Pattern Recognition 163 (2025) 111443

Contents lists available at ScienceDirect

Pattern Recognition

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

Adaptive transformer with Pyramid Fusion for cloth-changing Person Re-Identification

Guoqing Zhang [a,b], Jieqiong Zhou [a], Yuhui Zheng [a], Gaven Martin [c], Ruili Wang [b,d,e,*]

[a] School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, China
[b] School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand
[c] Institute for Advanced Study, Massey University, Auckland, New Zealand
[d] School of Computer Science, University of Nottingham, Ningbo, China
[e] School of Data Science and Artificial Intelligence , Wenzhou University of Technology, Wenzhou, China

ARTICLE INFO

Keywords:
Cloth changing
Person re-identification
Vision transformer

ABSTRACT

Recently, Transformer-based methods have made great progress in person re-identification (Re-ID), especially in handling identity changes in clothing-changing scenarios. Most current studies usually use biometric information-assisted methods such as human pose estimation to enhance the local perception ability of clothes-changing Re-ID. However, it is usually difficult for them to establish the connection between local biometric information and global identity semantics during training, resulting in the lack of local perception ability during the inference phase, which limits the improvement of model performance. In this paper, we propose a Transformer-based Adaptive-Aware Attention and Pyramid Fusion Network (A^3PFN) for CC Re-ID, which can capture and integrate multi-scale visual information to enhance recognition ability. Firstly, to improve the information utilization efficiency of the model in cloth-changing scenarios, we propose a Multi-Layer Dynamic Concentration module (MLDC) to evaluate the importance features at each layer in real time and reduce the computational overlap between related layers. Secondly, we propose a Local Pyramid Aggregation Module (LPAM) to extract multi-scale features, aiming to maintain global perceptual capability and focus on key local information. In this module, we also combine the Fast Fourier Transform (FFT) with self-attention mechanism to more effectively identify and analyze pedestrian gait and other structural details in the frequency domain and reduce the computational complexity of processing high-dimensional data in the self-attention mechanism. Finally, we build a new dataset incorporating diverse atmospheric conditions (for instance wind and rain) to more realistically simulate natural scenarios for the changing of clothes. Extensive experiments on multiple cloth-changing datasets clearly confirm the superior performance of A^3PFN. The dataset and related code are available on the website: https://github.com/jieqiongz1999/vcclothes-w-r.

1. Introduction

Person Re-Identification (Re-ID) strives to identify the same person across different cameras and plays a vital role in public safety. However, to date most person Re-ID methods [1–3] use clothing as discriminative information to deal with obstacles such as item occlusion and perspective changes. However, in a real-world scenario, such as criminal tracking, clothing change is a common evasion strategy, and traditional short-term Re-ID technology cannot effectively deal with this, as shown in Fig. 1. Therefore, it is important to study more targeted cloth-changing Person Re-ID methods.

The CC Re-ID task aims to extract identity information unaffected by clothing changes [4–6]. One category of approaches focus on identifying clothing-independent features, such as body outlines, posture key-points and gait information. For example, Yang et al. [7] proposed a network that can adapt to clothing changes around pedestrian silhouette sketches, but is affected by environmental factors such as lighting and occlusion, and may ignore key details such as facial features. The other category of approaches focus on separating identity and clothing features such as GAN and semantic-guided clothing erasure network [8]. However this usually brings challenges such as additional computational overhead, high computational requirements, and strong dependence on data quality.

* Corresponding author at: School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand.
E-mail addresses: guoqingzhang@nuist.edu.cn (G. Zhang), jieqiongz331@nuist.edu.cn (J. Zhou), zheng_yuhui@nuist.edu.cn (Y. Zheng), G.J.Martin@massey.ac.nz (G. Martin), ruili.wang@massey.ac.nz (R. Wang).

https://doi.org/10.1016/j.patcog.2025.111443
Received 16 July 2024; Received in revised form 15 January 2025; Accepted 5 February 2025
Available online 12 February 2025

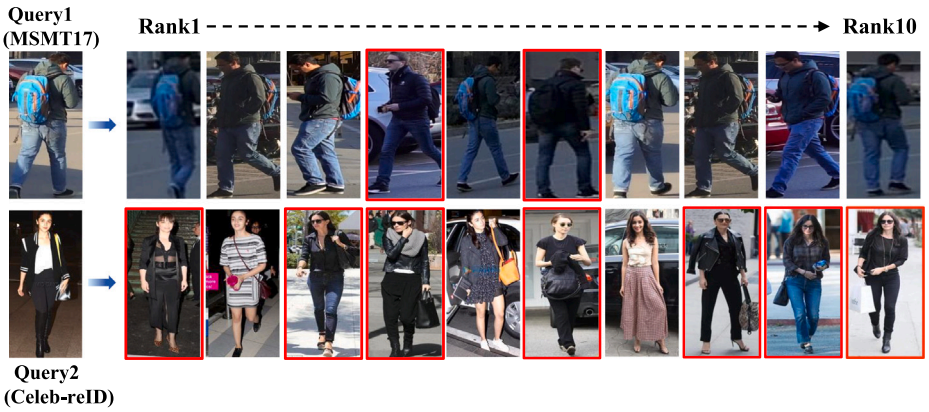

Fig. 1. Visualization of the Top-8 ranking lists generated by MGN [3] on the MSMT17 and Celeb-reID datasets. Images with red boxes indicate incorrect matches.

Recently, Vision Transformer [9] (ViT) has demonstrate remarkable performance in various computer vision tasks [10,11] with its multi-head self-attention mechanism to effectively capture inter dependencies within an image. However, a common limitation of the Transformer architecture is tendency to utilize the output of a specific layer for representation, often neglecting other valuable information embedded in the other layers. Fig. 2 shows the attention map visualization of the first three layers and the last three layers of the ViT model. It can be clearly observed that there are significant differences in the focus of attention in different layers. In addition, although existing methods often improve the local perception ability of pedestrian Re-ID after changing clothes through strategies such as posture estimation, but how to effectively coordinate local details with global semantic information remains a challenge.

To mitigate these limitations, we propose a Transformer based Adaptive-Aware Attention and Pyramid Fusion Network (A^3PFN) for CC Re-ID. We firstly design a Multi-Layer Dynamic Concentration module (MLDC) to integrate the characteristics of each layer of ViT and reduce the redundancy between layers. MLDC fuses different layer features through weighting and adjusts the importance of each layer in real time. Subsequently, recognizing that each layer of the ViT model concentrates on different aspects of the image, we propose a Local Pyramid Aggregation Module (LPAM) to extract multi-scale features, thereby maintaining attention to global perception and key local information. In this module, we also innovatively integrate a Fast Fourier Transform (FFT) into the self-attention mechanism to effectively identify subtle pedestrian differences in the frequency domain (such things as gait and clothing texture) to improve both computational efficiency and accuracy. Finally, since the existing Re-ID datasets do not consider the impact of weather, we propose the VC-Clothes-W&R dataset to fill this gap by introducing wind and rain elements.

Our primary contributions are the following:

- We propose a Transformer-based Adaptive Aware Attention and Pyramid Fusion Network for CC Re-ID;
- We integrate the Fast Fourier Transform into the self-attention mechanism to improve the model's ability of identifying pedestrian features in the frequency domain and optimize computing efficiency;
- We propose the VC-Clothes-W&R dataset, which fills the missing natural weather factors in existing pedestrian re-identification datasets by introducing wind and rain elements.

The remainder of the paper is organized as follow: Section 2 presents some related works and the details of our proposed framework are described in Section 3. Section 4 outlines the experimental setup and presents the results of extensive experiments on diverse datasets. Ablation studies are reviewed in Section 5, and Section 6 presents our conclusions.

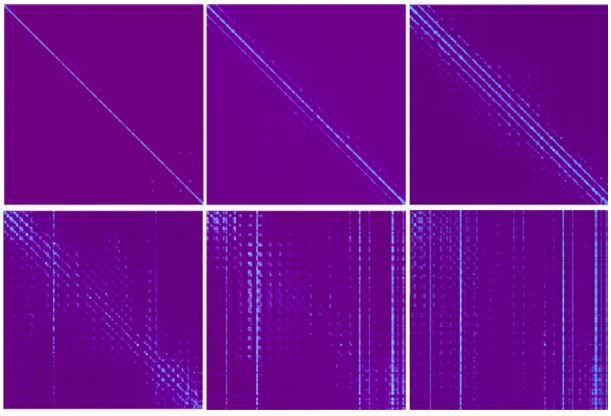

Fig. 2. Attention maps of the first three layers (first row) and the last three layers (second row) of ViT.

2. Related works

2.1. Classical person Re-ID

Current research has focused on solving problems such as lighting changes [12], occlusion [13], and cross resolution [14]. Jiang et al. [11] proposed a novel cross-modal Transformer (CMT) that jointly explores modal-level alignment modules and instance-level modules for visible-infrared person Re-ID, aiming to alleviate the loss of modality-specific information caused by existing methods integrating different modalities into a unified feature space. A Pose-guided Feature Decoupling (PFD) method proposed by Wang et al. [13] utilizes pose information to effectively decouple semantic components (such as human body or joint parts), and aligns unoccluded parts accordingly. Zhang et al. [14] proposed a Deep High-Resolution Pseudo-Siamese Framework (PS-HRNet), which introduces the VDSR-CA module to restore the resolution of low-resolution images and fully utilize the different channel information of feature maps, while using the new representation in HRNet to extract distinguishing features, thereby achieving excellent performance in cross-resolution scenarios. In addition, unsupervised Re-ID is also a key research focus: DHA [15] proposed an auto encoder-based method to generate deep latent attributes without extensive annotations, thus enhancing the ability to extract features from sparse but discriminative data to identify individuals within clues and reduce reliance on labeled data. IPES-GAN [16] adopts loop generation to adaptively balance environment and identity features to achieve domain adaptation, which significantly improves the robustness to environmental changes and camera settings in different domains.

2.2. Person Re-ID under intensive cloth variations

As public safety concerns become increasingly prominent, especially in the fields of monitoring and safety, there is a pressing need for effective identification of potential threats. Therefore, accurate identification of individuals who change their attire becomes crucial to promptly detect and intervene in potential security risks. These concerns have spurred many scholars to conduct in-depth research on CC Re-ID. In recent years, some related cloth-changing datasets have been released, such as VC-Clothes [17], Celeb-reID [18], Celeb-reID-light [19] and NKUP [20]. In these datasets, the same individual switches among multiple outfits, and wears various accessories, such as sunglasses, scarves, backpacks, etc. Frequent changes of clothing greatly reduce the reliability of traditional appearance-based matching methods.

To cope with the challenges brought by changing clothes, some works learn clothing-independent features with the help of identity-related auxiliary biological cues. For example, Hong et al. [4] proposed a shape–appearance mutual learning framework (FSAM), which is a dual-stream structure that acquires the detailed discriminative body shape information in shape stream and enriches the appearance stream with non-fabric-related details. Zhang et al. [21] proposed a novel Multi-Biometric Unified Network (MBUNet), which applies adaptive graph convolution to obtain relevant information between key points of the human body, and combines multiple biological features such as the person's head, neck, shoulders to mitigate the influence of clothing alterations. However, these methods have high requirements on image quality, and when the image is affected by occlusion, low illumination and so forth, this will limit the extraction of identity-related features, thus limiting the performance of the model. To further reduce the dependence on collecting a large amount of clothing change data, PosNeg [22] introduced an innovative data augmentation strategy, using positive augmentation and negative augmentation techniques to enrich the ID feature space and generate out-of-distribution synthetic samples, thereby enhancing the model's robustness to clothing changes.

Another very common methods seek to segregate clothing-related features from irrelevant features, enabling the model to concentrate on acquiring clothing-independent identity information. Xu et al. [8] proposed AFD-Net, which uses GAN and semantic perception models to distinguish the appearance and structural features of pedestrian images to achieve the separation of identity and clothing features, thereby enabling the model to learn identity Discriminating features. Similarly, SAVS [23] first locates the human body and clothing area according to the human body semantic segmentation, and introduces the human body semantic attention module to emphasize the human body information. Furthermore, it shields the clothing area to make the model focus on the extraction of visual semantic information unrelated to clothing. However, these kinds of methods generally face a challenge: in the process of separating clothing features from non-clothing features, distorted details are inevitably generated and the accurate expression of cloth-irrelevant features may be weakened, resulting in unstable training processes and poor model performance. Considering the limitation of the above two types of methods, we do not use any biological auxiliary branches or feature decoupling to help distinguish individuals, but make full use of the differences in features of each layer of Transformer to learn identity-related features. Specific introduction will be shown in the next section.

3. The proposed method

This section elaborates on our proposed approach. We first introduce a Multi-Layer Dynamical Concentration Module to evaluate the significance of features at each layer in real time while minimizing computational redundancy among highly correlated layers in Section 3.1. In Section 3.2, we further adopt a Local Pyramid Aggregation Module to enhance multi-scale features and integrate Fast Fourier Transform (FFT) to optimize the self-attention mechanism. Finally, the optimization of the overall framework is described in Section 3.3.

3.1. Multi-layer dynamical concentration module

In image processing, the Transformer architecture builds a visual feature hierarchy layer-by-layer, from edge and texture detection at the primary layer to scene comprehension at the high-level layer. However, previous Re-ID models often only focus on the information of the terminal layer, while ignoring the fine details of the primary and intermediate layers. To make up for this deficiency, we propose the Multi-Layer Dynamical Concentration Module (MLDC) (Fig. 3). This model dynamically synthesizes features across layers and also includes the key visual information from each layer in the final feature representation.

Calculation of weights. In order to effectively perform multi-layer feature fusion, in our method, we assign a weight coefficient w_i ($i = 1...12$) to each layer, the purpose of which is to evaluate the feature importance of each layer in real time and reduce the similarity redundancy of related layers, and the specific calculation process of w_i is as follows:

$$w_i = \frac{exp\left(f_i - \alpha \Sigma_{j=1,j\neq i}^{L} |\langle F_i, F_j \rangle|\right)}{\Sigma_{k=1}^{L} exp\left(f_k - \alpha \Sigma_{m=1,m\neq k}^{L} |\langle F_k, F_m \rangle|\right)}, \tag{1}$$

where $F_i \in \mathbb{R}^{N\times D}$ represents the output of the ith layer, N is the number of image blocks and D is the feature dimension of each token, $\langle\cdot,\cdot\rangle$ is the inner product, which measures the feature correlation of different layers, α is a regularization coefficient used to scale the impact of orthogonality constraints and reduce feature overlap between layers, L is the total number of layers. And f_i is a one-dimensional scalar that represents the importance of the output feature F_i of each layer, the specific calculation formula is as follows:

$$f_i = \frac{1}{h} \sum_{t=1}^{h} mean(A_{it}), \tag{2}$$

where h represents the number of attention heads in each layer, $mean(\cdot)$ represents the mean of all elements, $A_{it} \in \mathbb{R}^{N\times N}$ represents the attention score matrix of the tth head in the ith layer ($t, i = 1,2,\ldots,12$), and the formula is as follows:

$$A_{it} = \text{softmax}\left(\frac{Q_{it} K_{it}^{T}}{\sqrt{d_t}}\right), \tag{3}$$

where $Q_{it} \in \mathbb{R}^{N\times d_t}$ and $K_{it} \in \mathbb{R}^{N\times d_t}$ are the query and key matrices of the tth head in the ith layer, and $d_t = \frac{D}{h}$ is the dimension size of each head, which is used to scale the dot product result to prevent too large values from affecting the gradient of the $softmax(\cdot)$ function.

Enhanced Feature Fusion With Regularization. To mitigate the risk of model over-fitting that may occur due to the undue influence of specific layers, we incorporate an L^2 regularization term into our feature fusion formula:

$$F_{fusion} = \Sigma_{i=1}^{L} w_i F_i - \lambda \cdot \|W\|_F^2, \tag{4}$$

where λ is a non-negative regularization parameter employed to mitigate over-fitting by constraining the magnitude of the weights within the model. $\|W\|_F^2$ is the Frobenius norm of the weight matrix W and is the sum of the squares of all layer weights.

3.2. Local pyramid aggregation module

Although the adaptive weighted average method is very effective in integrating multi-layer features to achieve comprehensive representation, its global fusion method may ignore local information.

In order to capture image details at various scales, we design a local pyramid aggregation module (LPAM), as shown in Fig. 3. The module adopts a pyramid structure and gradually fuses the output of four different layers $F_{l_1}, F_{l_2}, F_{l_3}, F_{l_4}$ to achieve the extraction of different fine-grained information, the specific layer selection is provided in the ablation study section. For example, in the local feature

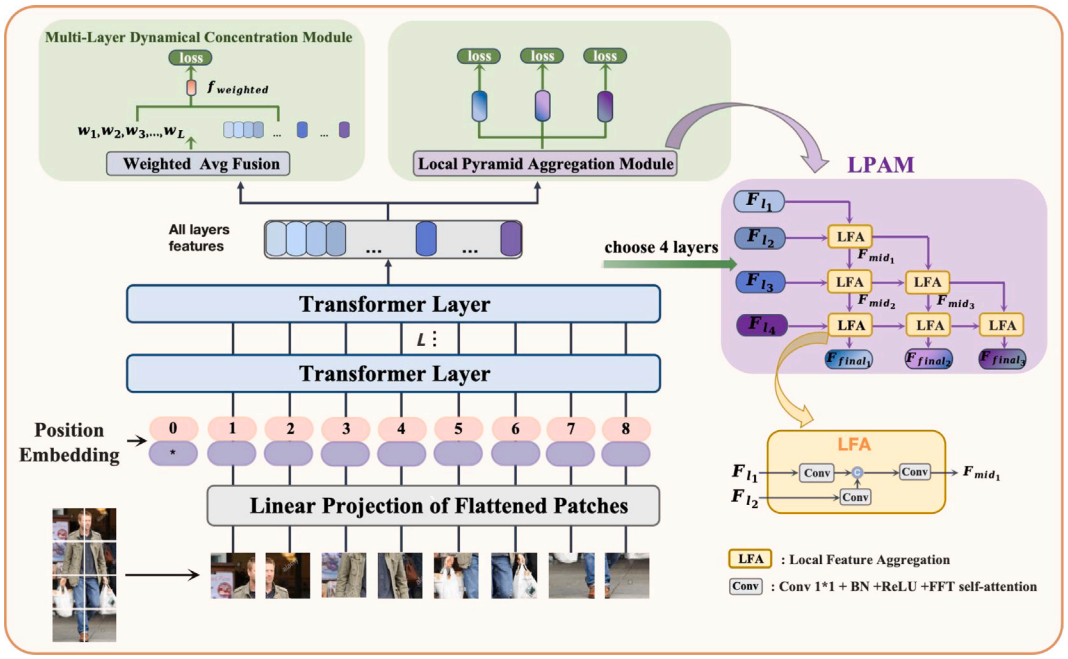

Fig. 3. The architecture of our proposed A^3PFN, which is built on ViT and contains two parallel modules–Multi-Layer Dynamical Concentration Module (MLDC) and Local Pyramid Aggregation Module (LPAM). MLDC aims to obtain aggregated global features by dynamically assigning weights to each layer. LPAM is designed to fuse multi-level features through a pyramid structure to obtain multi-scale information.

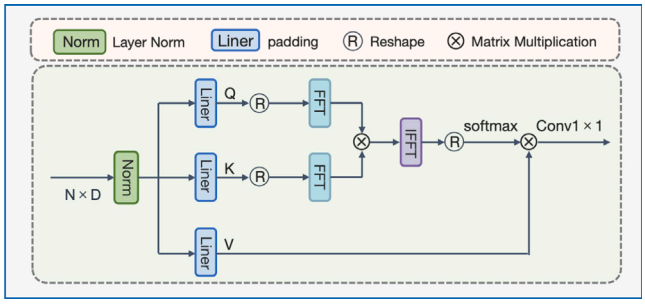

Fig. 4. Illustration of FFT Self-Attention, which aims to identify detailed information of pedestrians from a frequency domain perspective.

aggregation (LFA) of LPAM, we implement a 1×1 convolution layer and BatchNorm2D processing on features F_{l_1} and F_{l_2}, with the help of ReLU function for size adjustment and nonlinear enhancement. In addition, we introduce a self-attention mechanism to obtain enhanced pedestrian information from the frequency domain perspective (see Fig. 4). Finally, the convolved F_{l_1} and F_{l_2} are connected, and then input into the convolution block to achieve feature fusion. The formula for feature fusion is as follows:

$$F_{mid_1} = \varrho\left(\text{concat}\left(\varrho\left(F_{\ell_1}\right), \varrho\left(F_{\ell_2}\right)\right)\right), \tag{5}$$

where $\varrho(\cdot)$ represents the convolution block and $concat(\cdot,\cdot)$ refers to the splicing operation. The following fusion steps are similar to this. For the detailed process, please refer to the framework diagram of the local pyramid aggregation module (LPAM) shown in Fig. 3.

FFT Self-Attention. Fast Fourier Transform (FFT) is an effective algorithm for computing the Discrete Fourier Transform, as described in [24], which can significantly reduce the computational complexity from $O(N^2)$ to $O(N \log N)$, making it crucial in signal frequency domain analysis.

In our method, the self-attention module first accepts the input $X \in \mathbb{R}^{N \times D}$, where N is the number of image blocks (tokens) and D is the feature dimension of each token. Then, we perform different linear transformations on X to convert it into Q, $K$and V. Subsequently, Q, K and V are split into multiple heads. To improve the efficiency of Fast Fourier Transform (FFT), we appropriately fill the Q and K matrices to the integer power of 2, and then apply FFT on the filled Q_{padded} and K_{padded} and estimate their correlation in the frequency domain. The output is formulated as below:

$$Attn = Softmax(F^{-1}(F(Q_{padded}) \odot F(K_{padded}))[:,:,:,:Q.size(-1)]), \tag{6}$$

where $F(\cdot)$ and $F^{-1}(\cdot)$ represent FFT and inverse FFT (IFFT) respectively, $\odot$ is a dot product operation. Softmax function aims to normalize the result to produce attention weights $Attn$. Finally, we obtain the attention-weighted output through weighted calculation and residual connection.

3.3. Model optimization

We optimize the model through ID loss and triplet loss. The ID loss adopts the traditional cross-entropy loss function, excluding label smoothing, and its specific definition is as follows:

$$L_{ID} = -\sum_{j}^{C} y_j log(p_j), \tag{7}$$

where C is the number of categories, y_j is the one-hot encoding of the true label and p_j is the probability that the model predicts that the sample belongs to the jth category.

In triplet loss, we adopt a semi-hard sample mining strategy to solve the over-fitting problem caused by noisy samples in hard sample mining. This strategy selects samples that are predicted incorrectly but with low confidence during the training phase to improve the model's coverage of a broad sample set and mitigate the impact of noisy samples. The indicator function of semi-hard samples is as follows:

$$I_{\text{semi-hard}}(x_l^a, x_l^p, x_l^n) = \begin{cases} 1 & \text{if } d(ap) + m < d(an) < d(ap) + M \\ 0 & \text{otherwise}, \end{cases} \tag{8}$$

where $d(ap)$ represents the distance between the anchor sample x_l^a and the positive sample x_l^p, $d(an)$ represents the distance between the anchor sample and the negative sample x_l^n. m means the minimum distance difference between anchor samples and positive samples to

Table 1
Clothing change statistics of the long-term pedestrian datasets used in our experiments. Please note: 'SC' and 'CC' represent the two modes of same clothes and clothing change respectively.

Dataset	Source	Train(ID/Image)	Test(ID/Image)		Cameras	Data style	Weather backgrounds
			Query	Gallery			
VC-Clothes	Synthetic	256/9449	256/1020	256/8591	4	SC/CC	None
NKUP	Real	40/5336	39/332	67/4070	15	CC	None
Celeb-reID	Real	1052/34,186	420/2972	420/11,006	Many	CC	None
Celeb-reID-light	Real	100/887	100/934	590/10,842	Many	CC	None
PRCC	Real	150/17,896	71/3543	71/3384	3	SC/CC	None
LTCC	Real	77/9576	75/493	75/7050	12	SC/CC	None
VC-Clothes-W&R	Synthetic	256/9449	256/1020	256/8591	4	SC/CC	Wind & rain

prevent the model from only focusing on subtle differences between similar samples, M is the maximum distance between anchor samples and negative samples to avoid selecting overly simple negative samples and ensure that the model learns discriminative features. The triplet loss for semi-hard sample mining is defined as follows:

$$L_{tri} = -\sum_{l}^{N} I_{\text{semi-hard}}(x_l^a, x_l^p, x_l^n)\left(\|f(x_l^a) - f(x_l^p)\|_2^2 - \|f(x_l^a) - f(x_l^n)\|_2^2 + m\right)_+, \tag{9}$$

where $f(\cdot)$ denotes the feature extraction operator that maps the input image into an embedding space. $\|\cdot\|_2$ represents the L^2-norm, the Euclidean distance between two feature vectors. $(\cdot)_+$ is the hinge loss function, which means that the loss is calculated only when the value in the brackets is a positive number, otherwise the loss is 0. Consequently, the comprehensive loss function of the model is defined as follows:

$$L = \sum_{q}^{T} t_q(L_{ID_q} + L_{tri_q}) \quad (q = 1, \ldots, 4), \tag{10}$$

where $T = 4$ represents the total number of output features in our model and t_q represents the weight of each output feature. While assigning fixed weights to each part of the loss is simple, it may not produce the best model performance. Therefore, we adapt the loss function for each output feature to adaptive weights. Initially, the loss weight t_q for each output feature are expressed as the same size and are subsequently dynamically adjusted during training through back-propagation to obtain the optimal weights.

4. Experiments

4.1. Datasets and evaluation metrics

Datasets Details. To assess the performance of our proposed A^3PFN, we carry out experimental evaluations on the publicly available cloth-changing pedestrian datasets, including VC-Clothes [17], NKUP [20], Celeb-reID [18], Celeb-reID-light [19], LTCC [25] and PRCC [7]. Table 1 provides an overview of these datasets. Meanwhile, we build an enhanced dataset VC-Clothes-W&R for wind and rain scenes based on VC-Clothes to make up for the lack of environmental elements in existing datasets.

Evaluation Metrics. We evaluate the performance of A^3PFN using Rank-1 accuracy and mean precision (mAP) in three test scenarios: (1) general scenarios, covering clothing changes and consistent samples; (2) clothing change scenarios, only including clothing change samples; (3) clothing consistent scenes, only including consistent clothing samples. In the following tables, "sil", "ga", "dg", "pose" and "bs" represent pedestrian semantic segmentation, gait, data generation, Human Posture and body shape information.

Fig. 5. Some images from VC-Clothes-W&R.

4.2. VC-Clothes-W & R

As a synthetic dataset, VC-Clothes can provide a more controlled experimental environment, allowing us to keep other variables (such as lighting conditions and occlusions) relatively consistent across different scenes. In contrast, real image datasets often have the complexity of natural environments and introduce many uncontrollable variables, such as fluctuating lighting conditions, varying degrees of occlusion caused by unpredictable obstacles and other factors. These factors may mask the specific impact of weather on the recognition task, making it challenging to separate the variables required for testing. Therefore, to enhance the diversity of the image backgrounds and increase the model's robustness, we add wind and rain scenes to the VC-Clothes dataset [17]. The generated VC-Clothes-W&R dataset consists of images captured from four different cameras, segmented into training and testing sets. The training set comprises 256 unique identities with a total of 9449 images. Similarly, the test set is structured into query and gallery segments, maintaining the same 256 identities with 1020 images in the query and 8591 in the gallery set. Some sample images are shown in Fig. 5 and the construction process is as follows:

Considering the complexity of precipitation levels and changes in image brightness, we employ a refined atmospheric scattering model [26] to generate rainy and windy scenes for images. Specifically, we use $J(x, y)$ to represent a pixel in the original image and after adding the wind and rain scene, the corresponding pixel $I(x, y)$ can be calculated by the following formula:

$$I(x, y) = L[J(x, y)t(x, y) + R(x, y)(1 - t(x, y))M(x, y)], \tag{11}$$

where $L[\cdot]$ reflects the brightness coefficient, $R(x, y)$ represents the radiant brightness of raindrops on the spatial coordinates (x, y). Furthermore, $M(x, y)$ is the blur kernel used to simulate the width and diffusion properties of visual distortion caused by raindrops. $t(x, y)$ represents the medium transmission ratio at each point (x, y), which quantifies the light intensity due to the presence of raindrops at that specific location. The formula is as follows:

$$t(x, y) = e^{-\beta d(x,y)s(x,y)cos(\theta)}, \tag{12}$$

Table 2
Comparison with methods on VC-Clothes datasets (%).

Methods	Modality	Cross-clothes		Same-clothes	
		Rank-1	mAP	Rank-1	mAP
Short-term based methods					
PCB (ECCV 18) [27]	RGB	62.0	62.2	–	–
ISP (ECCV 20) [28]	RGB+sil	72.0	72.1	94.5	94.7
DG-Net (CVPR 19) [29]	RGB+dg	76.8	68.4	94.1	93.8
Cloth-changing based methods					
FSAM (CVPR 21) [4]	RGB+bs	78.6	78.9	94.7	94.8
MBUNet (TIP 22) [21]	RGB+pose	82.3	68.2	95.7	94.2
CAL (CVPR 22) [5]	RGB	81.4	81.7	95.1	95.3
ACID (TIP 23) [6]	RGB	84.3	74.2	95.1	94.7
AFL (TMM 23) [30]	RGB	82.5	83.0	–	–
DCR-ReID (TCSVT 23) [31]	RGB+bs	83.7	82.6	94.6	94.5
MGP (TMM 23) [32]	RGB+dg	81.8	81.7	94.7	94.9
Ours	RGB	**89.2**	**83.1**	**96.6**	**95.6**

where β is the attenuation coefficient of the medium rain, $d(x, y)$ is the depth of the medium at position (x, y), $s(x, y)$ denotes the density of rain at (x, y) and θ is the angle of light propagation relative to the viewer.

Next, we use the fuzzy kernel function $M(x, y)$ to simulate the scattering effect of raindrops on the imaging sensor and its effective radius, and quantitatively describes the optical diffusion characteristics caused by raindrops of different sizes. The kernel function is specifically defined as follows:

$$M(x, y) = \frac{1}{2} e^{-\frac{x^2+y^2}{2k^2}}, \tag{13}$$

where k represents the scale parameter of the raindrop scattering effect. By adjusting k, we can simulate the scattering distribution caused by raindrops of different sizes on the sensor, thus more accurately reflecting the impact of raindrops on image quality.

The entire model not only simulates raindrops, their direction and size, but also thoroughly considers the overall and local illumination variations under rainy conditions, forming highly realistic and diverse rainy scenarios.

4.3. Implementation details

Our method uses the Vision Transformer [9] (ViT-Base) as the baseline model, which has 12 Transformer layers and each containing 768-dimensional embeddings, and can effectively process image patches of 16 × 16 pixels. We resize the input images to [256, 128] to adapt to the model's processing requirements for patches. In addition, we apply data augmentation techniques including random horizontal flipping and random erasing, each with a probability of 0.5 to simulate visual changes in the real world. During training, we adopt a softmax triple sampling strategy with 4 instances per batch to enhance the generalization ability of the model and use 8 worker threads for data loading to improve data processing efficiency. The model is trained for 120 epochs by stochastic gradient descent (SGD), and the starting learning rate is set to 0.008. To optimize the learning rate adjustment process, we adopt the linear warm-up technique and set a learning rate decay factor of 1e−4. Every 10 epochs, we performed model evaluation using batches of 64 images to regularly monitor model performance.

4.4. Performance evaluations and comparisons

For the artificially synthesized datasets VC-Clothes, we compare the proposed A^3PFN with some short-term methods (i.e., PCB [27], ISP [28], DG-Net [29]) and cloth-changing Re-ID methods (i.e., FSAM [4], CAL [5], ACID [6], AFL [30], MGP [32]). In addition, we compare our method with five clothes-changing based methods (i.e., RCSAN [33], AFD-Net [34], MBUNet [21], SAFR [35], SAVS [23]) as well as some short-term methods on high-resolution datasets Celeb-reID and Celeb-reID-light collected by multiple cameras. Also, we compare with methods MVSE [36], UCAD [8] SAVS [23] based on clothing changes methods on the datasets NKUP collected in closed scenes. What is more, to further verify the generalization of our model, we also conduct experiments on VC-Clothes-W&R. More specifically, FSAM, MGP, MBUNet, SAVS, MVSE, and UCAD adopt multi-biological auxiliary modules to reduce the interference of variable appearance information; AFD-Net and SARF introduce generative adversarial networks to decouple identity-related and identity-independent features; other methods only use RGB modality to convey identity information that is not affected by clothing.

Results for VC-Clothes. We evaluate our proposed A^3PFN against seven methods based on cloth-changing and three short-term approaches on the VC-Clothes dataset, as illustrated in Table 2. We can notice that our method achieves the best results in both same-clothes and cross-clothes settings. Compared with FSAM [4], which transfers fine-grained body shape knowledge from the shape to appearance stream to enhance cloth-independent features, our method increases Rank-1 and mAP in cross-clothes scenarios by 10.6% and 4.2% respectively. This shows that the performance of FSAM is greatly affected by the quality of body shape information. Compared with ACID [6], a method that accumulates identity clues using a step-by-step competition strategy to accumulate identity clues, our method increases Rank-1 and mAP in cross-clothes scenarios by 4.9% and 8.9% respectively. It shows that our method effectively integrates fine-grained local information and global information, improving the information utilization efficiency in processing large-scale data. DCR-ReID [30] proposed a component reconstruction decoupling (CRD) module to separate clothing-related and unrelated features based on human body component reconstruction, but it is difficult to perform effectively when the image quality is low or there is occlusion. In addition, although its deep assembly decoupling (DAD) module enhances feature discrimination, its high computational requirements limit its application in resource-constrained environments. MBUNet [21] solves the problem of frequent clothing changes in clothing-changing scenes by utilizing biological cues that are not related to clothing (such as the head, neck, and shoulders), but these biological features may greatly reduce their effectiveness due to occlusion or angle changes between cameras. What is more, while the second best method AFL [30] achieves good performance, the construction of correlation factors requires to consider the differences and similarities of different identities, which greatly increases the complexity of the model. In contrast with AFL, our method not only achieves higher performance improvements in the cross-clothes scenarios with a 6.7% improvement in Rank-1 and a 0.1% improvement in mAP but also effectively reduces the cost of experimental settings. This shows that our method performs well in balancing high accuracy with lower computational and model complexity.

Results on Celeb-reID and Celeb-reID-light. We contrast our method with some recent short-term and cloth-changing Re-ID studies

Table 3
Comparison with SOTA methods on Celeb-reID and Celeb-reID-light (%).

Methods	Modality	Celeb-reID		Celeb-reID-light	
		Rank-1	mAP	Rank-1	mAP
Short-term based methods					
MGN (ACMMM 18) [3]	RGB	10.0	49.0	13.9	21.5
PCB (ECCV 18) [27]	RGB	8.2	37.1	9.0	16.7
DG-Net (CVPR 19) [29]	RGB+dg	50.1	10.6	23.5	12.6
Cloth-changing based methods					
RCSAN (ICCV 21) [33]	RGB	55.6	11.9	29.3	16.7
AFD-Net (IJCAL 21) [34]	RGB+GAN	52.1	10.6	22.2	11.3
MBUNet (TIP 22) [21]	RGB+pose	55.3	12.1	33.9	21.3
SAFR (TIP 22) [35]	RGB+GAN	56.0	14.2	29.5	16.7
DCR-ReID (TCSVT 23) [31]	RGB+bs	60.8	15.7	33.5	22.0
MGP (TMM 23) [32]	RGB+bs	60.5	16.1	32.8	21.5
SAVS (TNNLS 23) [23]	RGB+sil	**65.9**	**21.3**	–	–
Ours	RGB	61.4	16.9	**40.6**	**24.2**

Table 4
Comparison with SOTA methods on NKUP (%).

Methods	Modality	Rank-1	mAP
Short-term based methods			
PCB (ECCV 18) [27]	RGB	18.7	14.1
MGN (ACMMM 18) [3]	RGB	20.6	16.1
Cloth-changing based methods			
MVSE (ACMMM 21) [36]	RGB+sil	23.8	17.9
UCAD (IJCAI 22) [8]	RGB+bs	25.0	16.9
MBUNet (TIP 22) [21]	RGB+pose	24.5	17.7
DCR-ReID (TCSVT 23) [31]	RGB+bs	24.7	18.3
MGP (TMM 23) [32]	RGB+dg	25.1	18.0
SAVS (TNNLS 23) [23]	RGB+sil	25.3	18.6
Ours	RGB	**25.7**	**19.1**

on these two datasets, with the findings presented in Table 3. It is apparent that our method surpasses all the comparative methods on Celeb-reID-light. Relative to AFD-Net [34] and SARF [35], our Rank-1 accuracy shows an enhancement of 18.4% and 11.1%, and mAP increases by 12.9% and 7.5%, respectively. Since AFD-Net and SARF have in common that they decouple information by generating adversarial images and separate clothing regions through an additional human parsing model, which increases the computational overhead. In contrast, our method does not require the incorporation of additional biological branches, but makes full use of the information differences between different layers of ViT, effectively reducing the additional computational burden while improving the accuracy of the model.

From Table 3, it is seen that our method achieves the second highest performance on the Celeb-reID dataset and SAVS [23] obtains the best results. Specifically, ours is 4.5% and 4.4% lower than SAVS in Rank-1 and mAP, respectively. The main reason is that SAVS masks clothing clues through the visual masking module and reweights the visual feature map in the human semantic attention module to effectively utilize human semantic information. This method of combining clothing masking with biometrics provides new ideas for our future work.

Results on NKUP. Table 4 shows the comparison of our method with the short-term and CC Re-ID methods on the NKUP dataset. It can be seen that our methods significantly outperform all short-term methods, and gives the best performance. These three cloth-changing methods all use semantic segmentation or body shape as auxiliary cues to support the learning of clothing-independent features. But they all ignore the local structural cues of pedestrians. Relative to the second best method SAVS [23], our method improves Rank-1 and mAP by 0.4% and 0.5% respectively. This proves that our method successfully enhance the discrimination of human body parts by aggregating features at different scales.

Results for VC-Clothes-W&R. To further verify the generalization ability of our proposed method, we conduct comparative experiments on VC-Clothes-W&R. To the best of our knowledge, this is the first cloth-changing dataset that adds weather background. Table 5 presents the comparison results between our method with short-term and cloth-changing methods on this dataset, demonstrating that our method achieves the best performance. Compared with ACID [6], which achieves the second best performance, our method improves Rank-1 and mAP by 3.6% and 5.2% respectively in the cross-clothes setting. The main reason is that ACID gradually accumulates ID clues through global, channel, and pixel-level feature extraction, which may leads to feature redundancy and slow inference speed in large-scale data processing. However, as shown in the comparison results in Table 2, after introducing wind and rain scenarios, both our method and the comparison methods experience a decline in experimental accuracy, and our model does not achieve the minimal decrease. Therefore, further work is needed to address the impact of simulated weather conditions, such as incorporating adaptive image restoration techniques or introducing multi-modal data augmentation strategies to enhance the model's performance and robustness in weather environments.

Results for PRCC and LTCC. We compare with the state-of-the-art methods on the PRCC and LTCC datasets, the results are shown in Table 6. In cross-clothes settings, compared with the second-best method MBUNet [21], our method improves Rank-1 by 1.5% and 0.6% on PRCC and LTCC, and improves mAP by 3.4% and 2.7%. Compared with other methods, RCSAN [33] focuses on clothing state perception, but performs poorly under dynamic scene changes; FASM [4] enhances body shape and appearance features, but has limited effect when data is sparse or changes extremely; GIReID [37] relies on gait information from a single image, but its accuracy decreases when the gait is subtle or blurred; MBUNet focuses on extracting clothing-independent biometrics, but fails when occlusion is severe. Our model effectively improves the recognition accuracy and robustness under variable clothing conditions by integrating the dynamic information of each layer of the transformer and strengthening the capture of key local information. In the same-clothes setting, AIM [38] achieves the best results. It is an automatic intervention model based on causality, which simulates causal intervention through a dual-branch model and gradually separates clothing bias from entangled ID clothing representation without destroying semantic integrity. Compared with our method, AIM's Rank-1 is 0.2% and 4.2% higher on PRCC and LTCC, and mAP is 1.8% and 4.4% higher. In future work, we plan to draw on AIM's causal inference methods to enhance our model's ability to more effectively distinguish identity and clothing changes.

5. Ablation studies

Effectiveness of each module of our framework. To verify the effectiveness of our modules, we conducted experiments on datasets such as Celeb-reID, NKUP, Celeb-reID-light, VC-Clothes and VC-Clothes-W&R. The experimental results are shown in Table 7. We can see

Table 5
Comparison with SOTA methods on VC-Clothes-W&R (%).

Methods	Modality	Cross-clothes		Same-clothes	
		Rank-1	mAP	Rank-1	mAP
Short-term based methods					
ISP (ECCV 20) [28]	RGB+sil	66.3	63.1	92.2	91.4
DG-Net (CVPR 19) [29]	RGB+dg	70.7	65.9	91.8	90.2
Cloth-changing based methods					
MBUNet (TIP 22) [21]	RGB+pose	78.1	70.6	90.9	91.1
AFL (TMM 23) [30]	RGB	78.9	70.2	91.6	91.3
ACID (TIP 23) [6]	RGB	80.5	71.3	91.1	90.8
DCR-ReID (TCSVT 23) [31]	RGB+bs	80.1	72.0	91.9	91.4
MGP (TMM 23) [32]	RGB+dg	77.5	72.3	92.4	92.1
Ours	RGB	**84.1**	**76.5**	**93.1**	**92.7**

Table 6
Comparison with SOTA methods on PRCC and LTCC (%).

Methods	Modality	PRCC				LTCC			
		Cross-clothes		Same-clothes		Cross-clothes		Same-clothes	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Short-term based methods									
PCB (ECCV 18) [27]	RGB	22.9	–	86.9	–	23.5	10.0	61.8	27.5
ISP (ECCV 20) [28]	RGB+sil	36.6	–	92.8	–	27.8	11.9	66.3	29.6
MGN (ACMMM 18) [3]	RGB	53.5	53.3	98.2	98.4	25.0	12.6	68.4	34.6
Long-term based methods									
RCSAN (ICCV 21) [33]	RGB	50.2	48.6	100	97.2	–	–	–	–
GI-ReID (CVPR 22) [37]	RGB+ga	37.6	82.3	79.0	–	28.1	13.2	73.6	36.1
FSAM (CVPR 21) [4]	RGB+bs	54.5	–	98.8	–	38.5	16.2	–	–
MBUNet (TIP 22) [21]	RGB+pose	67.6	65.3	**100**	99.6	39.5	14.7	67.1	34.4
Chan et al (ACM 23) [39]	RGB+dg	65.8	61.2	99.5	96.7	32.9	15.3	73.4	36.8
AIM (CVPR23) [38]	RGB	54.7	55.0	**100**	**99.9**	38.3	17.0	**76.1**	**39.1**
Ours	RGB	**69.1**	**68.7**	99.8	98.1	**40.1**	**17.4**	71.9	34.7

Table 7
Ablation study of components in our framework in Celeb-reID, NKUP, VC-Clothes (cross clothes), VC-Clothes-W&R (cross clothes), and Celeb-reID-light (%).

Methods	Celeb-reID		NKUP		VC-Clothes		Celeb-reID-light		VC-Clothes-W&R	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Baseline	57.3	15.6	21.5	15.5	83.7	80.6	29.1	20.3	80.3	74.4
+MLDC	59.1	16.3	22.5	16.9	85.3	81.9	33.4	21.7	81.9	75.2
+LPAM	60.2	15.9	23.3	17.7	84.8	80.3	32.7	21.2	81.5	74.9
+MLDC+LPAM	60.9	16.8	24.2	18.3	87.1	82.3	36.1	23.9	83.1	75.8
+MLDC+LPAM (FFT)	**61.4**	**16.9**	**25.7**	**19.1**	**89.2**	**83.1**	**40.6**	**24.2**	**84.1**	**76.5**

that after the introduction of the MLDC module, the performance of all datasets has improved, especially on Celeb-reID-light, where the accuracies of Rank-1 and mAP have increased by 4.3% and 1.4% respectively. This shows that the MLDC module effectively improves the model's adaptability to clothing changes by dynamically fusing the multi-layer information of Transformer. When the proposed LPAM is added alone, the performance of the baseline on most datasets is further improved, especially the Rank-1 on NKUP and Celeb-reID-light are increased by 1.8% and 3.6% respectively, highlighting the role of the local pyramid aggregation module in extracting multi-scale features and capturing the key role of local information.

When MLDC and LPAM are used simultaneously, the model's performance of all datasets is significantly improved, especially on Celeb-reID-light, where the accuracies of Rank-1 and mAP are increased by 7.0% and 3.6% respectively. These results show that the simultaneous use of multi-layer information fusion and multi-scale information can further resist the interference caused by clothing changes. Next, to further verify the superiority of the FFT self-attention mechanism we proposed, we compare the performance of the attention mechanism combined with FFT and the ordinary attention mechanism in the local feature aggregation of the LPAM module. As can be seen from this Table, after the introduction of FFT, the performance of the model on all datasets has been further improved, especially on the VC-Clothes-W&R dataset, where Rank-1 and mAP are increased by 1.0% and 0.7% respectively. This fully demonstrates that FFT can help the model more effectively process and identify structural details in the frequency domain by converting data to the frequency domain. These results fully validate the effectiveness and importance of each component in our framework.

To intuitively demonstrate the effectiveness of each module, we visualize the experimental results on the Celeb-reID dataset, as illustrated in Fig. 6. It can be seen that the matching results of using modules MLDC and LPAM separately are better than the baseline model. When these two modules are used in combination, its matching rate in Ranks 1–10 is significantly improved. Especially after the introduction of FFT self-attention, the accuracy of the model is further improved, which is completely consistent with our ablation experimental results.

Visualization of feature distribution. To substantiate the efficacy of the introduced components, we employ t-SNE [40] for visualizing the distribution of features extracted by the model across different components, as illustrated in Fig. 7. With this figure, the circles mean randomly selected image features from the training set of Celeb-reID dataset, with varying colors denoting distinct identities.

Specifically, Fig. 7(a) shows the extraction distribution of features by the baseline model ViT. It can be seen that the feature points are relatively scattered and the degree of identity aggregation is low, which reflects the significant challenge to the effective recognition of ViT

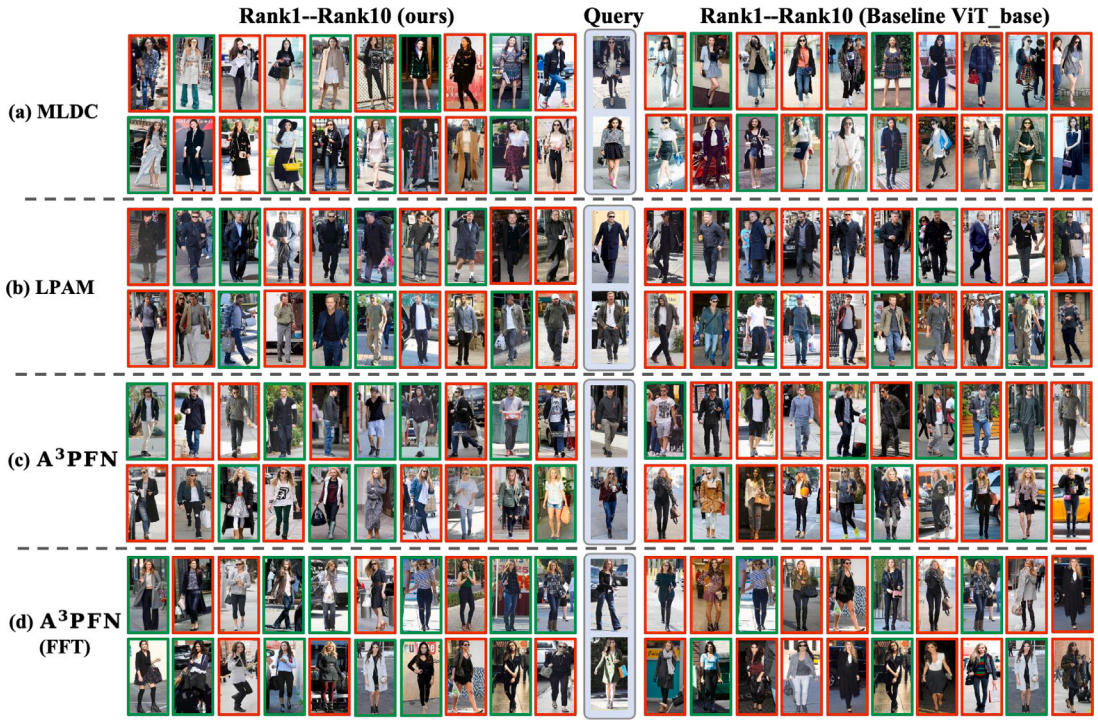

Fig. 6. Qualitative visualization of the baseline and our modules and combinations on the Celeb-reID dataset. Note that the green boxes highlight the correct results and the red boxes the incorrect results.

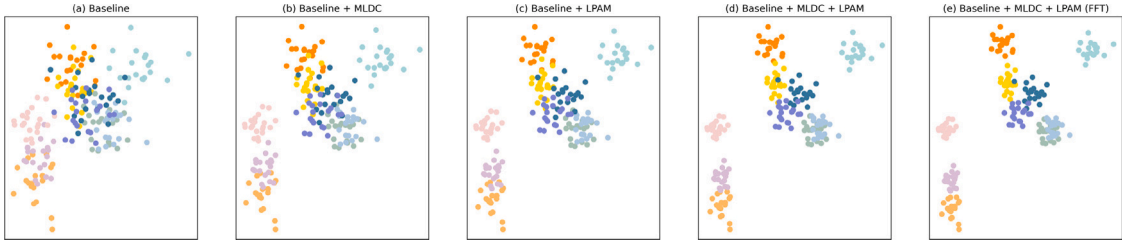

Fig. 7. Visual analysis of feature distribution on the Celeb-reID-light dataset. Circles represent sample features and different colors represent different identities. (a) The feature distribution of the baseline model, (b) the feature distribution of the baseline model after adding the MLDC module, (c) the feature distribution of the baseline model after adding the LPAM modules, (d) the feature distribution of the baseline model after adding the MLDC and LPAM modules, and (e) the feature distribution of the baseline model after adding the MLDC and LPAM (FFT) modules.

models by pedestrians changing clothes. Fig. 7(b) shows the feature distribution after introducing the MLDC module. This module significantly enhances the capture of identity information by weighted fusion of Transformer's multi-layer features, thereby making the boundaries between classes more obvious. Fig. 7(c) shows the feature distribution after introducing the LPAM module in the baseline model. By focusing on key local information and extracting multi-scale features, this module significantly improves the model's ability to capture and perceive local details, making the feature distribution more aggregated compared to the baseline. Fig. 7(d) shows the feature distribution after fusing the MLDC and LPAM modules at the same time, in which the feature point aggregation effect is better, which shows that the combination of these two modules not only maintains global perception capabilities, but also significantly enhances the recognition of details. and processing, demonstrating superior overall performance. Fig. 7(e) combines Fourier transform (FFT) on the basis of (d), which is the complete method A^3PFN we proposed, and its feature aggregation effect is the most outstanding. This fully illustrates the potential of FFT in helping models more effectively process and identify structural details in the frequency domain. In summary, our proposed component demonstrates its excellent performance.

Transformer layers selection of LPAM. The ViT-Base [9] model we selected contains 12 Transformer layers. Since the lower layers of the Transformer mainly focus on detail features such as edges, colors, and textures; the middle layers gradually turn to the local structure of the image; and the higher layers further focus on the global semantic understanding of the image. To enable the LPAM module to obtain comprehensive feature integration, we empirically selected 3rd, 6th, 9th, and 12th layers as inputs. To demonstrate the effectiveness of selecting these number of layers, we conduct partial experiments on the Celeb-reID, NKUP, and VC-Clothes (cross-clothing) datasets. The results are shown in Table 8 and we can see that our selected layers 3, 6, 9, and 12 achieve the best performance on all three datasets. This combination covers various feature stages from elementary to advanced, provides comprehensive feature integration for the LPAM module, and highlights the importance of considering local to global information in feature fusion.

Effectiveness of adaptive weights for loss functions. In order to explore the specific effects of adaptive weighting and fixed weight settings on the performance of the loss function, we implement five different fixed weight schemes on the NKUP dataset, including uniformly distributing weights t_1 to t_4 and letting t_1 to t_4 bear the maximum weights respectively. Through these representative experimental settings, we aim to evaluate the differences in the impact of different manual weight assignments on model performance. From Table 9, it is

Table 8
Comparison of ablation results on different Transformer layer selection.

Selected layers	Celeb-reID		NKUP		VC-Clothes	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
1 2 3 4	57.1	15.1	23.1	17.1	87.1	81.2
5 6 7 8	59.3	16.0	24.4	17.9	87.7	81.6
9 10 11 12	58.4	15.7	24.1	17.2	87.3	81.7
2 4 6 8	60.2	16.6	25.2	18.8	88.4	82.9
3 6 9 12	**61.4**	**16.9**	**25.7**	**19.1**	**89.2**	**83.1**

Table 9
Ablation study of manual weighting and adaptive weighting.

Weights				NKUP		
t_1	t_2	t_3	t_4	Rank-1	Rank-5	mAP
0.25	0.25	0.25	0.25	21.5	31.2	17.5
0.7	0.1	0.1	0.1	22.7	33.3	18.3
0.1	0.7	0.1	0.1	24.4	34.3	18.1
0.1	0.1	0.7	0.1	23.9	33.1	17.9
0.1	0.1	0.1	0.7	23.3	33.5	18.5
Ours (adaptive weights)				**25.7**	**35.1**	**19.1**

evident that the adaptive weights outperform the best manual weights by 1.3%, 1.6% and 0.6% on Rank-1, Rank-5 and mAP respectively. This shows that manual weight setting requires multiple debugging to determine better weight distribution, which consumes a lot of time and presents challenges in identifying the optimal weight. In contrast, experimental results fully demonstrate that adaptive weighting has a stronger ability to adapt to data and task differences.

6. Conclusion

In this paper we have proposed a Transformer-based Adaptive-Aware Attention and Pyramid Fusion Network for CC Re-ID. Our method utilizes a Multi-layer Dynamic Concentration Module to evaluate the importance of features at different levels in real time, effectively reducing computational redundancy and improving accuracy. In addition, our proposed Local Pyramid Aggregation Module optimizes the extraction process of multi-scale features, focusing on critical local information while maintaining global awareness capability. We also combine the Fast Fourier transform with a self-attention mechanism, aiming to enhance the ability to recognize fine pedestrian details. Finally, we add wind and rain scenes to the existing dataset to fill the lack of complex weather in existing pedestrian datasets.

Despite the good progress, our model still has some limitations in real-world environments: we find that the drop in accuracy of our method in rainy and windy scenes has no obvious advantage over other methods. In addition, frequent occlusion and complex lighting conditions may also hinder the accurate extraction of pedestrian identity features. Therefore, our future research will further integrate environmental factors such as occlusion and low light, and introduce more modal information such as depth map, thermal imaging or visual–inertial sensor data to help the model enhance its adaptability and robustness in various complex real-world scenarios.

CRediT authorship contribution statement

Guoqing Zhang: Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Funding acquisition, Conceptualization. **Jieqiong Zhou:** Writing – original draft, Visualization, Validation. **Yuhui Zheng:** Supervision, Funding acquisition. **Gaven Martin:** Writing – review & editing, Supervision. **Ruili Wang:** Writing – review & editing, Supervision.

Ethics approval

I have read and have abided by the statement of ethical standards for manuscripts submitted to the Journal of Pattern Recognition.

Declaration of competing interest

The authors confirm that 1. The work described is not under consideration for publication elsewhere; 2. All the necessary files have been uploaded by online; 3. Each author has participated sufficiently; 4. All the authors listed have approved the manuscript that is enclosed.

Acknowledgments

This research is supported by the National Natural Science Foundation of China under Grants 62172231, 92470202 and U22B2056, the Natural Science Foundation of Jiangsu Province, China under Grant BK20220107, the Preliminary Research Project on Leading Technologies by Wuxi Industrial Innovation Research Institute-Visual Intelligent Analysis of Worker Behavior and Anomaly Warning, Wenzhou Key Scientific and Technological Projects (No. ZG2024012), and the Ministry of Business Innovation and Employment 2020 Catalyst: Strategic – New Zealand-Singapore Data Science Research Programme Fund (grant number MAUX2002), New Zealand.

Data availability

Data will be made available on request.

References

[1] Y. Chen, H. Wang, X. Sun, B. Fan, C. Tang, H. Zeng, Deep attention aware feature learning for person re-identification, Pattern Recognit. 126 (2022) 108567.
[2] Y. Lu, W. Deng, Transferring discriminative knowledge via connective momentum clustering on person re-identification, Pattern Recognit. 126 (2022) 108569.
[3] 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, pp. 274–282.
[4] P. Hong, T. Wu, A. Wu, X. Han, W.-S. Zheng, Fine-grained shape-appearance mutual learning for cloth-changing person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10513–10522.
[5] X. Gu, H. Chang, B. Ma, S. Bai, S. Shan, X. Chen, Clothes-changing person re-identification with rgb modality only, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1060–1069.
[6] Z. Yang, X. Zhong, Z. Zhong, H. Liu, Z. Wang, S. Satoh, Win-win by competition: Auxiliary-free cloth-changing person re-identification, IEEE Trans. Image Process. (2023).
[7] Q. Yang, A. Wu, W.-S. Zheng, Person re-identification by contour sketch under moderate clothing change, IEEE Trans. Pattern Anal. Mach. Intell. 43 (6) (2019) 2029–2046.
[8] Y. Yan, H. Yu, S. Li, Z. Lu, J. He, H. Zhang, R. Wang, Weakening the influence of clothing: Universal clothing attribute disentanglement for person re-identification, in: IJCAI, 2022, pp. 1523–1529.
[9] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[10] G. Li, T. Zhao, Efficient image analysis with triple attention vision transformer, Pattern Recognit. (2024) 110357.
[11] K. Jiang, T. Zhang, X. Liu, B. Qian, Y. Zhang, F. Wu, Cross-modality transformer for visible-infrared person re-identification, in: European Conference on Computer Vision, Springer, 2022, pp. 480–496.
[12] P.K. Sarker, Q. Zhao, Enhanced visible–infrared person re-identification based on cross-attention multiscale residual vision transformer, Pattern Recognit. 149 (2024) 110288.
[13] T. Wang, H. Liu, P. Song, T. Guo, W. Shi, Pose-guided feature disentangling for occluded person re-identification based on transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 2540–2549.
[14] 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.
[15] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, S. Satoh, Learning sparse and identity-preserved hidden attributes for person re-identification, IEEE Trans. Image Process. 29 (2019) 2013–2025.
[16] A. Verma, A.V. Subramanyam, Z. Wang, S. Satoh, R.R. Shah, Unsupervised domain adaptation for person re-identification via individual-preserving and environmental-switching cyclic generation, IEEE Trans. Multimed. 25 (2021) 364–377.

[17] F. Wan, Y. Wu, X. Qian, Y. Chen, Y. Fu, When person re-identification meets changing clothes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 830–831.
[18] Y. Huang, J. Xu, Q. Wu, Y. Zhong, P. Zhang, Z. Zhang, Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification, IEEE TCSVT 30 (10) (2019) 3459–3471.
[19] Y. Huang, Q. Wu, J. Xu, Y. Zhong, Celebrities-reid: A benchmark for clothes variation in long-term person re-identification, in: 2019 IJCNN, IEEE, 2019, pp. 1–8.
[20] K. Wang, Z. Ma, S. Chen, J. Yang, K. Zhou, T. Li, A benchmark for clothes variation in person re-identification, Int. J. Intell. Syst. 35 (12) (2020) 1881–1898.
[21] G. Zhang, J. Liu, Y. Chen, Y. Zheng, H. Zhang, Multi-biometric unified network for cloth-changing person re-identification, IEEE Trans. Image Process. 32 (2023) 4555–4566.
[22] X. Jia, X. Zhong, M. Ye, W. Liu, W. Huang, Complementary data augmentation for cloth-changing person re-identification, IEEE Trans. Image Process. 31 (2022) 4227–4239.
[23] Z. Gao, H. Wei, W. Guan, J. Nie, M. Wang, S. Chen, A semantic-aware attention and visual shielding network for cloth-changing person re-identification, IEEE Trans. Neural Netw. Learn. Syst. (2023).
[24] H.J. Nussbaumer, H.J. Nussbaumer, The Fast Fourier Transform, Springer, 1982.
[25] X. Qian, W. Wang, L. Zhang, F. Zhu, Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, Long-term cloth-changing person re-identification, in: Proceedings of the Asian Conference on Computer Vision, 2020.
[26] E. Bruneton, F. Neyret, Precomputed atmospheric scattering, in: Computer Graphics Forum, Vol. 27, Wiley Online Library, 2008, pp. 1079–1086.
[27] 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, pp. 480–496.
[28] K. Zhu, H. Guo, Z. Liu, M. Tang, J. Wang, Identity-guided human semantic parsing for person re-identification, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer, 2020, pp. 346–363.
[29] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, J. Kautz, Joint discriminative and generative learning for person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2138–2147.
[30] Y. Liu, H. Ge, Z. Wang, Y. Hou, M. Zhao, Clothes-changing person re-identification via universal framework with association and forgetting learning, IEEE TMM (2023).
[31] Z. Cui, J. Zhou, Y. Peng, S. Zhang, Y. Wang, Dcr-reid: Deep component reconstruction for cloth-changing person re-identification, IEEE Trans. Circuits Syst. Video Technol. 33 (8) (2023) 4415–4428.
[32] Z. Zhao, B. Liu, Y. Lu, Q. Chu, N. Yu, C.W. Chen, Joint identity-aware mixstyle and graph-enhanced prototype for clothes-changing person re-identification, IEEE Trans. Multimed. (2023).
[33] Y. Huang, Q. Wu, J. Xu, Y. Zhong, Z. Zhang, Clothing status awareness for long-term person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11895–11904.
[34] W. Xu, H. Liu, W. Shi, Z. Miao, Z. Lu, F. Chen, Adversarial feature disentanglement for long-term person re-identification, in: IJCAI, 2021, pp. 1201–1207.
[35] S. Yang, B. Kang, Y. Lee, Sampling agnostic feature representation for long-term person re-identification, IEEE Trans. Image Process. 31 (2022) 6412–6423.
[36] Z. Gao, H. Wei, W. Guan, W. Nie, M. Liu, M. Wang, Multigranular visual-semantic embedding for cloth-changing person re-identification, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 3703–3711.
[37] X. Jin, T. He, K. Zheng, Z. Yin, X. Shen, Z. Huang, R. Feng, J. Huang, Z. Chen, X.-S. Hua, Cloth-changing person re-identification from a single image with gait prediction and regularization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14278–14287.
[38] Z. Yang, M. Lin, X. Zhong, Y. Wu, Z. Wang, Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1472–1481.
[39] P.P. Chan, X. Hu, H. Song, P. Peng, K. Chen, Learning disentangled features for person re-identification under clothes changing, ACM Trans. Multimed. Comput. Commun. Appl. 19 (6) (2023) 1–21.
[40] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).