Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 32(12): 6772–6791.
DOI: 10.3934/era.2024317
Received: 03 October 2024
Revised: 20 November 2024
Accepted: 10 December 2024
Published: 18 December 2024

Research article

A practical object detection-based multiscale attention strategy for person reidentification

Bin Zhang, Zhenyu Song*, Xingping Huang, Jin Qian and Chengfei Cai*

College of Information Engineering, Taizhou University, Taizhou 225300, China

* **Correspondence:** Email: songzhenyu@tzu.edu.cn, caichengfei@tzu.edu.cn.

Abstract: In person reidentification (PReID) tasks, challenges such as occlusion and small object sizes frequently arise. High-precision object detection methods can accurately locate small objects, while attention mechanisms help focus on the strong feature regions of objects. These approaches mitigate the mismatches caused by occlusion and small objects to some extent. This paper proposes a PReID method based on object detection and attention mechanisms (ODAMs) to achieve enhanced object matching accuracy. In the proposed ODAM-based PReID system, You Only Look Once version 7 (YOLOv7) was utilized as the detection algorithm, and a size attention mechanism was integrated into the backbone network to further improve the detection accuracy of the model. To conduct feature extraction, ResNet-50 was employed as the base network and augmented with residual attention mechanisms (RAMs) for PReID. This network emphasizes the key local information of the target object, enabling the extraction of more effective features. Extensive experimental results demonstrate that the proposed method achieves a mean average precision (mAP) value of 90.1% and a Rank-1 accuracy of 97.2% on the Market-1501 dataset, as well as an mAP of 82.3% and a Rank-1 accuracy of 91.4% on the DukeMTMC-reID dataset. The proposed PReID method offers significant practical value for intelligent surveillance systems. By integrating multiscale attention and RAMs, this method enhances both its object detection accuracy and its feature extraction robustness, enabling a more efficient individual identification process in complex scenes. These improvements are crucial for enhancing the real-time performance and accuracy of video surveillance systems, thus providing effective technical support for intelligent monitoring and security applications.

Keywords: person reidentification; object detection; YOLOv7; multiscale attention strategy

1. Introduction

With the development of smart cities, the demand for security has been growing, creating significant opportunities for the advancement of intelligent surveillance systems [1, 2]. Person reidentification (PReID) plays a crucial role in this context. PReID is a subproblem of image retrieval and is aimed primarily at determining whether a specific pedestrian appears in videos captured by different cameras [3]. The query inputs for such tasks can be images, video sequences, or even textual descriptions [4]. This paper focuses on the image retrieval task. Under various query conditions, PReID aims to overcome the visual limitations of fixed cameras, and it is widely applied in security fields in combination with pedestrian detection and tracking technologies. As a result, innovative reidentification technology has become a popular research focus in both academia and industry. In simple terms, the PReID task can be broken down into two key steps: feature extraction and feature matching. The goal of feature extraction is to obtain discriminative feature representations (FRs) from pedestrian images or video frames that can accurately reflect the identity of the target pedestrian. The objective of feature matching is to compare the extracted FRs with those in a gallery set to identify the same pedestrian. PReID leads to numerous challenges due to various factors, such as different camera angles [5], resolution variations [6], lighting condition changes [7], posture variations [8], occlusions [9, 10], and heterogeneous environments [11]. In real-world application scenarios, PReID often involves analyzing video sequences. The process typically begins by performing object detection to locate pedestrians in the input video; this is followed by the extraction of discriminative features and then the calculation of feature similarity values to identify objects. Therefore, the key challenges related to PReID lie in how to better extract discriminative FRs and design more effective metric learning loss functions.

Early PReID research relied primarily on manually extracting fixed discriminative features [12, 13] and learning better similarity metrics [14]. These methods are prone to errors and are time-consuming, which significantly affects their accuracy and real-time performance in PReID tasks. Moreover, owing to the challenges of cross-device variations and substantial differences among the captured images, traditional methods struggle to achieve satisfactory results. Currently, commonly used approaches in the field of PReID include supervised, semisupervised, and weakly supervised learning techniques. Among them, supervised learning methods can be further categorized into global FR, local FR, temporal FR, and auxiliary FR learning strategies.

To extract fine-grained features during global feature learning, Wu et al. [15] employed small convolutional kernels to capture fine-grained details from pedestrian images. Qian et al. [16] proposed a multiscale deep FR learning model that is capable of learning global features at different scales and adaptively matching them. In local FR learning methods, image partitioning strategies are typically divided into two categories: horizontal partitioning [17, 18] and pose estimation [19]. Sun et al. [17] and Zhang et al. [18] adopted the horizontal partitioning approach, where images are uniformly divided into six equal parts, with local features extracted from each segment. Su et al. [19] proposed a pose-driven deep convolutional model that utilizes human pose estimation for image segmentation purposes, addressing the challenge of pedestrian posture variations. Additionally, Nixon et al. [20] explored the use of gait for conducting video-based PReID. The key distinction between video-based methods and image-based methods is that the former not only consider the content information within individual frames but also account for the motion and temporal information between consecutive frames. Auxiliary FR learning enhances reidentification performance by extracting semantic information from pedestrian images or by using generative adversarial networks (GANs) to improve the FR learning process. In recent years, attention mechanisms have been widely used to enhance FR learning methods because of their strong performance. In [21], a harmonious attention network

(HA-CNN) was proposed; this approach jointly learns “soft” pixel attention and “hard” region attention to capture both global and local features, thereby attaining improved reidentification accuracy. A soft attention mechanism can be implemented in three ways: channel models, spatial models, and hybrid models that combine both spatial and channel attention. The squeeze-and-excitation network (SENet) [22] is a typical example of a channel attention network. It explicitly models the interdependencies between channels and adaptively recalibrates channel feature responses, thereby enhancing the representation ability of the network. To address the PReID problem, Wang et al. [23] designed a fully attentive module that addresses the loss of spatial structural feature information by SENet. Akin to SENet, this module can be integrated into different backbone networks to enhance their recognition capabilities. Tay et al. [24] proposed the attribute attention network (AANet), which is a novel architecture that integrates person attributes and attribute attention maps into a classification framework to solve PReID challenges. In summary, the existing PReID methods have achieved improved recognition accuracy through combinations of various FR learning techniques and attention mechanisms.

PReID technology is crucial for enhancing the real-time performance and accuracy of video surveillance systems in the intelligent monitoring and security fields. While the existing methods have achieved progress in specific scenarios, challenges remain with respect to handling complex scenarios, such as those with small targets, occlusions, and various lighting conditions. This study introduces a PReID method based on object detection and attention mechanisms (ODAMs); the proposed approach is designed to significantly improve the recognition performance achieved in these challenging scenes by enhancing the accuracy of target localization and the effectiveness of feature extraction. To address issues such as object occlusion and small object sizes, this paper proposes a PReID method that combines high-precision object detection with a multiscale attention strategy (MAS). In the literature [25–27], various strategies have been proposed for addressing cross-domain person reidentification (PReID). However, these methods require extensive labeled data and tend to be less effective when addressing occlusions and scale variations. In contrast, our approach integrates multiscale attention mechanisms and residual attention mechanisms (RAMs), significantly enhancing the robustness and accuracy of the model in complex scenarios. In this study, You Only Look Once version 7 (YOLOv7) is selected as the object detection algorithm due to its significant speed and accuracy advantages over other models such as YOLOv4 and RetinaNet. The YOLOv7 architecture optimizes the feature extraction layers and computational units of the network, enhancing its efficiency in terms of handling objects with various scales and complexity levels. Furthermore, YOLOv7 maintains the highest accuracy in real-time object detection tasks exceeding 30 FPS, making it an ideal choice for our study. Specifically, we employ YOLOv7 as the detection algorithm and integrate a size attention strategy into the backbone of the detection network to improve its detection accuracy. During the feature extraction process, ResNet-50 is used as the base network, and an RAM is introduced to construct a new feature extraction network for PReID. This approach allows the extracted information to focus more on the key local details of the target object, resulting in more effective FRs. The main contributions and innovations of this study are as follows.

- We propose a PReID method that combines high-precision object detection with an MAS, enabling accurate localization and extraction of the critical local information from target objects, thereby yielding improved detection accuracy.
- By integrating the RAM into the base network, we enhance the capacity of the feature extraction

network, making the extracted features more distinctive.

- Extensive experimental results obtained on public datasets demonstrate the effectiveness of the proposed method, particularly its robustness in terms of handling small targets and occluded objects.

The remainder of this paper is organized as follows. Section 2 provides a detailed introduction to the proposed method. Section 3 presents the dataset description, experimental setup, evaluation metrics, and experimental results, along with a discussion and an analysis. Finally, Section 4 offers the conclusion of this study and outlines potential future work ideas.

2. Methodology

2.1. Multiscale attention strategy

The MAS processes the input data with different scales by using a "pyramid attention" approach. As shown in Figure 1, the MAS model divides the input data into multiple layers at different scales, where the lower scale layers capture coarser global information, while the higher scale layers retain finer, more detailed local information. By applying attention mechanisms at various scales, the lower-scale attention focuses on global features, helping the model understand the overall target structure. On the other hand, at higher scales, the attention mechanism emphasizes detailed features, enabling the model to capture the crucial details and local characteristics. As a result, the MAS enhances the ability of the deep learning model to accurately differentiate between important information and background noise when processing inputs with different scales, improving the overall capabilities of the model. In PReID tasks, object size and posture variations are common, and detecting and recognizing small objects presents challenges due to information losses. By introducing an MAS, the constructed model can accurately distinguish key information from background noise across different scales, thus improving its ability to recognize small objects and objects in complex scenes. This makes MASs particularly crucial for complex visual tasks such as PReID. To address these challenges, this paper proposes a network architecture based on an MAS, as shown in Figure 1. The design of this network leverages the strengths of multiscale attention to significantly enhance both its accuracy and robustness in PReID tasks. The innovative design of this network enables the model to achieve notable performance improvements, particularly when addressing small objects and complex scenes. By effectively capturing and analyzing pedestrian features at different scales, the model is able to provide more precise recognition results.

Assuming that the MAS model has feature maps at different scales $X_1, X_2, ..., X_n$, where each X_i corresponds to a different scale, the MAS can be represented as follows:

$$\mathcal{A}(\mathcal{S}) = \sum_{i=1}^{n} Conv\left(\alpha_i(\mathcal{S})X_i\right), \tag{2.1}$$

where $\mathcal{A}(\mathcal{S})$ is the feature map produced after performing attention allocation and *Conv* refers to the convolution operation applied to the weighted feature map $\alpha_i(\mathcal{S})X_i$. Here, $\alpha_i(\mathcal{S})$ is the attention weight for the i-th feature map at the current scale $\mathcal{S}$. The calculation formula for $\alpha_i(\mathcal{S})$ is as follows:

$$\alpha_i(\mathcal{S}) = \frac{\exp\left(Q(\mathcal{S})\mathcal{K}\left(X_i\right)\right)}{\sum_{j=1}^{n} \exp\left(Q(\mathcal{S})\mathcal{K}\left(X_j\right)\right)}, \tag{2.2}$$

where $\mathcal{Q}(\mathcal{S})$ represents the query vector, which encodes the current state of the model or the information it is seeking, and $\mathcal{K}(X_i)$ represents the key vector, which captures the relevant information from different parts of the input data; the attention mechanism computes a similarity score between the queries $\mathcal{Q}$ and the keys $\mathcal{K}$. Typically, both $\mathcal{Q}$ and $\mathcal{K}$ are derived from the current state $\mathcal{S}$ of the model.

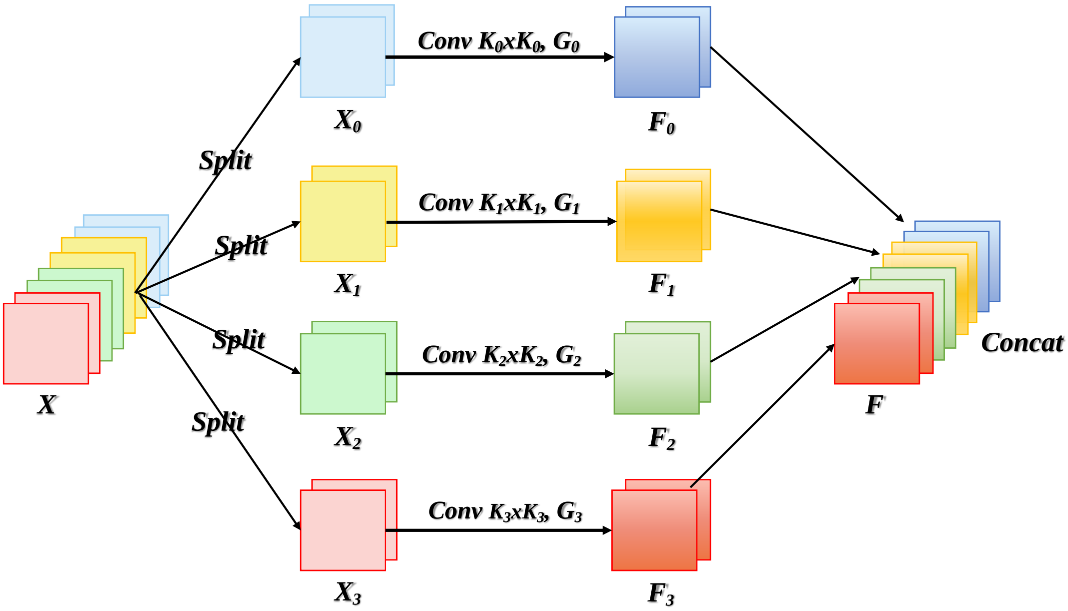


Figure 1. Architectural description of the MAS. (F_0 and F_3 represent feature maps with different scales, $ConvK_ix$ and G_i represent the convolution operations and attention weights applied at the i-th scale.)

2.2. *Object detection*

YOLOv7 is a significant version within the YOLO series and is known for being a lightweight and efficient object detection model with outstanding performance in terms of both speed and accuracy. Notably, within the range of 5 to 160 frames per second (fps), YOLOv7 outperforms most known object detectors, making it a top choice for real-time object detection tasks. Specifically, when the V100 GPU is used, YOLOv7 achieves the highest accuracy among real-time object detectors operating at over 30 fps. Compared with other versions of YOLO, YOLOv7 retains a detection approach similar to that of YOLOv4 and YOLOv5, but it optimizes the network architecture to improve both its detection efficiency and accuracy. As shown in Figure 2, the YOLOv7 network architecture incorporates deeper feature extraction layers and more effective computational units, allowing the model to excel when handling objects with varying scales and complexities. The YOLOv7 network model is composed of four main components.

- Input: The input section preprocesses the given image through techniques such as data augmentation and normalization, preparing it for subsequent processing steps.
- Backbone: After preprocessing, the image is passed through the backbone network, where its essential features are extracted.

- Neck: The extracted features are then fed into the neck module, where multiscale feature fusion is performed. This module generates feature maps at three different scales (large, medium, and small) to handle objects with various sizes.
- Head: Finally, these fused features are sent to the head module, where object localization and classification are carried out, producing the final detection results.

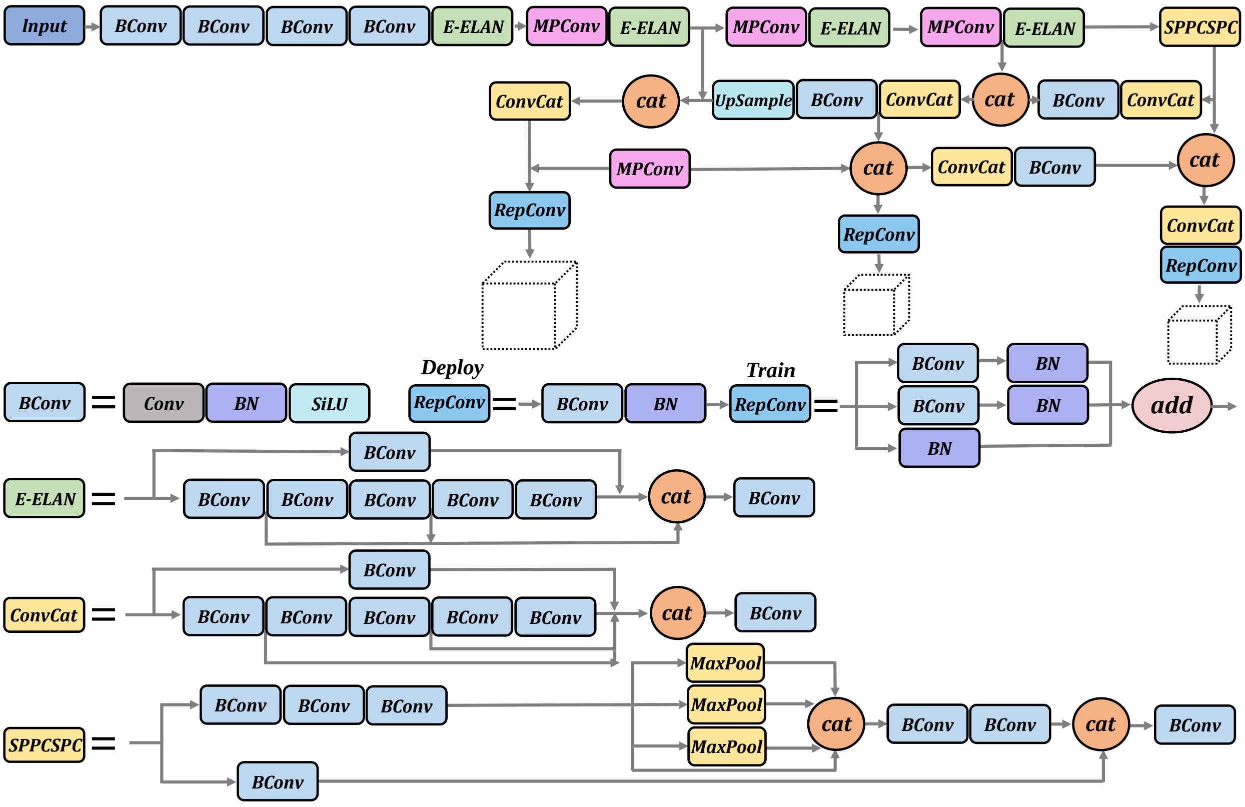


Figure 2. Architectural description of the YOLOv7 network.

This multipart architecture allows YOLOv7 to efficiently detect objects of various sizes with high accuracy and speed. The backbone of the YOLOv7 network model is meticulously designed and integrates several innovative modules, including convolutional layers, an extended efficient layer aggregation network (E-ELAN) module, an MPConv module, and a spatial pooling pyramid-based cross-stage partial convolution (SPPCSPC) module.

- E-ELAN Module: This module builds on the original ELAN transition structure and incorporates strategies such as expansion, shuffling, and cardinality merging to significantly enhance the learning capacity of the network while preserving the original gradient path.
- SPPCSPC Module: By introducing multiple parallel maximum pooling (MaxPool) operations within convolutional operations, the SPPCSPC module effectively prevents image distortions and resolves the feature redundancy issue that is commonly encountered in convolutional neural networks (CNNs).
- MPConv Module: This module expands the receptive field of the feature layer through MaxPool operations, which are then fused with standard convolutional features, further improving the

generalization performance of the network.

In the neck module, YOLOv7 adopts the path aggregation feature pyramid network (PAFPN) structure from YOLOv5, ensuring an efficient multiscale feature fusion process. Finally, in the head section, YOLOv7 employs the IDetect detection head, which is optimized to handle large, medium, and small objects. Different configurations of the RepConv module are used in the head design to enhance the performance and efficiency of both the training and inference processes.

2.3. Residual attention network for feature extraction

Attention mechanisms can be categorized into two types. One type is top-down conscious attention, known as focused attention, which refers to goal-driven, task-oriented active attention. The other type is bottom-up unconscious attention, referred to as saliency-based attention, which is driven by external stimuli and operates automatically without deliberate intervention. This type of attention typically shifts its focus to prominent objects through mechanisms such as "winner-takes-all" or gating. The residual attention network is composed of multiple stacked attention modules. Each attention module consists of two parts: a mask branch and a trunk branch. The trunk branch is responsible for feature processing and can adopt any network model. In this paper, preactivated residual units, ResNet, and inception are used as the basic units.

Given an input x, the feature map output by the trunk branch is denoted as $\mathcal{T}_{i,c}(x)$. The mask branch, which combines bottom-up and top-down attention mechanisms, learns a mask $\mathcal{M}_{i,c}(x)$ with the same size as that of the output of the trunk branch. The mask $\mathcal{M}_{i,c}(x)$ acts as a weight for $\mathcal{T}_{i,c}(x)$. Therefore, the final feature map output by the attention module is as follows:

$$\mathcal{H}_{i,c}(x) = \mathcal{T}_{i,c}(x) \cdot \mathcal{M}_{i,c}(x), \tag{2.3}$$

where $\cdot$ denotes elementwise multiplication. This operation effectively applies the learned attention mask to weight the output of the trunk branch, enhancing the relevant features while suppressing irrelevant features. In the attention module, the attention mask functions as a feature selector during the forward pass, whereas during the backward pass, it acts as a filter for performing gradient updates:

$$\frac{\partial \mathcal{M}(x,\theta)\mathcal{T}(x,\emptyset)}{\partial \emptyset} = \mathcal{M}(x,\theta)\frac{\partial \mathcal{T}(x,\emptyset)}{\partial \emptyset}, \tag{2.4}$$

where θ represents the parameters of the mask branch and $\emptyset$ represents the parameters of the trunk branch. The noise resistance provided by the attention module effectively reduces the impact of noise on the gradient updates.

In image processing tasks, training images often pose challenges such as background occlusion, complex scene layouts, and significant appearance variations. This requires the utilized model to integrate multiple attention strategies to effectively capture key information. Without a stacked attention module architecture, a substantial increase in the number of feature channels would be needed to capture and integrate the complex interactions and attention weights among various factors. However, a single attention module can optimize features only once, which may limit the overall robustness and fault tolerance of the constructed model when addressing complex and variable image data. The mask branch consists of two key steps: a fast feedforward sweep and a top-down feedback process. The fast feedforward sweep rapidly gathers global information from the entire input image, whereas the top-down feedback step combines this global information with the original feature map. In CNNs, these two steps are implemented as a combination of bottom-up and top-down fully convolutional structures.

In this paper, the mask branch adopts a network architecture similar to that of fully convolutional networks. Figure 3 illustrates the structure of the mask branch. First, several maximum pooling operations are applied to quickly expand the receptive field. After the lowest resolution is reached, a symmetric network structure is used to restore the features to their original resolution through linear interpolation, with the number of interpolations matching the number of maximum pooling operations to ensure that the input and output dimensions are the same. Next, two consecutive 1 × 1 convolutional layers are added, and the output is normalized to the [0, 1] range via a sigmoid layer. Additionally, skip connections are introduced between the bottom-up and top-down processes to capture multiscale information, thereby facilitating feature extraction to compute object similarity values.

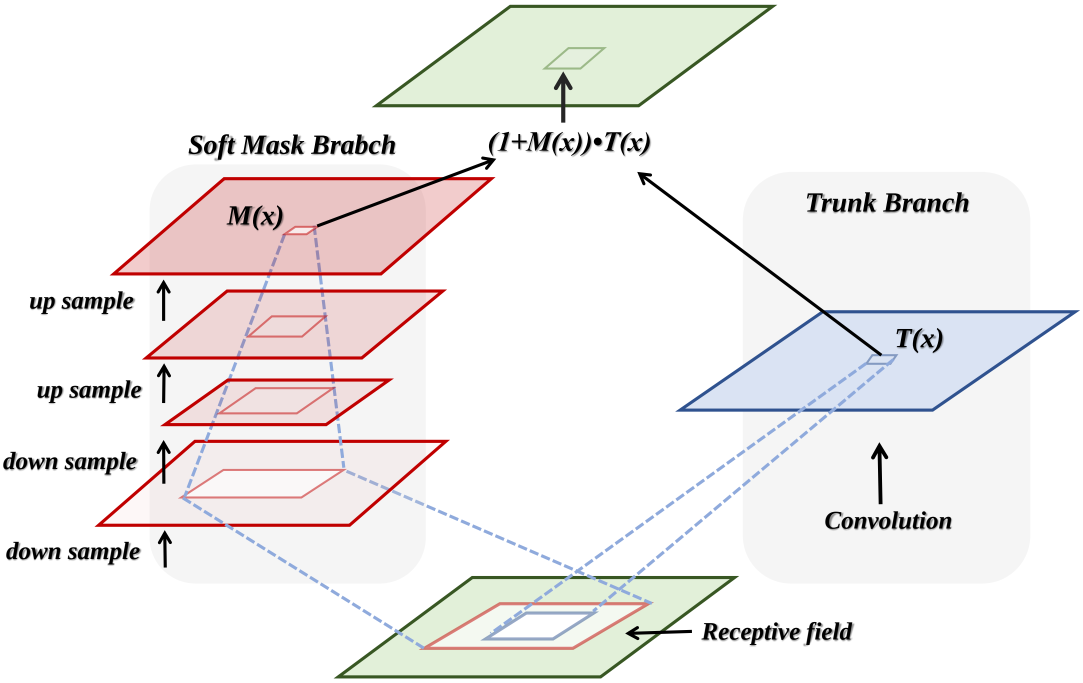


Figure 3. Receptive field comparison between the mask and trunk branches.

2.4. Proposed method

In this study, we propose an innovative fusion strategy for object detection tasks that combines the YOLOv7 framework with an MAS to effectively address the challenges associated with varying object sizes in images. Specifically, we design a multiscale attention module that enables the model to adaptively focus on objects with different sizes, thereby improving the resulting detection accuracy. Additionally, during the feature extraction phase, we integrate an RAM to further enhance the feature extraction capabilities of the model. This fusion technique not only preserves inherent detection efficiency of YOLOv7 but also optimizes its FRs, allowing the model to more accurately and comprehensively capture object features in complex scenes, thus significantly improving its overall detection performance. The implementation procedure is outlined in Algorithm 1.

Algorithm 1: Object detection-based MAS.

```
Input: images
Output: label
begin
    Initialize the YOLOv7 backbone network.
        network = YOLOv7().
    Define the multiscale attention module.
        class MultiScaleAttentionModule: ...
            return attention_map.
    Integrate the multiscale attention module into YOLOv7.
        class YOLOv7WithAttention(YOLOv7): ...
            return detections.
    Define the RAM for attaining enhanced feature extraction.
        class ResidualAttentionModule: ...
            return enhanced_features.
    Integrate the RAM into the feature extraction network.
        class FeatureExtractionNetworkWithResidualAttention: ...
            return enhanced_features.
    Integrate the entire system into an end-to-end detection model.
        class EndToEndDetectionModel: ...
            return detections.
    Implement a training loop and evaluate the model on the validation set.
        model = EndToEndDetectionModel().
            evaluate(model, validation_data).
    end.
```

3. Experimental studies

3.1. Experimental setup

To ensure the fairness of the experiments and the reliability of the results, all experiments in this study were conducted under a unified software environment and identical hardware conditions. The specific experimental setup was as follows: The network model was trained using Python 3.8 and the PyTorch-based deep learning framework, the integrated development environment (IDE) used was Visual Studio Code (VS), and the hardware environment was a Windows Server 2016 system equipped with an NVIDIA RTX 4090 GPU with 24 GB of VRAM. ResNet-50 was selected as the backbone CNN, and common data augmentation strategies, including random cropping, horizontal flipping, and random deletion, were employed. The input image size for all datasets was set to 256 × 128. The backbone network was pretrained on the ImageNet dataset. In the experiments conducted in this study, the following key hyperparameters were set: The initial learning rate was $8{\times}10^{-4}$, and the weight decay rate was 5×10^{-4}. The model began saving after 200 iterations, with subsequent saves occurring every 10 iterations. The maximum number of iterations was set to 600. When using batch normalization, the selected batch size N is critical to the effectiveness of the model training process. A batch size that is too small may result in the statistical estimates used for normalization failing to accurately reflect the distribution of the entire dataset, which can negatively impact the generalizability of the model and increase the risk of overfitting to small batch data. On the other hand, a batch size that is too large leads to increased memory usage and computational resource consumption, and the delayed gradient updates may reduce the efficiency of training. To strike a balance between computational efficiency and model performance, through an experimental validation and a resource optimization analysis, this paper set the batch size N to 64. This value ensured that batch normalization could effectively smooth the data distribution while promoting the stability and convergence speed of the model during training. This strategy avoided the potential issues caused by batch sizes that were too large or too small. With this configuration, the model demonstrated good stability and efficiency throughout the training process.

3.2. Dataset description

In this study, multiple public PReID datasets were used, including VIPeR [28], CUHK01 [29], CUHK02 [30], CUHK03 [31], Market-1501 [32], DukeMTMC-reID [33], and MSMT17 [34]. Detailed information about these datasets is provided in Table 1.

Table 1. Image-based PReID datasets.

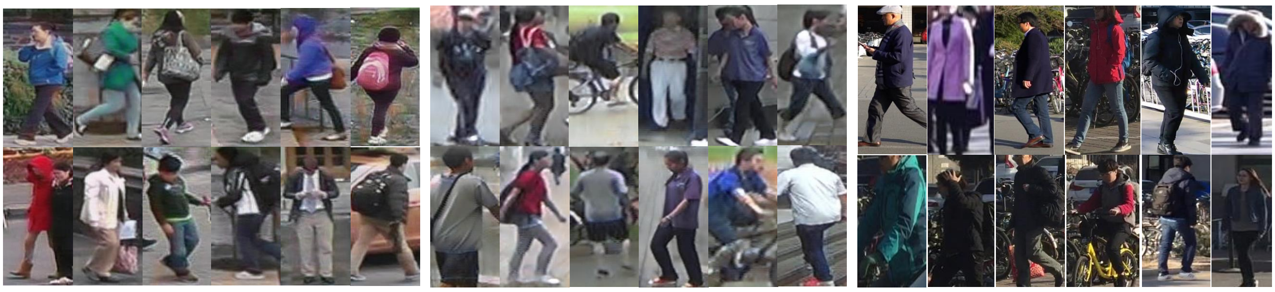
Datasets	Years	ID	Boxes	Cameras	Labeled
ViPeR	2007	632	1264	2	Handcrafted
CUHK01	2012	971	3384	2	Handcrafted
CUHK02	2013	1816	7264	10	Handcrafted
CUHK03	2014	1360	13164	10	DPM+Handcrafted
Market-1501	2015	1501	32217	6	DPM+Handcrafted
DukeMTMC-reID	2017	1812	36441	8	Handcrafted
MSMT17	2018	4101	126441	15	Faster R-CNN

During the experiments, we used Market-1501 [32], DukeMTMC-reID [33], and CUHK03 [31] as test sets to evaluate the reidentification performance of the model, while the remaining datasets were used as training sets. Example images from these datasets are shown in Figure 4.

- VIPeR is a challenging PReID dataset consisting of images captured by two cameras from different viewpoints. Each camera captures only one image per pedestrian, resulting in a total of 1264 images involving 632 unique pedestrians. Each pedestrian has two images, one from each camera. Due to the significant differences between the images in this dataset and its high level of reidentification difficulty, VIPeR is considered one of the most challenging PReID datasets.
- The CUHK01 dataset contains 971 unique pedestrians with a total of 3884 manually cropped images. A key feature of this dataset is that each pedestrian appears in at least two different camera viewpoints, making it useful for evaluating the performance of algorithms across varying perspectives. Specifically, Camera A captures more diverse viewpoints and posture variations, while Camera B provides frontal and rear views of the pedestrians, enhancing the diversity of the dataset.
- CUHK02 is an extended version of CUHK01, with increased scale and complexity. This dataset includes 1816 pedestrians and a total of 7264 images, introducing five pairs of camera viewpoints.

Compared to CUHK01, CUHK02 not only expands the number of pedestrians and images but also covers more diverse pedestrian image configurations, including different viewpoints, postures, image resolutions, lighting conditions, and illumination settings. This makes the dataset more suitable for evaluating PReID algorithms in complex scenarios.

- The CUHK03 dataset also uses five pairs of cameras to capture images and features two pedestrian annotation methods: traditional manual labeling and automatic pedestrian bounding box detection using the deformable part model (DPM) detector. This dataset contains 1360 unique pedestrians with a total of 13,164 images, and it possesses various image sizes to better simulate real-world scenes. By capturing pedestrian images from more viewpoints and incorporating automatic detection algorithms for annotation purposes, CUHK03 exhibits enhanced realism and practical applicability.
- The Market-1501 dataset was collected from a supermarket on the Tsinghua University campus using five high-resolution cameras and one low-resolution camera. Pedestrian bounding boxes were automatically detected and annotated using the DPM detector. The dataset contains a total of 1501 unique pedestrians and 32,668 images, with each image uniformly resized to 128 × 64 pixels. Compared to CUHK03, Market-1501 offers a larger number of annotated images and introduces 2793 query images and 500,000 distractors, providing a more realistic simulation of the randomness encountered in real-world scenarios.
- The DukeMTMC-reID dataset captures dynamic pedestrian images using 8 high-definition cameras. The dataset is divided into a training set (16,522 images), a query set (2228 images), and a gallery (17,661 images). This partitioning scheme allows for a comprehensive evaluation of algorithmic performance across different tasks.
- The MSMT17 dataset features images captured by 15 cameras across a university campus, covering more diverse scenes. An advanced pedestrian detection method using the faster region-based CNN (R-CNN) was employed to implement automatic detection and annotation, ensuring accurate and consistent data. MSMT17 includes 4101 unique pedestrians with a total of 126,441 images, making it one of the largest datasets in the PReID field. Compared to earlier datasets, MSMT17 presents more viewpoint variations and significant lighting differences, providing strong support for evaluating the performance of algorithms in complex environments.



(a) DukeMTMC-reID (b) Market-1501 (c) MSMT17

Figure 4. Sample images derived from the different datasets.

3.3. Evaluation metrics

PReID technology has become increasingly mature, and the associated evaluation metrics have also been progressively refined. Common evaluation standards include Rank-*n*, cumulative matching characteristic (CMC) curves, and mean average precision (mAP). Rank-*n* represents the probability that the correct pedestrian image appears within the top *n* retrieved images (sorted by confidence). Typically, evaluations range from Rank-1 to Rank-10, with Rank-1, Rank-5, and Rank-10 being the most commonly used metrics. Since Rank-1 is the most frequently used measure and holds significant reference value, it is one of the key evaluation metrics employed in this study.

Additionally, mAP is calculated by averaging the average precision values produced across multiple classification tasks. mAP is widely used for evaluating the performance of PReID models. Specifically, the formula for calculating mAP is as follows:

$$mAP = \frac{\sum_{i=1}^{N} Precision_i \times Recall_i}{\sum_{i=1}^{N} Recall_i}, \tag{3.1}$$

where $Precision_i$ represents the proportion of true-positive samples among all positive samples identified by the model at a recall rate of $Recall_i$. $Recall_i$ indicates the proportion of true-positive samples identified by the model at that recall rate out of the total number of positive samples. N represents the number of recall points used. This metric provides a comprehensive measure of how well the model ranks relevant images across the entire dataset.

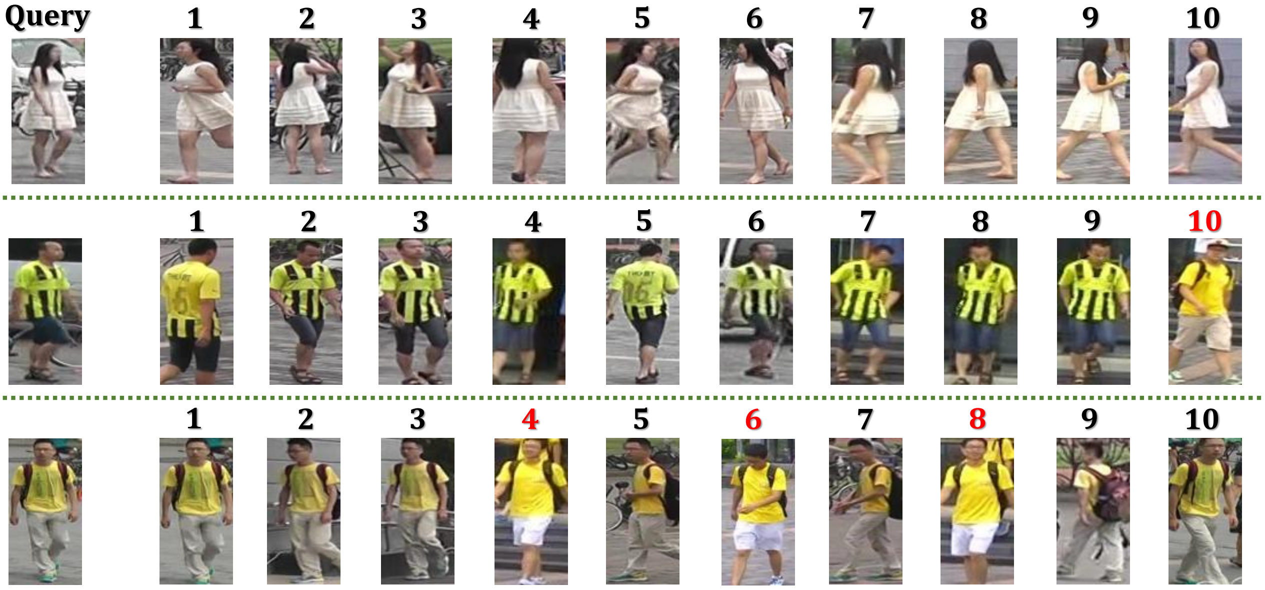


Figure 5. Reranking results obtained on the Market-1501 dataset.

3.4. Performance comparison

3.4.1. Comparison with the baseline

We conducted a comparative analysis of the performance of the proposed algorithm on three datasets and compared it with mainstream algorithms developed in recent years. The baseline network was ResNet-50, which was only pretrained, and the comparison results are shown in Table 2. By analyzing the experimental data produced for the Rank-1 and mAP evaluation metrics across the three datasets, it is evident that the proposed method demonstrated significant improvements over the baseline network while also exhibiting strong robustness. Figure 5 presents a visualization of the pedestrian reranking results obtained on the Market-1501 dataset, where black numbers indicate correct matches, red numbers indicate incorrect matches, and "Query" represents the query image.

Table 2. Performance comparison with the baseline network on three datasets.

Method	Market-1501		DukeMTMC-reID		CUHK03-L		CUHK03-D	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Baseline	73.8	79.3	82.5	67.4	69.4	61.5	61.2	59.4
Ours	**97.2**	**90.1**	**91.4**	**82.3**	**85.1**	**80.4**	**82.4**	**77.2**

3.4.2. Comparison with other models

In this study, we selected 11 commonly used PReID algorithms. These methods exhibit varying characteristics in terms of their performance and complexity, reflecting their adaptability across different scenarios. The reason for selecting these models was to comprehensively evaluate the performance of our method in various complex settings and to demonstrate its accuracy and robustness advantages. Overall, these methods can be categorized into feature extraction-based methods (MGCAM [23], Deep-Person [21], and HPM [35]), attention mechanism-based methods (HA-CNN [21], DuATM [36], and Mancs [17]), an adversarial learning-based method (AACN [18]), a model structure optimization method (SVDNet [37]), and methods that incorporate semantic segmentation (MaskReID [38], DSA-ReID [39], and SPReID [40]). Specifically, feature extraction methods rely on deep learning and hierarchical models to extract high-level and local features, thereby improving their reidentification performance. However, these methods often demand substantial data and computational resources and face limitations when addressing complex backgrounds and posture variations. By introducing spatial or channel attention, attention mechanisms significantly enhance the focus of a model on key features, improving its performance in complex environments. However, these methods typically come with higher computational complexity levels, longer training times, and greater hardware resource demands. In adversarial learning tasks, adversarial training reduces the domain differences between cameras, yielding enhanced robustness in cross-camera recognition scenarios. Nevertheless, the complexity of adversarial training increases the difficulty of hyperparameter tuning and extends the required training time. In contrast, model structure optimization techniques, such as the use of singular value decomposition to reduce feature redundancy, improves the resulting recognition performance while maintaining a relatively simple training process. Methods that incorporate semantic segmentation or parsing effectively reduce the degree of background interference and address occlusion issues, but they rely heavily on the accuracy of the semantic segmentation results and add to the complexity of the constructed model.

The PReID method proposed in this paper, which is based on a spatial attention mechanism, effectively addresses the aforementioned issues. Table 3 shows the experimental results produced by this method on the Market-1501 dataset. From the results, it is evident that different algorithms exhibited significant performance variations in terms of the Rank-1 and mAP metrics. Methods such as DSA-ReID, SVDNet, HPM, and Mancs all achieved excellent results for both metrics. Notably, DSA-ReID stood out with a Rank-1 score of 95.7% and an mAP of 87.6%, indicating that this method is highly stable and accurate when identifying pedestrians in complex scenes. In contrast, AACN and MaskReID yielded lower mAP values, at 66.9% and 70.3%, respectively. While MaskReID performed well in terms of Rank-1 (90.0%), its lower mAP suggests that its overall matching accuracy across multiple candidate images was inferior to that of other methods, possibly due to adversarial training and background interference handling limitations. Additionally, methods such as SPReID and Deep-Person also performed exceptionally well, with SPReID achieving a Rank-1 score of 92.5% and an mAP of 81.3%, highlighting the importance of semantic parsing in attaining enhanced recognition performance. Our proposed method achieved the highest Rank-1 (97.2%) and mAP (90.1%), significantly outperforming the other algorithms. This demonstrates that our approach offers clear advantages in terms of accuracy and comprehensiveness, enabling stable and efficient PReID across various complex scenarios.

Table 3. Performance comparison with other methods on the Market-1501 dataset.

Methods	**Rank-1**	**mAP**
MGCAM	83.8	74.4
MaskReID	90.0	70.3
AACN	85.9	66.9
SPReID	92.5	81.3
HA-CNN	91.2	75.7
DuA TM	91.4	76.6
Deep-Person	92.3	79.5
Mancs	93.1	82.3
HPM	94.2	82.7
SVDNet	94.8	86.0
DSA-ReID	95.7	87.6
Ours	**97.2**	**90.1**

Additionally, the experimental results obtained on the DukeMTMC-reID dataset are presented in Table 4. Compared with methods such as IDE [17], Gp-reID [41], MaskReID [38], SVDNet [37], PAN [17], AACN [18], SPReID [40], HA-CNN [21], MGN [42], and DSA-reID [39], our method achieved significant improvements in both the Rank-1 and mAP metrics, with a Rank-1 score of 91.4% and an mAP of 82.3%. Specifically, IDE attained lower performance, with a Rank-1 value of 73.2% and an mAP of 52.8%, indicating that as a basic PReID method, IDE struggles with achieving high matching accuracy across multiple candidate images. Gp-reID performed better, achieving a Rank-1 value of 85.2% and an mAP of 72.8%, demonstrating a notable improvement in its overall recognition effect, particularly when handling complex backgrounds and diverse scenes. MaskReID and SVDNet had mid-level performance on this dataset, with MaskReID reaching a Rank-1 value of 78.9% and an mAP of 61.9% and SVDNet achieving Rank-1 and mAP values of 76.7% and 56.8%, respectively. This indicates that while these methods showed some effectiveness in pedestrian recognition, they still face limitations when addressing complex scenarios. AACN, PAN, and

HA-CNN also exhibited moderate performance. PAN performed relatively well in terms of mAP (66.7%), exhibiting some advantage in the overall matching task, but its Rank-1 score (75.9%) was lower, suggesting that there is room for improvement with respect to identifying the most accurate matches. SPReID, MGN, and DSA-ReID performed remarkably well on this dataset. MGN led the other methods with a Rank-1 score of 88.7% and an mAP of 78.4%, highlighting its strong advantage when handling complex backgrounds and diverse features. DSA-ReID achieved a Rank-1 score of 86.2% and an mAP of 74.3%, demonstrating its adaptability and robustness in practical applications. Our proposed method achieved the highest scores for both the Rank-1 (91.4%) and mAP (82.3%) metrics, significantly surpassing the other methods. This indicates that our approach not only excels in terms of single-image matching accuracy but also demonstrates superior overall matching precision. It has proven to be highly reliable and accurate in handling complex PReID tasks.

Table 4. Performance comparison with other methods on the DukeMTMC-reID dataset.

Methods	**Rank-1**	**mAP**
IDE	73.2	52.8
Gp-reID	85.2	72.8
MaskReID	78.9	61.9
SVDNet	76.7	56.8
PAN	75.9	66.7
AACN	76.8	59.3
SPReID	84.4	71.1
HA-CNN	80.5	63.3
MGN	88.7	78.4
DSA-ReID	86.2	74.3
Ours	**91.4**	**82.3**

The experimental results obtained on the CUHK03 dataset are shown in Table 5. Compared to methods such as MGCAM, HA-CNN, Mancs, MGN, and DSA-ReID, our method yielded a significant Rank-1 improvement on this dataset. As shown in Table 5, the tested PReID algorithms generally performed slightly worse under the "detected" condition than under the "labeled" condition, with most algorithms producing lower Rank-1 and mAP values in the detected setting. This indicates that noise and errors present in automatically detected data can impact the recognition performance of the algorithms. However, advanced methods such as DSA-ReID and the proposed method maintained high accuracy and precision even under the detected condition. Notably, our method demonstrated strong potential in complex scenarios and automated applications. Under the labeled condition, our method achieved a Rank-1 score of 85.1% and an mAP of 80.4%. Under the detected condition, its Rank-1 and mAP were 82.4% and 77.2%, respectively. Although a slight performance decline was observed under the detected condition, our method excelled in both scenarios, significantly outperforming the other methods. This indicates that our method exhibits outstanding stability and precision when addressing automatically detected data.

To assess the contributions of the MAS and the RAM to the performance of the proposed model, we conducted ablation experiments. Specifically, we removed these modules and evaluated the performance achieved by the model on the Market-1501 dataset. The experimental results showed that after removing the MAS, the mAP of the model decreased from 90.1% to 85.2%, and its Rank-1 accuracy dropped from 97.2% to 93.4%. Similarly, after removing the RAM, the mAP decreased from 90.1% to 87.3%, and the Rank-1 accuracy dropped from 97.2% to 94.5%. These results highlight the importance of both modules for enhancing the performance of the developed model.

Table 5. Performance comparison with other methods on the CUHK03 dataset.

Method	**Labeled**		**Detected**	
	Rank-1	mAP	Rank-1	mAP
MGCAM	50.1	50.2	46.7	46.9
HA-CNN	44.4	41.0	41.7	38.6
Mancs	69.0	63.9	65.5	60.0
MGN	68.0	67.4	66.8	66.0
DSA-ReID	78.9	75.2	78.2	73.1
Ours	**85.1**	**80.4**	**82.4**	**77.2**

4. Conclusions

This paper proposes a PReID algorithm for completing PReID tasks based on an ODAM, focusing on two key aspects: multiscale detection and deep feature extraction. First, an MAS is introduced to address the limitations exhibited by the traditional methods when detecting small objects. This strategy captures detailed features in pedestrian images at various scales, significantly improving the detection accuracy achieved for small or partially occluded objects. Second, by integrating an RAM, the feature extraction network employs a ResNet structure, which not only extracts deeper features but also effectively explores the relationships between features, thereby enhancing the accuracy and robustness of PReID. Extensive experiments were conducted in comparison with the mainstream models developed in recent years, and the results showed that the proposed method performed exceptionally well in terms of both the Rank-1 and mAP metrics, surpassing most of the existing methods. These results validate the effectiveness of our approach, especially in terms of its stability and accuracy when handling complex scenes and automatically detected data.

Our method is not only applicable to static images but can also be extended to video-based PReID and object tracking tasks, demonstrating its generalizability to practical applications. Nevertheless, the proposed method has some limitations. First, the introduction of multiscale attention mechanisms and RAMs increases the complexity of the model, which in turn raises its demand for computational resources, making it less suitable for deployment in resource-constrained environments. Second, although the method performs well on datasets in a controlled laboratory setting, its robustness and generalizability in more complex real-world scenarios still require further validations. Moreover, while the method exhibits potential for use in video applications, further optimization is needed to address video-specific challenges such as motion blur, occlusion, and camera angle variations. In future work, we plan to optimize the model structure to reduce its computational complexity, making it more applicable to real-world scenarios. We also aim to validate and improve the robustness and generalizability of the model by using larger, real-world datasets. Additionally, we will explore the integration of dynamic feature extraction and cross-frame information fusion techniques for video-based PReID tasks to enhance the recognition capabilities of the model in video analysis scenarios.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This study was partially supported by the Qinglan Project of Jiangsu Universities, the Talent Development Project of Taizhou University (No. TZXY2018QDJJ006 and TZXY2019QDJJ006), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 22KJB520036 and 23KJB510033) and the Young Science and Technology Talent Support Project of Taizhou.

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. W. Sun, Q. Li, C. Zhao, S. K. Nguang, Mode-dependent dynamic output feedback h∞ control of networked systems with markovian jump delay via generalized integral inequalities, *Inf. Sci.*, **520** (2020), 105–116. https://doi.org/10.1016/j.ins.2020.02.023
2. F. Tung, J. S. Zelek, D. A. Clausi, Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance, *Image Vision Comput.*, **29** (2011), 230–240. https://doi.org/10.1016/j.imavis.2010.11.003
3. L. Zheng, Y. Yang, A. G. Hauptmann, Person re-identification: Past, present and future, preprint, arXiv:1610.02984. https://doi.org/10.48550/arXiv.1610.02984
4. M. Ye, C. Liang, Z. Wang, Q. Leng, J. Chen, J. Liu, Specific person retrieval via incomplete text description, in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, (2015), 547–550. https://doi.org/10.1145/2671188.2749347
5. M. Wang, B. Lai, J. Huang, X. Gong, X. S. Hua, Camera-aware proxies for unsupervised person re-identification, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021), 2764–2772. https://doi.org/10.1609/aaai.v35i4.16381
6. G. Zhang, Y. Ge, Z. Dong, H. Wang, Y. Zheng, S. Chen, Deep high-resolution representation learning for cross-resolution person re-identification, *IEEE Trans. Image Process.*, **30** (2021), 8913–8925. https://doi.org/10.1109/TIP.2021.3120054
7. G. Zhang, Z. Luo, Y. Chen, Y. Zheng, W. Lin, Illumination unification for person re-identification, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2022), 6766–6777. https://doi.org/10.1109/TCSVT.2022.3169422

8. X. Shu, G. Li, X. Wang, W. Ruan, Q. Tian, Semantic-guided pixel sampling for cloth-changing person re-identification, *IEEE Signal Process. Lett.*, **28** (2021), 365–1369. https://doi.org/10.1109/LSP.2021.3091924

9. E. Ning, C. Wang, H. Zhang, X. Ning, P. Tiwari, Occluded person re-identification with deep learning: A survey and perspectives, *Exp. Syst. Appl.*, **239** (2024), 122419. https://doi.org/10.1016/j.eswa.2023.122419

10. J. Miao, Y. Wu, Y. Yang, Identifying visible parts via pose estimation for occluded person re-identification, *IEEE Trans. Neural Networks Learn. Syst.*, **33** (2021), 4624–4634. https://doi.org/10.1109/TNNLS.2021.3059515

11. T. Si, F. He, P. Li, Y. Song, L. Fan, Diversity feature constraint based on heterogeneous data for unsupervised person re-identification, *Inf. Process. Manage.*, **60** (2023), 103304. https://doi.org/10.1016/j.ipm.2023.103304

12. M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (2010), 2360–2367. https://doi.org/10.1109/CVPR.2010.5539926

13. A. Bedagkar-Gala, S. K. Shah, A survey of approaches and trends in person re-identification, *Image Vision Comput.*, **32** (2014), 270–286. https://doi.org/10.1016/j.imavis.2014.02.001

14. W. S. Zheng, S. Gong, T. Xiang, Person re-identification by probabilistic relative distance comparison, in *CVPR 2011*, (2011), 649–656. https://doi.org/10.1109/CVPR.2011.5995598

15. L. Wu, C. Shen, A. Hengel, Personnet: Person re-identification with deep convolutional neural networks, preprint, arXiv:1601.07255. https://doi.org/10.48550/arXiv.1601.07255

16. X. Qian, Y. Fu, Y. G. Jiang, T. Xiang, X. Xue, Multi-scale deep learning architectures for person re-identification, in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), 5399–5408. https://doi.org/10.1109/ICCV.2017.577

17. 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 480–496. https://doi.org/10.1007/978-3-030-01225-0

18. X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, et al., Alignedreid: Surpassing human-level performance in person re-identification, preprint, arXiv:1711.08184. https://doi.org/10.48550/arXiv.1711.08184

19. C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven deep convolutional model for person re-identification, in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), 3960–3969. https://doi.org/10.1109/ICCV.2017.427

20. M. S. Nixon, J. N. Carter, Automatic recognition by gait, in *Proceedings of the IEEE*, **94** (2006), 2013–2024. https://doi.org/10.1109/JPROC.2006.886018

21. W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 2285–2294. https://doi.org/10.1109/CVPR.2018.00243

22. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 7132–7141. https://doi.org/10.1109/CVPR.2018.00745

23. C. Wang, Q. Zhang, C. Huang, W. Liu, X. Wang, Mancs: A multi-task attentional network with curriculum sampling for person re-identification, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 365–381.

24. C. P. Tay, S. Roy, K. H. Yap, Aanet: Attribute attention network for person re-identifications, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 7134–7143. https://doi.org/10.1109/CVPR.2019.00730

25. H. Li, N. Dong, Z. Yu, D. Tao, G. Qi, Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2021), 2814–2830. https://doi.org/10.1109/TCSVT.2021.3099943

26. S. Wang, R. Liu, H. Li, G. Qi, Z. Yu, Occluded person re-identification via defending against attacks from obstacles, *IEEE Trans. Inf. Forensics Secur.*, **18** (2022), 147–161. https://doi.org/10.1109/TIFS.2022.3218449

27. Y. Wang, G. Qi, S. Li, Y. Chai, H. Li, Body part-level domain alignment for domain-adaptive person re-identification with transformer framework, *IEEE Trans. Inf. Forensics Secur.*, **17** (2022), 3321–3334. https://doi.org/10.1109/TIFS.2022.3207893

28. D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, **3** (2007), 1–7.

29. W. Li, R. Zhao, X. Wang, Human reidentification with transferred metric learning, in *Computer Vision—ACCV 2012: 11th Asian Conference on Computer Vision*, (2013), 31–44. https://doi.org/10.1007/978-3-642-37331-2_3

30. W. Li, X. Wang, Locally aligned feature transforms across views, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2013), 3594–3601. https://doi.org/10.1109/CVPR.2013.461

31. W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2014), 152–159. https://doi.org/10.1109/CVPR.2014.27

32. L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in *Proceedings of the IEEE International Conference on Computer Vision*, (2015), 1116–1124. https://doi.org/10.1109/ICCV.2015.133

33. Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), 3754–3762. https://doi.org/10.1109/ICCV.2017.405

34. L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, (2018), 79–88. https://doi.org/10.1109/CVPR.2018.00016

35. M. Guo, E. Chou, D. A. Huang, S. Song, S. Yeung, F. F. Li, Neural graph matching networks for fewshot 3D action recognition, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 653–669.

36. J. Si, H. Zhang, C. G. Li, J. Kuen, X. Kong, A. C. Kot, et al., Dual attention matching network for context-aware feature sequence based person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 5363–5372. https://doi.org/10.1109/CVPR.2018.00562

37. Y. Sun, L. Zheng, W. Deng, S. Wang, Svdnet for pedestrian retrieval, in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), 3800–3808. https://doi.org/10.1109/ICCV.2017.410

38. L. Qi, J. Huo, L. Wang, Y. Shi, Y. Gao, Maskreid: A mask based deep ranking neural network for person re-identification, preprint, arXiv:1804.03864. https://doi.org/10.48550/arXiv.1804.03864

39. D. Chen, H. Li, T. Xiao, S. Yi, X. Wang, Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 1169–1178. https://doi.org/10.1109/CVPR.2018.00128

40. M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, M. Shah, Human semantic parsing for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 1062–1071. https://doi.org/10.1109/CVPR.2018.00117

41. J. Almazan, B. Gajic, N. Murray, D. Larlus, Re-id done right: Towards good practices for person re-identification, preprint, arXiv:1801.05339. https://doi.org/10.48550/arXiv.1801.05339

42. G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, in *Proceedings of the 26th ACM International Conference on Multimedia*, (2018), 274–282. https://doi.org/10.1145/3240508.3240552