗	✓	✓	51.3	87.6	96.4	93.7
✓	✓	✓	✓	✓	✓	**52.4**	**88.2**	**96.7**	**93.9**

Table 7
Effect of different combination strategies on model performance.

Baseline	BIS	Gauss	Method	ReID	
				mAP	top-1
✓	✓	✗	CFM+SFM	51.7	87.3
✓	✓	✗	SFM+CFM	51.8	87.9
✓	✓	✗	SFM&CFM in parallel	51.2	86.9
✓	✓	✓	CFM+SFM	51.8	87.8
✓	✓	✓	SFM+CFM	**52.4**	**88.2**
✓	✓	✓	SFM&CFM in parallel	51.4	87.6

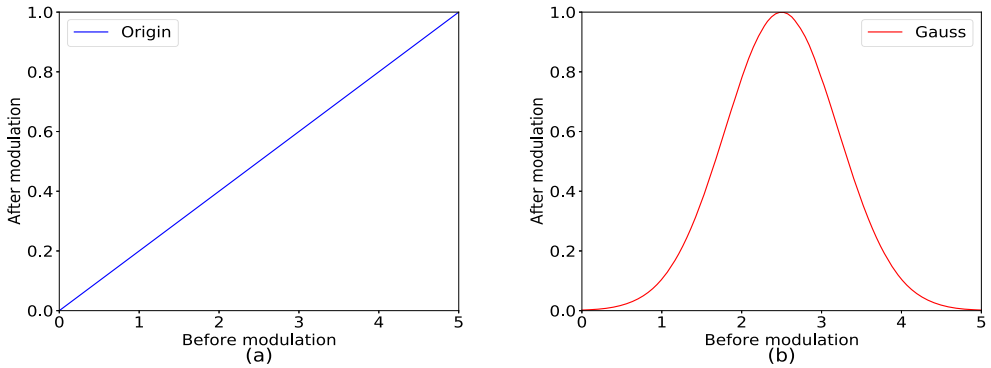

Fig. 6. Illustration of modulation function. The values on the axis represent the distribution range of activations. (a) represents the original activation distribution. (b) represents the activation distribution after Gaussian modulation.

components are used simultaneously, the model achieves optimal performance. This emphasizes the effectiveness of the synergistic interplay among these core components.

4.4.2. Comparison of different combining methods

We further investigate the impact of the CFM and SFM under different combination methods. Analysis of Table 7 shows that the sequential structure (SFM+CFM) is the most effective combination strategy. This means that spatial features are first reweighted and mixed through the SFM operation, followed by further enhancement of channel features through the CFM operation. This sequential order may help preserve important spatial details and perform finer adjustments at the channel level. From the data in the *third* and *sixth* rows of Table 7, we notice that although the parallel structure performs slightly lower than the sequential structure, it still brings some performance improvements. This could be attributed to the sequential structure's superior utilization of parameter learning capabilities.

4.4.3. Effectiveness of Gaussian function

Analysis of the last three rows of Table 7 shows that replacing the activation function for spatial attention in the SFM and channel attention in the CFM from the original sigmoid function to the Gaussian function has improved ReID performance. To further investigate the reasons behind this improvement, we visualize the activation value distributions before and after modulation in Fig. 6. We discover that the Gaussian function suppresses the highest and lowest activation values while highlighting the minor activation

Table 8
Comparison of other feature augmentation methods over PRW dataset. "Tokens" means token-level attention enhancement, and "Feats" means feature-level enhancement.

Method	Tokens	Feats	ReID	
			mAP	top-1
MLP-Mixer [48]	✓	✗	49.1	86.8
ViT Transformer [40]	✓	✗	48.9	85.8
PCB [33]	✗	✓	48.6	86.8
RFB [49]	✗	✓	49.4	87.1
OSNet [35]	✗	✓	50.7	86.7
MFE(Ours)	✗	✓	**52.4**	**88.2**

Table 9
Investigating the impact of different convolution scales on performance. "Granularity" represents the sizes of the used convolution kernels.

Granularity					ReID	
1×1	3×3	5×5	7×7	9×9	mAP	top-1
✓	✓	✓	✓	✓	50.9	87.3
✓	✓	✓	✓	✗	**52.4**	**88.2**
✗	✓	✓	✓	✗	51.9	87.7
✗	✗	✓	✓	✗	51.5	87.3
✗	✗	✗	✓	✗	50.6	87.2

Table 10
Results of different attention mechanisms over PRW dataset.

Method	PRW	
	mAP	top-1
AMPN w/o Attention(baseline)	50.8	87.2
baseline+CBAM [37]	51.0	87.5
baseline+ECA [36]	51.3	87.5
baseline+CA [38]	51.2	87.6
AMPN w/Attention	**52.4**	**88.2**

values, thereby directly extracting important but easily overlooked details. This is very important for our intricate person search challenges.

4.4.4. Analysis of feature augmentation

Table 8 shows the comparison of our designed MFE with other methods. MLP-Mixer [48] and ViT Transformer [40] employ token-level attention enhancement. However, experimental results indicate poor performance on such small 14×14 basic feature maps. This limitation likely arises from the reduced feature size, which hampers the accurate capture of crucial feature information. On the contrary, the PCB [33] block may potentially disrupt the holistic information of pedestrians, while the OSNet [35] block is complex and requires more computational resources. In contrast, the RFB [49] block employs varying dilation rates to achieve lightweight feature representations. Despite these methods incorporating feature-level enhancement, their performance falls short of our MFE, which excels at capturing multi-granularity details in images, particularly on small-sized feature maps.

4.4.5. Effect of multiple granularities

We further explore the impact of multi-granularity convolutions within the MFE. Table 9 shows that increasing the granularity of convolutions gradually, from bottom to top, significantly improves performance. However, the performance decreases noticeably when adding a $\{9\times9\}$ convolution. This suggests that the larger convolution size may cause information loss, indicating that excessive granularity can negatively impact performance. Additionally, combining 5×5 and 7×7 convolutions results in a 0.9% improvement in mAP over using a 7×7 convolution alone, highlighting the critical role of the 5×5 convolution in enhancing performance. These findings reinforce the rationale behind employing four levels of granularity in our approach.

4.4.6. Comparison of different attention mechanisms

We compare the performance of our attention module with other attention mechanisms in Table 10. To establish a baseline, we start with a configuration lacking any attention mechanism (achieved by removing spatial and channel attention, as shown in Fig. 4). Subsequently, we individually introduce CBAM, ECA, and CA attention to the baseline. Although these attention mechanisms yield performance gains, they are not as effective as our method.

In Fig. 7, the visualization of activation maps highlights the effectiveness of our attention mechanism. Our model more accurately focuses on pedestrian areas and effectively suppresses background noise, a capability that becomes particularly salient in scenarios

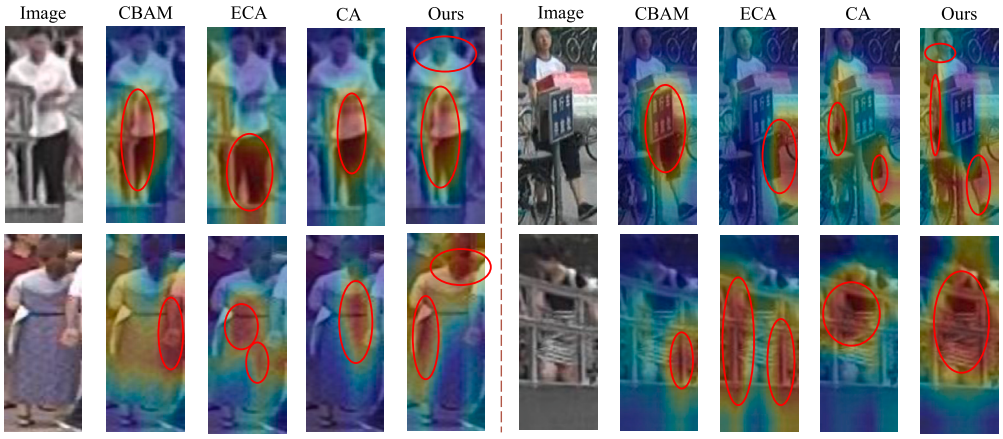

Fig. 7. Visualization of activation maps of different attention mechanisms *in occluded scenes*. From these images, we can see that CBAM tends to focus on the prominent areas within an image, yet these are not always our intended pedestrian targets. Both ECA and CA attention, being more spatially oriented, often only locate non-critical areas of the target pedestrian. In contrast, our method better focuses on pedestrian areas and effectively suppresses occlusion interference.

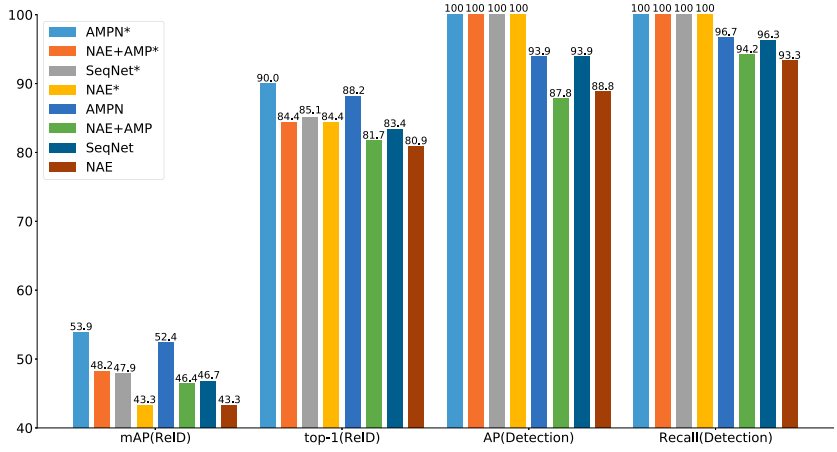

Fig. 8. Person search and person detection results on PRW with and without providing ground-truth detection boxes. The * denotes the ideal results using the ground-truth boxes.

with occlusions. A key factor contributing to this efficacy is the Gaussian activation function characteristic of our attention mechanism. This function assists in modulating activation values, diminishing the most extreme and least significant ones. Consequently, it accentuates the subtle activation signals, which often contain essential but overlooked details of pedestrian forms. Thus, this selective emphasis on important features leads to the extraction of more nuanced feature representations, significantly aiding the model in accurately identifying.

4.4.7. Relation between person detection and ReID

Improved detection results typically lead to enhanced performance in person search tasks. We conduct a comparison between our method and two Faster R-CNN based approaches, namely NAE and SeqNet. As shown in Fig. 8, when focusing solely on person ReID instead of person search, specifically when providing local ground-truth boxes, our AMPN outperforms other competitors, achieving a 6% improvement in mAP and a 5% increase in top-1 score. Furthermore, our detection AP values are comparable to SeqNet. These results indicate that the proposed method gains more advantages from the improved ReID features, rather than more accurate detection.

4.4.8. Analysis of hyper-parameters

We analyze the impact of various hyper-parameters on model performance, including the learning rate lr, NMS threshold u, and batch size b. As depicted in Fig. 9, concerning the learning rate, a value that is too high may cause instability during training, while one that is too low may slow down the training process. The model reaches optimal performance when the learning rate is set to 0.003. Regarding the NMS threshold, a higher threshold may result in filtering out too many overlapping bounding boxes, thereby weakening the model's ability to detect pedestrians. When the threshold is set to 0.4, the model achieves its best performance. As for

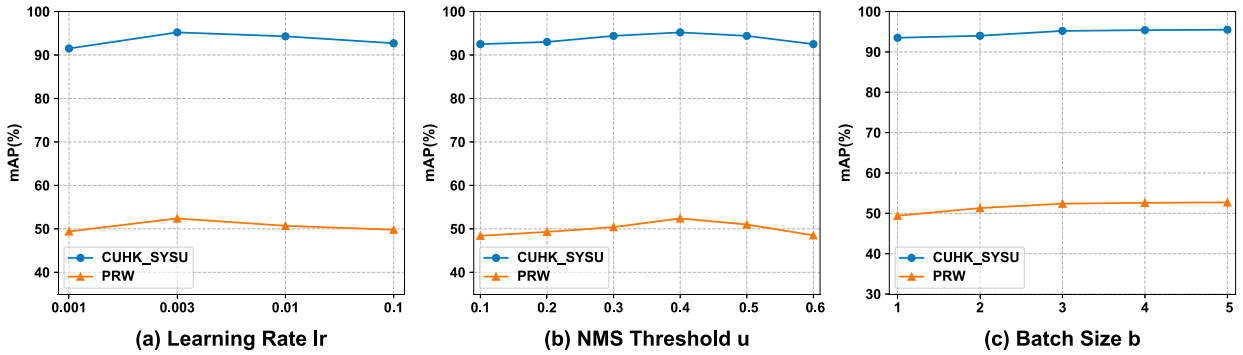

Fig. 9. Analysis the impact of hyper-parameters on model performance.

Table 11
Evaluate the proposed method in a cross-dataset scenario."CUHK-SYSU → PRW" means that the model is trained on CUHK-SYSU dataset while tested on PRW.

Method	CUHK-SYSU → PRW		PRW → CUHK-SYSU	
	mAP	top-1	mAP	top-1
OIM [4]	20.4	42.2	49.2	54.8
SeqNet [1]	25.6	71.8	50.6	55.6
AMPN(Ours)	27.6	76.8	52.5	57.3

the batch size, as it increases, the performance gradually improves and eventually stabilizes. However, considering that a batch size of 5 requires 30 GB of computational resources, we chose a batch size of 3 after a trade-off between performance and resources.

4.4.9. Generalizability on cross-dataset scenario

To validate the generalization ability of our model, we perform cross-dataset comparisons. Specifically, we directly use the model trained on the source dataset (e.g., CUHK-SYSU) to evaluate its performance on a different target dataset (e.g., PRW). We compare our AMPN model with the SeqNet baseline and OIM, and the results are presented in Table 11. It can be observed from the table that although the mAP decreases in both cross-dataset scenarios, the model trained on CUHK-SYSU outperforms the one on PRW. Since the CUHK-SYSU dataset contains a more diverse range of scenes, it demonstrates better transferability.

4.5. Qualitative results

In Fig. 10, we present visualized quantitative results comparing the AMPN method with SeqNet [1] and SAT [25]. Our method exhibits a distinct advantage in challenging scenarios, effectively eliminating interference from surrounding individuals in crowded scenes and accurately matching the target person. In contrast, SeqNet and SAT yield inaccurate retrieval results in complex scenarios with pose variations due to appearance deformations. In contrast, our approach utilizes the PPE block to consider the discriminative relationships and rich features within the RoI region, achieving correct matching results and obtaining more compact boxes. Even in the presence of scale variations and extreme lighting conditions, our method successfully retrieves the target. This is attributed to our MFE module's ability to capture detailed features of a person at different scales.

We also present some failure cases in Fig. 11. The first case illustrates a situation where the query person is heavily occluded, and the second case shows that the query person and the retrieval result have similar appearances. The analysis suggests two main reasons for these failures. Firstly, the training data include only a limited number of occlusion instances, potentially impeding the model's ability to develop effective strategies for handling occlusions. Secondly, our model lacks adequate comprehension and discriminative capabilities in complex scenarios.

4.6. Performance in various challenging scenes

4.6.1. Performance in the pose/viewpoint variations

We further assess performance on PRW's cross-camera gallery, as shown in Table 12. Our approach surpasses the performance of HOIM†, NAE+†, SeqNet†, and AGWF†. This superior performance is attributed to our AMP module, which produces more distinguishable ReID features, especially notable in cross-camera scenarios with pose/viewpoint variations. The remarkable cross-camera results underscore the potential of our method to be applicable in diverse locations, enabling accurate matching and recognition of pedestrians across different cameras, rather than being limited to similar scenes with the same camera ID.

4.6.2. Performance in the occlusions or scale variations

We analyze our AMPN method on these specific subsets of the CUHK-SYSU dataset, which include 187 and 290 representative sample queries for the occluded gallery and scale variations gallery, respectively. The outcomes are illustrated in Table 13, aware

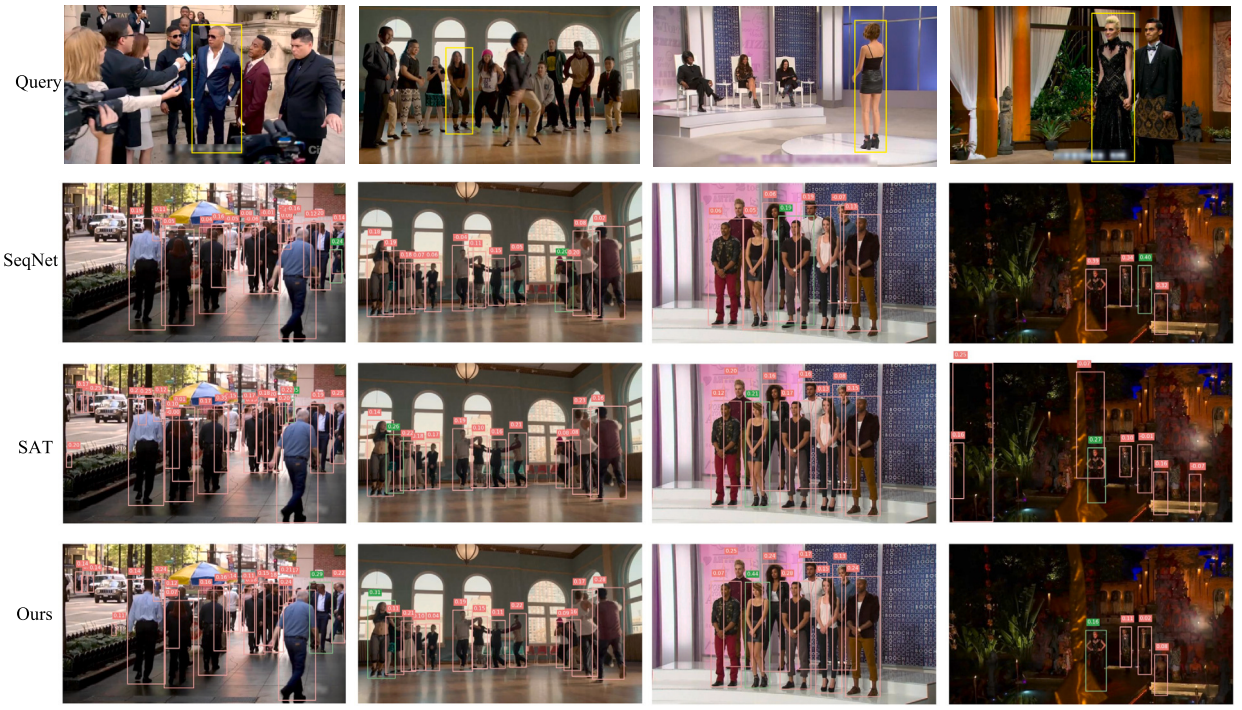

Fig. 10. Qualitative results on CUHK-SYSU dataset. Yellow boxes denote queries, green boxes denote correct top-1 matches and red boxes denote other detected persons. Compared with SeqNet and SAT, our AMPN is more robust to occlusion crowding (a, b), pose variation (c), and scale variation (d). Zoom in for better viewing.

Fig. 11. Failure cases on the PRW dataset. We illustrate failure cases to show the potential limitations of our method in cases of heavy occlusions and similar appearances.

Table 12
Performance on PRW test dataset for query person in scenarios with pose/viewpoint variations. † indicates the result tested on the cross-camera gallery.

Method	Cross-Cam ID	
	mAP	top-1
HOIM† [47]	36.5	65.0
NAE++† [15]	40.0	67.5
SeqNet† [1]	44.3	70.6
AGWF† [17]	48.0	73.2
AMPN†(Ours)	**49.3**	**74.6**

Table 13
Performance on two subsets of CUHK-SYSU using occluded (left) or scale variations (right) gallery to query person.

Method	Occluded		Scale Variations	
	mAP	top-1	mAP	top-1
SeqNet	88.25	89.29	85.06	85.81
SeqNet+AMP	89.09	89.69	85.79	86.24

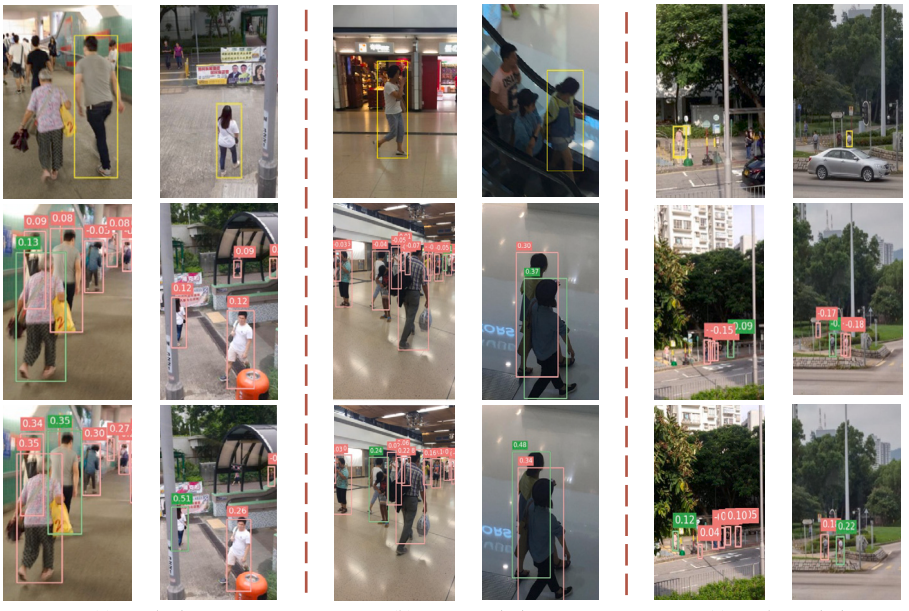

Fig. 12. Qualitative comparison between SeqNet and our method in three challenging scenes: (a) Occlusion, (b) Pose Variation, and (c) Scale Variation. Our approach consistently delivers the correct top-1 match in all instances.

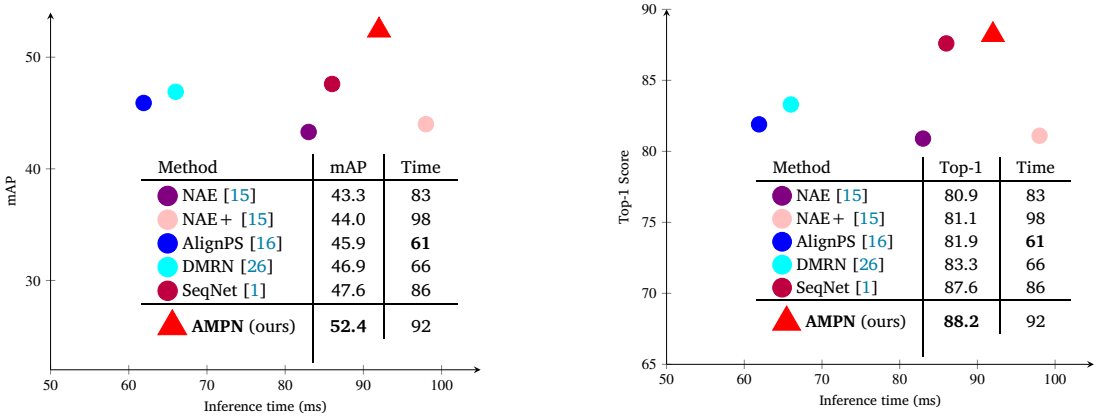

Fig. 13. Accuracy (AP) vs Time (ms) comparison with other end-to-end approaches on PRW dataset. We present accuracy as mAP (left) and top-1 score (right). These methods all use ResNet50 as the backbone and run on a single Tesla V100.

that the size of the search gallery is 100 in both cases. By addressing the challenges posed by occlusions and scale variations images, the AMP module significantly improves the accuracy and robustness of person search. Our approach enhances feature representation, mitigates the impact of occlusions, and captures key pedestrian information within the RoI, demonstrating outstanding performance compared to SeqNet. Moreover, we present the visualization results for these two galleries in Fig. 12. These visualizations further validate our method's capability to address common real-world person search tasks.

4.7. Efficiency comparison

We assess the efficiency of AMPN in comparison to other typical end-to-end networks using the same scale test image and hardware condition. The findings are presented in Table 14, comparing the Number of Parameters (M), Multiply-Accumulate Operations (MACs), and Inference Time (ms). Our method exhibits lower computational complexity and achieves a 4.8% mAP gain over SeqNet while maintaining a desirable speed. In Fig. 13, we visualize the forward inference time of our method. It can be observed that our AMPN method takes 92 ms to process a single image, which is even faster than NAE+.

Table 14
Comparison of person search efficiency.

Method	GPU(TFLOPs)	Params(M)	MACs(G)	Time(ms)	mAP	top-1
NAE [15]	V100(14.1)	**33.43**	414.16	83	43.3	80.9
NAE+ [15]	V100(14.1)	36.52	430.46	98	44.0	81.8
AlignPS [16]	V100(14.1)	42.18	**316.76**	**61**	45.9	81.9
DMRN [26]	V100(14.1)	-	-	66	46.9	87.6
SeqNet [1]	V100(14.1)	48.41	401.89	86	47.6	87.6
AMPN(Ours)	V100(14.1)	50.88	404.27	92	**52.4**	**88.2**

5. Conclusion

In this work, we present AMPN, an end-to-end framework specifically designed to address the challenges of occlusions and pose/scale variations in person search. Central to our framework is the novel AMP module, which effectively captures discriminative relation features within the ROI and exhibits robust performance against occlusions. This module includes a Pedestrian Perception Enhancement block, employing both Spatial-wise Feature Mixer and Channel-wise Feature Mixers to capture discriminative relation features. Additionally, a Background Interference Suppressor is introduced to enhance foreground/background discriminability in the joint space. The AMP can be seamlessly integrated with other person search models, and extensive experiments validate the merits of our AMPN, confirming its state-of-the-art performance.

Although our proposed AMPN achieves satisfactory results, there remains room for improvement in certain areas. (i) AMPN primarily analyzes the foreground (i.e., pedestrians) while neglecting the background (i.e., scene) information. However, scene information often provides valuable cues for person search. Therefore, introducing advanced scene-aware technologies such as Graph Convolutional Networks (GCN) can enhance performance by more deeply analyzing the potential relationships between scenes and pedestrians. (ii) While AMPN is designed to address occlusions and pose/scale variations, the overall network architecture has not been optimized, resulting in slightly slower inference speeds compared to some current methods. Employing heuristic/evolutionary search optimization intelligent algorithms, which continuously search and optimize model parameters, represents a promising direction for future research to improve inference speed and accuracy.

CRediT authorship contribution statement

Qixian Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Jun Wu:** Writing – review & editing, Supervision, Methodology. **Duoqian Miao:** Writing – review & editing, Supervision, Funding acquisition. **Cairong Zhao:** Writing – review & editing, Supervision, Resources. **Qi Zhang:** Writing – review & editing, 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.

Data availability

Data will be made available on request.

Acknowledgement

The work is supported by the National Key Research and Development Program (Grant No.2022YFB3104700), the National Natural Science Foundation of China (Grant Nos. 61976158, 62006172, 62376198, 62076182, 62163016), the Jiangxi Double Thousand Plan (No.jxsq2019102088), the Jiangxi Provincial Natural Science Foundation (No.20212ACB202001).

References

[1] Z. Li, D. Miao, Sequential end-to-end network for efficient person search, in: Proc. AAAI Conf. Artif. Intell. (AAAI), 2021, pp. 2011–2019.
[2] 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 (ACM Multimedia), 2014, pp. 937–940.
[3] 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.
[4] 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.
[5] K. Yuan, D. Miao, Y. Yao, H. Zhang, X. Zhao, Feature selection using zentropy-based uncertainty measure, IEEE Trans. Fuzzy Syst. 32 (4) (2024) 2246–2260, https://doi.org/10.1109/TFUZZ.2023.3347757.
[6] G. Zhang, W. Lin, A. kumar Chandran, X. Jing, Complementary networks for person re-identification, Inf. Sci. 633 (2023) 70–84, https://doi.org/10.1016/j.ins.2023.02.016.

[7] K. Wang, S. Dong, N. Liu, J. Yang, T. Li, Q. Hu, PA-Net: learning local features using by pose attention for short-term person re-identification, Inf. Sci. 565 (2021) 196–209, https://doi.org/10.1016/j.ins.2021.02.066.
[8] 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.
[9] X. Lan, X. Zhu, S. Gong, Person search by multi-scale matching, in: Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 536–552.
[10] 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.
[11] C. Wang, B. Ma, H. Chang, S. Shan, 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.
[12] H. Yao, C. Xu, Joint person objectness and repulsion for person search, IEEE Trans. Image Process. 30 (2020) 685–696, https://doi.org/10.1109/TIP.2020.3038347.
[13] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei, J. Feng, IAN: the individual aggregation network for person search, Pattern Recognit. 87 (2019) 332–340, https://doi.org/10.1016/j.patcog.2018.10.028.
[14] 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.
[15] 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.
[16] 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.
[17] B.-J. Han, K. Ko, J.-Y. Sim, End-to-end trainable trident person search network using adaptive gradient propagation, in: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 925–933.
[18] C. Zhao, Z. Chen, S. Dou, Z. Qu, J. Yao, J. Wu, D. Miao, Context-aware feature learning for noise robust person search, IEEE Trans. Circuits Syst. Video Technol. 32 (10) (2022) 7047–7060, https://doi.org/10.1109/TCSVT.2022.3179441.
[19] L. Jaffe, A. Zakhor, Gallery filter network for person search, in: Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2023, pp. 1684–1693.
[20] J. Qin, P. Zheng, Y. Yan, R. Quan, X. Cheng, B. Ni, MovieNet-PS: a large-scale person search dataset in the wild, in: Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2023, pp. 1–5.
[21] 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, https://doi.org/10.1109/TPAMI.2016.2577031.
[22] T. Wang, H. Liu, P. Song, T. Guo, W. Shi, Pose-guided feature disentangling for occluded person re-identification based on transformer, in: Proc. AAAI Conf. Artif. Intell. (AAAI), 2022, pp. 2540–2549.
[23] G. Wang, S. Yang, et al., High-order information matters: learning relation and topology for occluded person re-identification, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 6449–6458.
[24] R. Yu, D. Du, R. LaLonde, 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.
[25] 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.
[26] 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, https://doi.org/10.1109/TPAMI.2022.3221079.
[27] 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.
[28] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, X. Yang, Learning context graph for person search, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 2158–2167.
[29] Z. Song, C. Zhao, G. Hu, D. Miao, Learning scene-pedestrian graph for end-to-end person search, IEEE Trans. Ind. Inform. 20 (2) (2024) 2979–2990, https://doi.org/10.1109/TII.2023.3298473.
[30] C. Yan, B. Gong, Y. Wei, Y. Gao, Deep multi-view enhancement hashing for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 43 (4) (2020) 1445–1451, https://doi.org/10.1109/TPAMI.2020.2975798.
[31] Y. Li, D. Miao, H. Zhang, J. Zhou, C. Zhao, Multi-granularity cross transformer network for person re-identification, Pattern Recognit. 150 (2024) 110362, https://doi.org/10.1016/j.patcog.2024.110362.
[32] Y. Li, Y. Liu, H. Zhang, C. Zhao, Z. Wei, D. Miao, Occlusion-aware transformer with second-order attention for person re-identification, IEEE Trans. Image Process. 33 (2024) 3200–3211, https://doi.org/10.1109/TIP.2024.3393360.
[33] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline), in: Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 480–496.
[34] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, in: Proc. 26th ACM Int. Conf. Multimedia (ACM Multimedia), 2018, pp. 274–282.
[35] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Omni-scale feature learning for person re-identification, in: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 3702–3712.
[36] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: efficient channel attention for deep convolutional neural networks, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 11534–11542.
[37] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: convolutional block attention module, in: Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19.
[38] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 13713–13722.
[39] C. Yan, Y. Hao, L. Li, J. Yin, A. Liu, Z. Mao, Z. Chen, X. Gao, Task-adaptive attention for image captioning, IEEE Trans. Circuits Syst. Video Technol. 32 (1) (2021) 43–51, https://doi.org/10.1109/TCSVT.2021.3067449.
[40] A. Dosovitskiy, L. Beyer, et al., An image is worth 16x16 words: transformers for image recognition at scale, in: Proc. Int. Conf. Learn. Representations (ICLR), 2021.
[41] C. Yan, L. Meng, L. Li, J. Zhang, Z. Wang, J. Yin, J. Zhang, Y. Sun, B. Zheng, Age-invariant face recognition by multi-feature fusionand decomposition with self-attention, ACM Trans. Multimed. Comput. Commun. Appl. 18 (1s) (2022) 1–18, https://doi.org/10.1145/3472810.
[42] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2961–2969.
[43] P.-T. Jiang, Q. Hou, Y. Cao, M.-M. Cheng, Y. Wei, H.-K. Xiong, Integral object mining via online attention accumulation, in: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 2070–2079.
[44] L. Yang, R.-Y. Zhang, L. Li, X. Xie, SimAM: a simple, parameter-free attention module for convolutional neural networks, in: Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 11863–11874, http://proceedings.mlr.press/v139/yang21o.html.
[45] 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.

[46] X. Chang, P.-Y. Huang, Y.-D. Shen, 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.
[47] D. Chen, S. Zhang, W. Ouyang, J. Yang, B. Schiele, Hierarchical online instance matching for person search, in: Proc. AAAI Conf. Artif. Intell. (AAAI), 2020, pp. 10518–10525.
[48] I.O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al., MLP-Mixer: an all-mlp architecture for vision, in: Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2021, pp. 24261–24272.
[49] S. Liu, D. Huang, et al., Receptive field block net for accurate and fast object detection, in: Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 385–400.
[50] Q. Zhang, L. Cao, C. Shi, Z. Niu, Neural time-aware sequential recommendation by jointly modeling preference dynamics and explicit feature couplings, IEEE Trans. Neural Netw. Learn. Syst. 33 (2022) 5125–5137, https://doi.org/10.1109/TNNLS.2021.3069058.