ITC 3/54

Information Technology and Control
Vol. 54 / No. 3/ 2025
pp. 976-991
DOI 10.5755/j01.itc.54.3.40913

A Novel Employee Re-identification Method Based on Attention-free Capsule Network for Factory Surveillance Images

Received 2025/03/21 | Accepted after revision 2025/05/28

HOW TO CITE: Zhao, X., Zhang, N., Ding, W. (2025). A Novel Employee Re-identification Method Based on Attention-free Capsule Network for Factory Surveillance Images. *Information Technology and Control*, 54(3), 976-991. https://doi.org/10.5755/j01.itc.54.3.40913

A Novel Employee Re-identification Method Based on Attention-free Capsule Network for Factory Surveillance Images

Xinbo Zhao, Nan Zhang*

Liaoning University of International Business and Economics, School of Management, Dalian, China, 116052

Wei Ding

Dalian Polytechnic University, School of Fashion, Dalian, China, 116034

Corresponding author: zhangnan202406@126.com

Traditional methods have low recognition rate and poor robustness when dealing with complex factory employee monitoring images, so a new employee re recognition method based on Attention-free Capsule Network for factory monitoring images is proposed. Least Squares Generative Adversarial Network (LSGAN) is used to restore the factory monitoring image to repair the missing or damaged image caused by lighting, occlusion, noise, etc. Wavelet Contourlet transform is used to improve the details and clarity of images and the accuracy of subsequent staff re recognition. The mixed Gaussian model (GMM) is used to accurately segment the employee foreground in the image, and the segmented employee foreground image is input into the Attention-free Capsule Network. The feature is extracted through multi-layer convolution and pooling operations, and the dynamic routing mechanism is used to extract and aggregate employee identity features. After training, the employee identity tags are output to achieve efficient employee re recognition of factory monitoring images. The experimental results show that compared with visual attention, KISS+, and center and scale prediction methods, the proposed method introduces a novel approach for employee re identification in factory monitoring images. The proposed method demonstrates strong anti-interference and adaptability. In the comparison of key indicators, the proposed method achieved a recognition accuracy of 95.8 on Rank-1, 98.4 on Rank-5, and 99.5 on Rank-10. This series of numerical comparison results fully demonstrates that the proposed method has strong image processing capabilities, as well as high recognition accuracy and robustness.

KEYWORDS: Least Squares Generative Adversarial Network, Wavelet Contourlet transform, Image preprocessing, Gaussian mixture model, Foreground segmentation, Attention-free Capsule Network

1. Introduction

The re-recognition of factory surveillance image is a key and challenging research field in intelligent surveillance system. With the continuous development of industrial automation and intelligence, the demand of factory monitoring system for employee behavior analysis, safety monitoring and automatic management is increasing. As an important part, employee re identification technology aims to accurately identify specific employees from monitoring images and provide strong support for factory management. Traditional employee re recognition methods mainly relies on artificial feature design and classifier training [22]. These methods often have poor effects in complex monitoring scenes, and are vulnerable to changes in lighting, occlusion, posture and other factors.

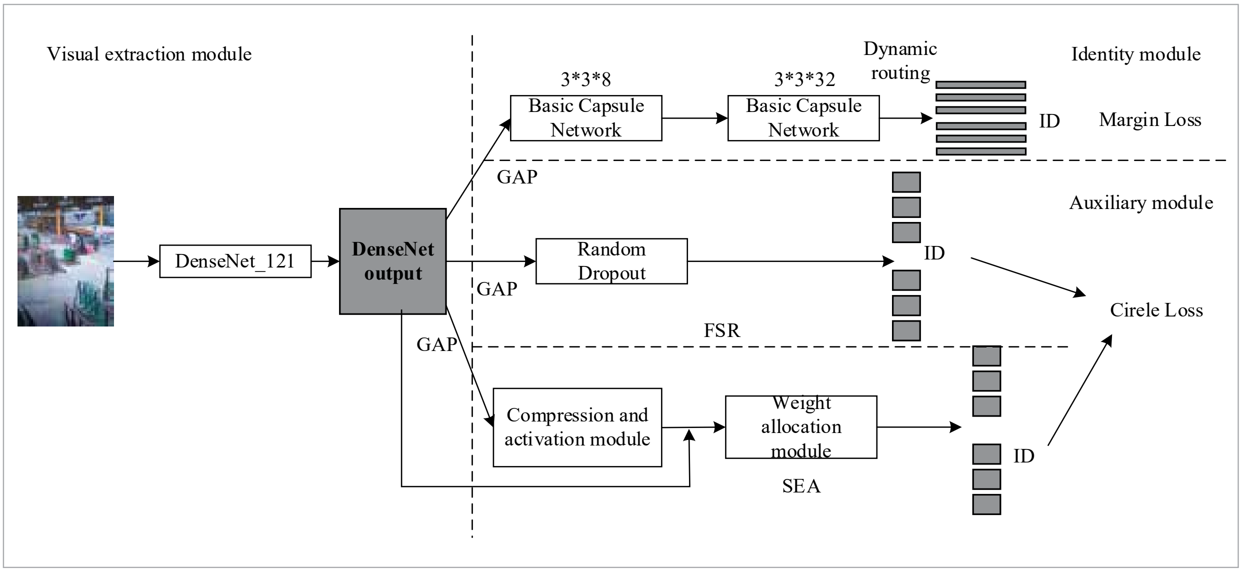
An et al. [1] proposed are recognition method for employees in factory surveillance images based on visual attention: Extracting salient region features of image through depth learning; Using the common attention mechanism to identify the key areas in the image pair; Combining global and local features to form joint features; Finally, the positive sample generation network is used to add training samples to complete the re identification of employees. This method is limited to the saliency detection of specific scenes, and is not robust to complex situations such as occlusion and light changes. Han et al. [5] introduced orthogonal basis vectors to generate virtual samples to enhance the stability of the covariance matrix, in which combined with multi feature fusion, discriminant features are extracted from surveillance images; High dimensional features are transformed into low dimensional expressions through dimension reduction and input into KISS+method to realize re recognition of factory monitoring images. Multi feature fusion may introduce redundant information, and the dimension reduction process may also lead to information loss, affecting the final recognition accuracy. Zhang et al. [21] proposed F-CSP detector that works as follows: Capture the video stream in real time through the factory's surveillance camera and decompose it into a single frame image; The feature pyramid network (FPN) in the F-CSP detector is used to reduce the number of channels in the feature map of each image frame; The balanced feature pyramid (BFP) is used to fuse these multi-scale feature maps into a more comprehensive feature map to capture the characteristics of employees at different scales; Compare these feature maps with the features in the known employee feature database to realize real-time employee re recognition. This method has poor recognition effect for employees with extreme scale changes or posture changes, and is not sensitive to complex background interference. Based on the above analysis, the challenges and shortcomings faced by existing work include: limited saliency detection in specific scenarios, insufficient robustness to occlusion and lighting changes; Information loss may occur due to redundant information or dimensionality reduction, affecting recognition accuracy; Poor adaptability to extreme scale changes, attitude changes, and complex backgrounds. Overall, the generalization ability, feature discriminability, and real-time performance of existing methods in complex scenarios still need to be further improved.

Therefore, we propose an employee re-identification method based on Attention-free Capsule Network for factory surveillance images (as shown in Figure 1), with the following main content:

1 The LSGAN is used to restore the original collected factory monitoring image, and repair the missing or damaged image information caused by lighting, occlusion, noise and other factors.
2 Wavelet Contourlet transform is used to enhance the details and clarity of the monitoring image, so as to improve the recognition accuracy of staff recognition.
3 The mixed Gaussian model (GMM) is used to accurately segment the employee foreground in the factory monitoring image after pretreatment.
4 The segmented image of the employee's foreground region is used as the input of the Attention-free Capsule Network. The network extracts the features of the image through multi-layer convolution, pooling and other operations, and uses the dynamic routing mechanism in the capsule network to implement aggregation and classification of the features. Finally, the network outputs the employee identity label corresponding to each input image.
5 Experimentation and discussion.
6 Concluding remarks.

Figure 1

Framework for employee re identification method in factory monitoring images.



2. Factory Surveillance Image Preprocessing

In practice, factory surveillance images may not be clear enough due to various reasons (e.g., camera quality, lighting conditions, compression during transmission, etc.), or there are missing, noise and other problems. To address these issues, the process is to first restore the image, and then implement enhancement, denoising and other processes to improve the accuracy of the subsequent re-identification of employees.

2.1. Image Restoration

The purpose of image restoration is to restore the information lost in the acquisition or transmission of factory monitoring image as much as possible, so that the image is closer to the original state. Take the least square methods to generate countermeasure network (LSGAN) [2, 10]. The schematic diagram of the generative adversarial network structure is shown in Figure 2.

In Figure 2, $G(z,\vartheta_g)$ is generator with multi-layer perception; $D(x,\vartheta_d)$ is a discriminator with a classification function. ϑ_g,ϑ_d is the parameter within the generator and discriminator, respectively. The steps are as follows:

1 **Build LSGAN model**

Generator: Design a generator network that can input original factory surveillance images (damaged, blurred) captured by different angles and cameras and generate restored images, the structure of the generator is adjusted according to the specific task and data set.

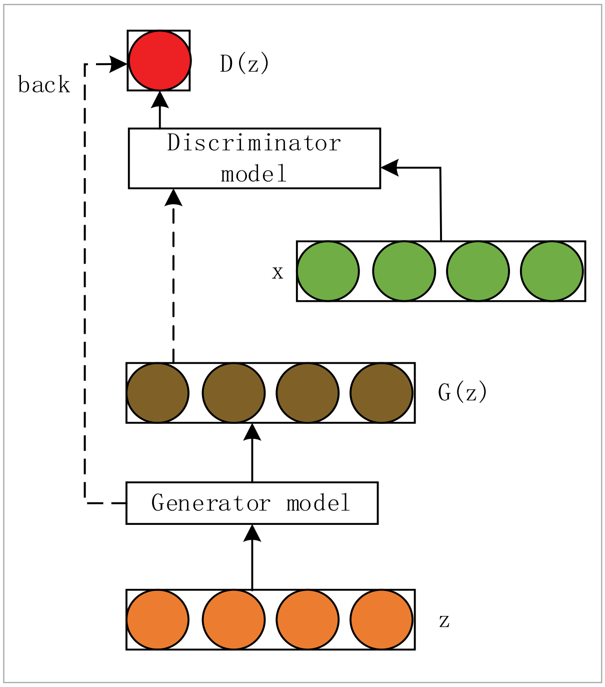
Discriminator: A discriminator network is constructed to distinguish between the generated image and the real image. The discriminator is a binary classifier that can output the probability that the input image is a real image.

2 **Define the loss function that**

The least squares loss function [17] is used as the loss function of the generator. When the discriminator is applied to the generated image, the $G(z)$ the discriminant value is close to the set generating graph labeling a, the loss function is minimized. The generator loss function aims to guide the generator to generate more realistic images, making it difficult for the discriminator to distinguish between generated and real images. When the loss function reaches its minimum value, it means that the image generated by the generator has higher quality and less difference from the real image. It is calculated as follows:

Figure 2

Schematic diagram of the structure of the generative adversarial network.



$$\min_D V(D)=\frac{1}{2}R_{x\sim p_{data}(x)}\left[\left(D(x)-b\right)^2\right]. \\ +\frac{1}{2}R_{z\sim Pz(z)}\left[\left(D(G(Z))-a\right)^2\right], \tag{1}$$

where a is label for generating graph. b is label for real diagrams. c is label for the generator expects the discriminator to discriminate against the generated graph. Again, a least squares loss function is used as the loss function for the discriminator. The function of the discriminator loss is to enable the discriminator to accurately distinguish between real images and generated images. It prompts the discriminator to approach the set real image label for the discrimination value of the real image, while approaching the set generator expected label for the discrimination value of the generated image. The formula is as follows:

$$\min_G V(G)=\frac{1}{2}R_{z\sim Pz(z)}\left[\left(D(G(Z))-c\right)^2\right], \tag{2}$$

3 Training LSGAN model

The alternating training law [8] is followed during the training process so that the generator and the discriminator are trained alternately: fixing the generator and updating the parameters of the discriminator to maximize its loss function (i.e., to better distinguish between the real and the generated image); fixing the discriminator and updating the parameters of the generator to minimize its loss function (i.e., to generate the more indistinguishable image), computed by the following formula:

$$\min_G \max_D V(D,G)=R_{z\sim P_{data}(x)}\left[\lg D(x)\right] \\ +R_{z\sim P_z(z)}\left[\lg\left(1-D\left(G(x)\right)\right)\right]. \tag{3}$$

When the generator is fixed, if the function $V(D,G)$ takes the maximum value, the following conditions must be satisfied:

$$D_G^*(x)=\frac{bP_{data}(x)+aP_g(z)}{P_{data}(x)+P_g(z)}=\frac{P_{data}(x)}{P_{data}(x)+P_g(z)}. \tag{4}$$

Generator in $\max_D V(D,G)$ takes the minimum value, the whole network reaches the Nash equilibrium, when both the generator and the discriminator reach the optimum, in this state, the image generated by the generator is almost indistinguishable from the discriminator, i.e., it achieves the purpose of image restoration, namely:

$$C(G)=\max_D V(D,G)=R_{X\sim P_{data}}\left[\lg\frac{P_{data}(x)}{P_{data}(x)+P_g(z)}\right] \\ +R_{X\sim P_g}\left[\lg\frac{P_{data}(x)}{P_{data}(x)+P_g(z)}\right]. \tag{5}$$

4 Image restoration

Evaluate the trained LSGAN model on the verification set or test set to ensure that it can effectively restore damaged factory monitoring images. Input the damaged factory monitoring image into the trained generator to generate the restored image.

Based on the above process, generate LSGAN model training pseudocode as follows:

```
Initialize parameters θ_G and θ_D for generator G and discriminator D
Set generated image label c=0, real image label a=1, and generator expected label b=0
Set number of training epochs and batch size
for epoch in 1 to epochs:
for batch in data_loader(batch_size):
# Train discriminator D
Fix parameters θ_G of generator G
Real images x, generated images G(z) = forward_propagation(batch)
Calculate discriminator loss D_loss
Backpropagate to update θ_D to minimize D_loss
# Train generator G
Fix parameters θ_D of discriminator D
Generated images G(z) = forward_propagation(batch)
Calculate generator loss G_loss
Backpropagate to update θ_G to minimize G_loss
```

2.2. Image Enhancement

In the factory surveillance image employee re recognition method, the surveillance image after image restoration still needs further processing to enhance the useful information in the image, especially for employee characteristics. Contourlet transform is very suitable as a multi-resolution, local and multi-directional image representation method. The following are the enhanced processes:

1 Wavelet Contourlet Transform

The factory monitoring image is decomposed into low frequency subband and high frequency subband through wavelet decomposition [15, 16], and each high frequency subband is subject to directional decomposition using directional filter banks to further extract the directional information of the image. The wavelet Contourlet coefficient is adjusted to enhance the texture and detail of the image in a specific frequency and direction range, so as to improve the clarity and recognition of the image. By adjusting the wavelet Contourlet coefficients to enhance the texture and details of specific frequency and directional ranges in the image, the clarity and recognition of the image can be improved. The low-frequency sub-band contains the main energy and overall structural information of the image, while the high-frequency subband contains the details and edge information of the image. Directional decomposition of high-frequency sub bands can better capture the directional features of images, and enhancing these features can help improve the clarity of employee features in images.

2 Image Fusion in Wavelet Contourlet Transform Domain

The image fusion is to obtain better map image enhancement effect. The fusion method is to use the upper wavelet Contourlet transform to realize the enhancement processing of the restored factory monitoring image. Through the enhancement transform, the transformed image A and transformed image B from the same image are obtained, and then the wavelet Contourle fusion is implemented for these two images. The fusion rule is based on pixel points:

a For low-frequency coefficients, the fusion rule is:

Weighted sum of transform image A and B coefficients [14]: when image fusion is implemented, the weighted value of the low-frequency coefficients of transform image A and B after wavelet-Contourlet decomposition is selected as the fused coefficient, namely:

$$g_N^F(i,j)=e_1g_N^A(i,j)+e_2g_N^B(i,j), \tag{6}$$

where: $e_1 + e_2 - 1, N$ is a sequence constant. The fusion of low-frequency coefficients adopts a weighted sum method, which can comprehensively consider the low-frequency information of two images, making the fused image more reasonable in overall structure and energy distribution. By adjusting the value of λ, the contribution of low-frequency information between the two images can be balanced, resulting in a fusion result that better meets the requirements.

b For high frequency coefficients, the fusion rule is:

The fusion rule of selecting the largest coefficient: select the wavelet-Contourlet coefficient with a larger value from the corresponding position in the wavelet coefficient matrix of each transformed image as the wavelet-Contourlet coefficient of the fused image, namely:

$$F(i,j)=\begin{cases}F_A(i,j),\text{if}F_A(i,j)>F_B(i,j)\\F_B(i,j),else\end{cases}. \tag{7}$$

High frequency coefficients usually contain the details and edge information of the image. Using a fusion rule with larger coefficient selection can preserve more prominent features in the image, enhance the clarity and detail representation of the image.

3 **Cyclic panning methods**

Since wavelet-Contourlet transform does not have translation invariance, it will lead to image quality degradation and visual artifacts. Therefore, the circular translation method [12] is introduced in the fusion process to implement circular translation of the original image for a certain distance, and then try wavelet-Contourlet transform and fusion processing for the translated image, and finally implement reverse translation to obtain the final enhanced image. The expression formula is as follows:

$$\tilde{D} = \frac{1}{N_1 N_2} \sum_{i=1, j=1}^{N_1, N_2} D_{-i,-j} \left(T^{-1} \left(g \left[T \left(D_{i,j} \left(x \right) \right) \right] \right) \right), \quad (8)$$

where: N_1N_2 is the maximum translation; D is the cyclic translation operator; the subscript $i, j, –i, –j$ are the translations in the row and column directions, respectively; T is the transformation operator; T^{-1} is the inversion operator; g is the denoising operator. The cyclic translation method performs wavelet Contourlet transform and fusion processing at multiple translation positions, and then superimposes and inverts the results, which can effectively improve the translation invariance of the transform, reduce image quality degradation and visual artifacts, and obtain higher quality enhanced images.

3. Factory Surveillance Image Employee Re-identification Method

3.1. Employee Foreground Image Segmentation

The mixed Gaussian model (GMM) [4, 7] is used to accurately segment the employee foreground in the factory monitoring image after preprocessing. GMM is a multimodal probability density function estimation method, which can well handle multiple distribution patterns of pixel values in images. In factory monitoring, due to various factors such as employees, machines, lighting, etc., pixel values may present multiple different distribution patterns. GMM can use multiple Gaussian distributions to simulate these patterns, so as to accurately separate the employee foreground area and the factory background area.

1 **Parameter initialization**

For each pixel point x_t in the factory surveillance images at time t, initialize number of K Gaussian distribution:

$$\begin{cases} A\left(x_t\right) = \sum_{i=1}^{K} \frac{\xi_{i,t}}{\left(2\pi\right)^{d/2} \left|Cov_{i,t}\right|^{1/2}} \times e^{-\frac{1}{2}\left(x_t - v_{i,t}\right)^T \sum_{i,t}^{-1}\left(x_t - v_{i,t}\right)} \\ Cov_{i,t} = \varsigma_i^2 O \end{cases}, \quad (9)$$

where $\xi_{i,t}$, $v_{i,t}$, $Cov_{i,t}$ are weights, means and covariance matrices of the i individual Gaussian distributions [19] at the time t; T is the background threshold. The maximum number of Gaussian distributions per pixel point is $K_{max} = 4$, initialize the background model with the number of Gaussian models per pixel point as $K = 1$, the pixel values at each point of the first frame are used to initialize the Gaussian distribution mean $v_{K,0}$, standardized variance ς taking the relatively larger value, i.e., the $\varsigma_{K,0} = 20$, the mixed Gaussian weights is $\frac{1}{K_{\max}}$.

2 **Background model learning and updating**

When detecting the scenic spots in front of the employee area, the general x_t matches each Gaussian distribution one by one according to the priority order ξ/ς. If no Gaussian distribution of the background model matches x_t, the point is judged as the former scenic spot, otherwise it is the background point. If no Gaussian distribution matches with x_t are found, the corresponding processing is performed according to the new Gaussian distribution generation criteria. The specific implementation is provided as follows:

a Matching guidelines

The existing K Gaussian distribution parameters are matchedbyprioritywiththecurrentpixelvaluex_t,that is, whether the $\left|v_{i,t} - x_t\right| \left\langle \max\left(2\varsigma_{K,0}, \upsilon\right)\right\rangle, i = 1, 2, ..., K$ is satisfied, of which υ is a threshold constant.

b Background learning and updating

Background learning and updating are carried out simultaneously using the same iterative equation. Using the current observation with the pre-existing i Gaussian model matched, if successful, update the matched i-th Gaussian model distribution parameter as below:

$$\begin{cases} \nu_{i,t+1} = \left(1-\beta_i\right)\nu_{i,t} + \beta_i x_t \\ \varsigma^2_{i,t+1} = \left(1-\beta_i\right)\varsigma^2_{i,t} + \beta_i\left(\nu_{i,t} - x_t\right)^2 \\ \beta_i = \beta e^{-\left(x_t - \nu_{i,t}\right)^2/2\varsigma^2_{i,t}} \end{cases} . \qquad (10)$$

For the Gaussian distribution with no successful match, its ν, ς remain unchanged. The weights of the K Gaussian distributions are updated as follows:

$$\begin{cases} \xi_{i,t+1} = \left(1-\alpha_i\right)\xi_{i,t} + \alpha_i Q_{i,t} \\ \alpha_i = \alpha e^{-\left(x_t - \nu_{i,t}\right)/2\varsigma^2_{i,t}} \end{cases} , \qquad (11)$$

where α is the rate at which the weights are updated, which is used to prioritize the Gaussian component weights in the background. The smaller α it is, the more stable the background component. β_i is the rate of updating for the background, the larger β_i it is, the faster the background component converges. For the matched Gaussian component as $Q_{i,t} = 1$, other mismatches K–1 components are $Q_{i,t} = 0$. After updating the parameters of the Gaussian distribution and the weights of the distributions, the distributions were reprioritized and reordered, and the number of background distributions was determined.

c The new Gaussian distribution generation criterion

When none of the existing K Gaussian distributions can be matched to the current pixel value, and K<$K_{\max}$, adding a new Gaussian distribution to the background model, i.e., the K=K–1. Its mean is initialized with the current pixel value,the standard deviation and the weights are set to, respectively, $20, \dfrac{1}{K_{\max}}$. If the number of Gaussian distributions has reached the upper limit,i.e., the K=$K_{\max}$, temporarily create a new Gaussian distribution, set its mean value to the current pixel value, initialize the standard deviation and weight to 20 and 0.01, respectively, and carry out iterative updates according to the GMM update rules, when the weight of this new component is greater than a certain threshold $T_\xi = \max\left(\xi_{i,t}\right), i = 1,2,\ldots,K_{\max}$, the distribution in which offspring replace the original with the least weight. In the iterative updating process, the mean and variance of the new distribution will be updated according to Formula (10), and the weight of the new distribution will be dynamically adjusted and updated using the sigmoid function [9, 23], that is $\alpha_i = \dfrac{1}{1+e^{0.1(Term-c)}}$, where, constant *Term* is used to control how long foreground pixels are blended into the background after they remain stationary, variable c is used to count the number of times an observation matches the new Gaussian distribution. When the new distribution weight update increases to the threshold value T_ξ, the Gaussian distribution parameterized by the iteratively updated mean, variance and weights will be substituted for the one with the smallest weight among the original components. This larger weight T_ξ can guarantee that the newly added distribution can become a stable background component. Finally, when the new Gaussian component becomes the background model, the counting parameter c clear to zero, and for the new K weights are renormalized [11]. In the process of employee re-identification, the pixel value of the current frame is matched with the background model, and if the matching fails, the pixel point is judged as the foreground point of the employee area, so as to realize the dynamic real-time segmentation of the foreground area of the employee.

In summary, the implementation process of GMM parameter update is as follows:

```
Initialize parameters (ω_k, μ_k, σ_k) for K Gaussian distributions at each pixel
Set weight update rate α, background update rate ρ, and matching threshold T
for each frame in video stream:
for each pixel in image:
current_pixel = get current pixel value
matched = False
# Match Gaussian distributions by priority
for k in 1 to K:
current_pixel - μ_k
# Match successful, update parameters
matched = True
break
else:
ω_k = (1-α)*ω_k
# Create new Gaussian distribution if no match
```

```
if not matched and K < max_distributions:
K += 1
ω_K = 0.01
μ_K = current_pixel
σ_K = initial_std
# Resort and normalize weights
Sort Gaussian distributions by ω_k/σ_k
Normalize weights so that Σω_k = 1
# Foreground segmentation
if not matched:
Mark as foreground point
else:
Mark as background point
```

3.2. Employee Re-identification

An employee re-identification method based on ReIDCaps for factory surveillance images is proposed. For the segmented employee foreground area, it includes three core modules: 1) visual feature extraction module; 2) Employee identity information perception module; and 3) Auxiliary module.

1 Visual feature extraction module

Use O_m^x denote the input image of the foreground region of the employee, where m is the index of employee identity, use DenseNet121 to extract the employee foreground image O_m^x of the underlying visual features, outputting a feature map $P(O_m^x)\epsilon^{\sim 7*7*1024}$. The feature map is transferred to the subsequent capsule network layer (main body) [3, 20], feature sparse representation (FSR) layer and soft embedded attention (SEA) layer. These three branch losses are expressed as L_{CAPS}, L_{FSR}, L_{SEA}. In these three modules, FSR and SEA are two branches of the auxiliary module, so the objective function of ReIDCaps network is obtained as follows:

$$L = L_{CAPS} + \eta * \left(L_{FSR} + L_{SEA} \right), \tag{12}$$

where η is to balance the contribution weights of the capsule layer and auxiliary modules. This objective function comprehensively considers the losses of the capsule network layer, FSR layer, and SEA layer. By adjusting the weight values, it can balance the impact of different modules on network performance, enabling the network to better extract the features of employee foreground images and improve the accuracy of employee re identification.

2 Employee identity awareness module

Given the input set of images $P(O_m^x)$ of the foreground regions of employees, of which, O_m^x is foreground images of the m-th group. The Attention-free Capsule Network (basic capsule layer P-Caps and classified capsule layer C-Caps) is introduced to re identify employees through the dynamic routing process.

a Basic capsule layer (P-Caps)

Implement processing for feature maps $P(O_m^x)$, construct P-Caps layer, and obtain multiple 8-dimensional vector capsules, which are recorded as b_k^{8D}, of which $k\,\epsilon\,[1{,}288]$, the length of each vector capsule is normalized using a nonlinear squeezing function to ensure that its length is between [0,1]:

$$b_k^{8D} = \frac{\left\| b_k^{8D} \right\|^2}{1 + \left\| b_k^{8D} \right\|^2} \cdot \frac{b_k^{8D}}{\left\| b_k^{8D} \right\|}. \tag{13}$$

The nonlinear squeezing function can compress the length of the capsule vector to between [0,1], so that the length of the vector can represent the probability of the entity represented by the capsule, while retaining the directional information of the vector, which is helpful for subsequent classification and recognition.

b Classification capsule layer (C-Caps)

Based on the output of the P-Caps layer, the employee identity capsule in the C-Caps layer is obtained by identifying the employee identity through the dynamic routing process. In C-Caps layer, there are N employee identity capsules. N is the number of different employee identities in the training set. Convert each 8-dimensional vector in P-Caps layer b_k^{8D} which maps the dimensions to 24 dimensions V_m^{24D}. Use coupling coefficient and mapped vector capsule to calculate each employee identity capsule V_m^{24D} in C-Caps layer:

$$V_m^{24D} = \sum_{k=1}^{K} u_k^n \cdot b_k^{24D}, \tag{14}$$

where $m\,\epsilon\,[1{,}N]$, k is the total number of C-Caps layer weight vector capsules; u_k^n is the coupling coefficient of the corresponding employee identity capsule in the C-Caps layer, which is determined by the dynamic routing process between the P-Caps and CCaps layers. The dynamic routing process iteratively updates the coupling coefficient, allowing capsules in the P-Caps layer that have a higher correlation with employee identity capsules in the C-Caps layer to transmit information more effectively, thereby improving the accuracy of employee identity recognition.

c Loss function training

Use an appropriate loss function (marginal loss) to train the ReIDCaps network to minimize the difference between the predicted employee identity and the actual identity. The mathematical formula is as follows:

$$L_{CAPS} = \sum_{m=1}^{N}\left\{ y_m \cdot \max\left(0, q^{+} - \left\|V_m^{24D}\right\|^{2}\right) + \mu\cdot\left(1-y_m\right)\cdot\max\left(0,\left\|V_m^{24D}\right\| - q^{-}\right)^{2}\right\}, \tag{15}$$

where y_m is the input image O_m^x whether or not the employee's status exists, and if so, the y_m=1, otherwise y_m=0. Use μ=0.5 to balance the weights between the two components, using $q^{+} = \frac{N-1}{N}$ and $q^{-} = \frac{1}{N}$ to control the length of V_m^{24D}. The marginal loss function encourages the length of correctly classified capsule vectors to be greater than that of misclassified capsule vectors, and the difference between the two is greater than a set marginal value, thereby enabling the network to better learn the characteristics of employee identity and improve classification accuracy.

3 Supporting modules

The auxiliary module mainly includes FSR and SEA modules. The output through the backbone network is pooled in a global average way to obtain the input image O_m^x CNN of $F(O_m^x)$:

$$O_m^x = \frac{1}{J\times W}\sum_{i=1}^{J}\sum_{j=1}^{W} P\left(O_m^x\right)\left(i,j\right), \tag{16}$$

where J,W is, respectively, the height and width of $P(O_m^x)$; After the global average pooling layer, the inputs of FSR and SEA are $F_{FSR}(O_m^x)$ and $F_{SEA}(O_m^x)$.

The FSR module is used to improve the generalization ability based on the output of convolutional neural network (CNN). CNN network has higher adaptability than the traditional neural network. Then apply the Dropout layer $F(O_m^x)$ to obtain sparse representation $F_{FSR}(O_m^x)$. Dropout layer randomly sets some neurons to zero, which helps to prevent over fitting and improve network robustness.

The SEA module passes the $P(O_m^x)$ upper global average pooling to obtain features that $F_{SEA}(O_m^x)$, then using the squeeze and excitation module, the features are compressed into a low-dimensional space and processed through a fully connected layer to generate weight vectors $F'(O_m^x)$. After the weight vector is activated through the sigmoid, it is used to weight the original feature in the channel direction and get the output $P'\left(O_m^x\right) = F'\left(O_m^x\right)\otimes P\left(O_m^x\right)$, thereby enhancing the network's ability to discriminate between critical features, where the $\otimes$ is the channel direction multiplication between $F'(O_m^x)$ and $P(O_m^x)$.

In the FSR and SEA auxiliary modules, the basic cross entropy loss is combined with the above two losses to jointly optimize the model, namely $L_{FSR,SEA} = L_{crossentropy} + L_{labelsmooth} + L_{cirele}$. Label smoothing regularized cross-entropy loss helps to prevent overfitting of the model to the training data and improves the generalization ability, which is calculated as follows:

$$L_{labelsmooth} = -\frac{1}{I\times K}\sum_{i=1}^{I}\sum_{j=1}^{K} p_{i,j}\ln\left(\left(1-\varepsilon\right)q_{i,j} + \frac{p_{i,j}}{B}\right), \tag{17}$$

where I is the number of employees; K is the number of pictures. B is the number of employees. $p_{i,j}$ is the predicted probability; $q_{i,j}$ is the true probability; ε is the smoothing factor. Label smoothing regularization cross entropy loss introduces smoothing factors to smooth real labels, reducing the model's excessive dependence on training data and improving its generalization ability.

Circle Loss is a loss function suitable for re recognition tasks. It takes into account the angular relationship between feature vectors, which helps to optimize the feature space. The formula is:

$$
\begin{aligned}
L_{cirele} &= \log\left[1+\sum_{i=1}^{U}\sum_{j=1}^{Z}\exp\left(\eta\left(s_n^i - s_p^i + m\right)\right)\right] \\
&= \lg\left[1+\sum_{j=1}^{Z}\exp\left(\eta\left(s_n^i + m\right)\right)\sum_{i=1}^{U}\exp\left(\eta\left(-s_p^i\right)\right)\right],
\end{aligned}
\tag{18}
$$

where s_p is intraclass similarity; s_n is interclass similarity; U is the number of intra-class similarity scores; Z is the number of similarity scores between classes; η is the scale parameter; m is the rigor of the optimization. Circle Loss redefines the optimization objectives of intra class similarity and inter class similarity, allowing the network to more flexibly adjust the distribution of samples in the feature space and improve the performance of re identification.

Eventually, the model will output a vector of probability distributions, where each element represents the probability that the input image of the foreground area of a factory employee belongs to the corresponding employee identity. By comparing these probability values, the employee identity predicted by the model can be determined, and usually the identity with the largest probability value is selected as the prediction result.

4. Experiments and Discussions

In order to verify the overall effectiveness of the employee re-identification method based on Attention-free Capsule Network for factory surveillance images, relevant tests need to be carried out.

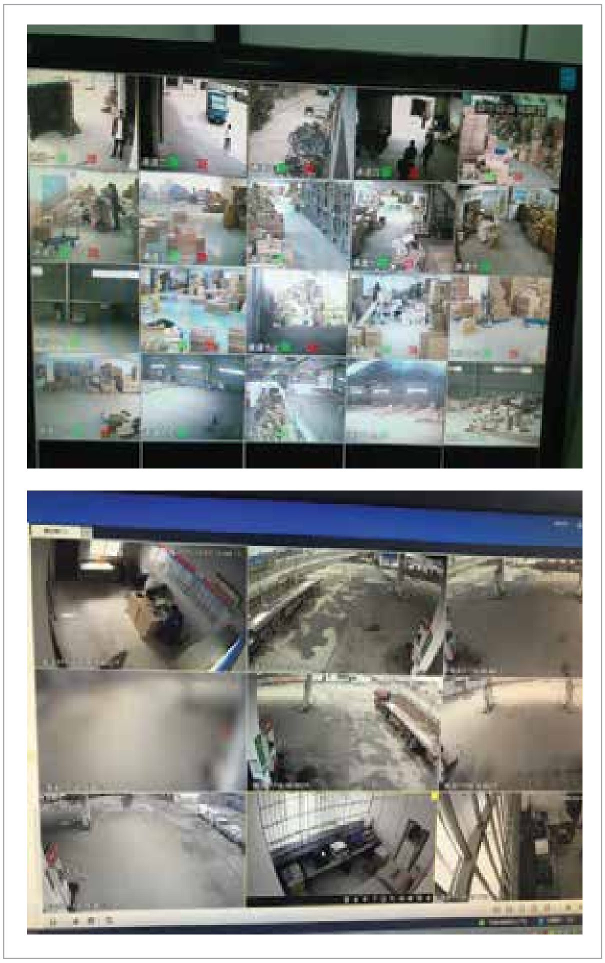
4.1. Experimental Environment

This experiment focuses on the re recognition task of employees in factory monitoring images. In order to verify the effectiveness and generalization ability of the method, public data sets similar to the factory environment were selected for pre training and testing, and the actual monitoring images of the factory (Figure 3) were collected as the final experimental data set. The experimental hardware environment includes a high-performance GPU server to support rapid training of models and processing of large-scale data sets.

4.2. Introduction to the Data Set

Factory Staff ReID (user-defined data set): in order to be closer to the actual factory environment, the factory monitoring images are collected and the employee identity tags are marked. The dataset contains 50 different employee identities, each of which contains images taken from multiple camera perspectives. It is divided into training set, verification set and test set for training and evaluating re recognition methods.

Figure 3
Actual monitoring images of the factory.



4.3. Parameterization

The experimental configuration parameters were pre-set, as shown in Table 1.

Depending on the characteristics of the methodology and the characteristics of the data set, the following parameter settings are developed:

1 Learning rate: adopting a segmented constant learning rate strategy, initially set at 0.001, this val-

Table 1

Experimental environment and configuration parameters.

Indicators	Configuration
Memory	48GB
GPU	RTX3090(24GB)
Computing platforms	CUDA11.3
Compiled language	Python3.8
Open-source framework	Python1.10.0
Input image size	256*128
Resolution	640*480
Pixel Depth	8Bit
Image element size	3~10
Signal-to-noise ratio	45~55dB
The length of the video	24h
Storage capacity	2.4TB
Network bandwidth	1080P

ue is stable and reasonable in most tasks, avoiding parameter updates that are too fast or too slow; Decay to 1/10 every 5 or 10 rounds to finely adjust parameters in the later stages of training, helping the model converge better.

2 Batch Size: set to 32 or 64 based on memory and computing resources. 32 increases randomness to help escape local optima and has low memory requirements, while 64 facilitates GPU parallel computing and improves efficiency. The two can balance resource utilization.

3 Number of wavelet transform levels: select 4 levels based on image size and detail requirements. The series is often finely decomposed but computationally complex, with 4 levels achieving a good balance between detail capture and computational efficiency.

4 Number of directions of Contourlet transform: select 4-8 directions to capture multi-directional features, balancing feature extraction and computational complexity within this range.

5 GMM parameter convergence iterations: set to 100 times, according to experimental summary, this number can ensure accurate parameter estimation and avoid excessive calculation.

6 Capsule dimension: set to 8 or 16 dimensions based on task and dataset characteristics. Dimensions affect feature complexity and computational complexity, balancing feature representation and computational efficiency within this range.

7 Routing iterations: set to 3 times in capsule networks, dynamic routing requires iterative weight adjustment. 3 times can ensure accuracy and avoid excessive computation.

4.4. Experimental Content

In order to comprehensively evaluate the performance of an employee re-identification method based on Attention-free Capsule Network for factory surveillance images, the following experiments are designed:

1 Pre-processing experiments: On the preset factory monitoring image data set, the proposed method is used to carry out image restoration and enhancement related pre-processing to verify the image processing capability of the proposed method.

2 Foreground segmentation experiment: continue to use the proposed method for the pre-processed image to start the foreground segmentation of the employee area, and analyze whether the proposed method can effectively distinguish the employee area and the background area of the factory reasonably.

3 Comparative experiment: the proposed method is compared with the methods in other studies[2-4] in terms of visualization, cumulative matching curve (CMC) index and mean of average precision (mAP) index, and the advantages of the proposed method are evaluated by comparing the performance of different methods. Among them, the cumulative matching curve (CMC) is the hit probability (*Rank–k*) of *top–k*, means that before in the gallery *k* sub-matching success, and thus the probability of finding the target identity employee. Adopting *Rank*–1 as the main indicator, *Rank*–5 and *Rank*–10 as a supporting indicator which is used to assess the probability of a successful *k*-th amplitude match, thus recording its success in different *N* CMC curve and first matching success rate under value. In addition, the mean precision (mAP) is used as another indicator to measure the effectiveness of the recognition method, which is used to measure the average precision of the method for 10 groups of images. The calculation formula is:

$$mAP = \frac{\sum_{k=0}^{N} AP_k}{N}, \quad (19)$$

where $\sum_{k=0}^{N} AP_k$ is the average precision for each category; N is the total number of categories.

4.5. Experimental Results and Analysis

1 Pre-processing results

Randomly select a group of factory monitoring images as the original image, we can see that the original factory monitoring images appear fuzzy, pixel missing and more colorful phenomena, using the proposed method to carry out pre-processing, the results are shown in the Figure 4.

From the results of preprocessing (as shown in Figure 4), it can be seen that the proposed method effectively solves a variety of problems in the original image. Aiming at the blur phenomenon of the image, through the restoration processing of the least squares generation antagonism network (LSGAN), the clarity and detail information of the image are successfully restored, the missing pixels are filled, and the image is more complete. In the case of more color, the wavelet-Contourlet transform image enhancement method is used to effectively suppress the interference of noise and noise, make the image more pure and clear, and provide a better basic environment for subsequent staff's re recognition task.

2 Foreground segmentation results

Two groups of factory monitoring images containing different lighting conditions, occlusion conditions and noise interference were screened, and the segmentation test of employee foreground area was carried out using the Gaussian mixture model (GMM) to show the following segmentation result images.

Figure 5 shows the test results of employee foreground region segmentation through Gaussian mixture model (GMM). Under different lighting conditions, occlusion and noise interference, the GMM method can more accurately separate the foreground area of employees from the factory background. In the scene with large illumination changes, the method can still maintain a good segmentation effect, showing strong robustness. Despite some challenges, GMM method can still capture the prospects of most employees under occlusion and circumstances, providing valuable information for subsequent employee re identification. In general, GMM method performs well in the task of employee foreground segmentation of factory monitoring images, laying a solid foundation for improving the accuracy and efficiency of employee recognition.

Figure 4
Pre-processing results of the proposed method.

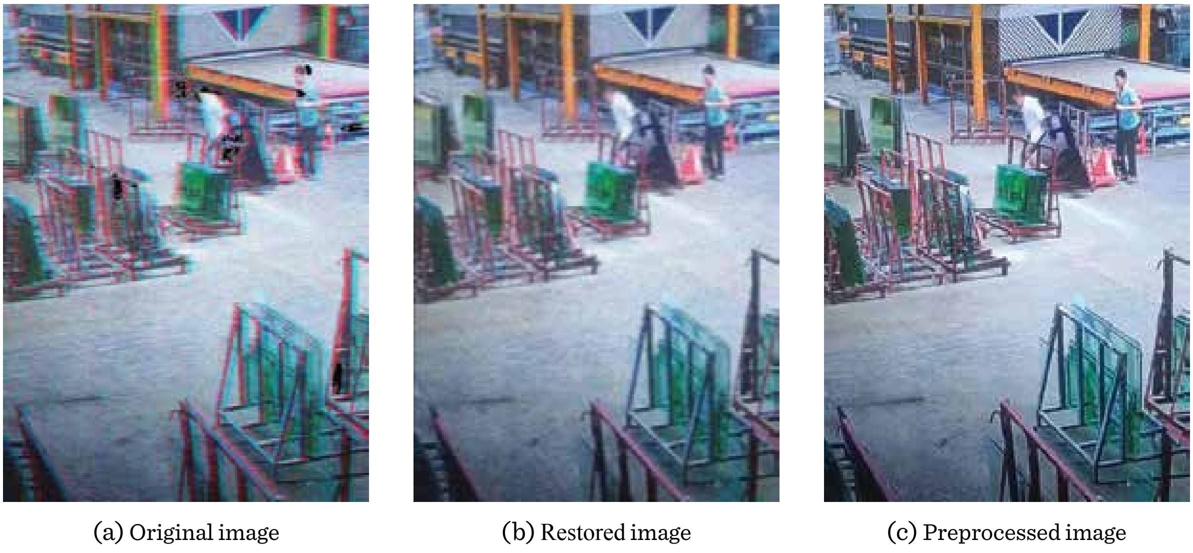
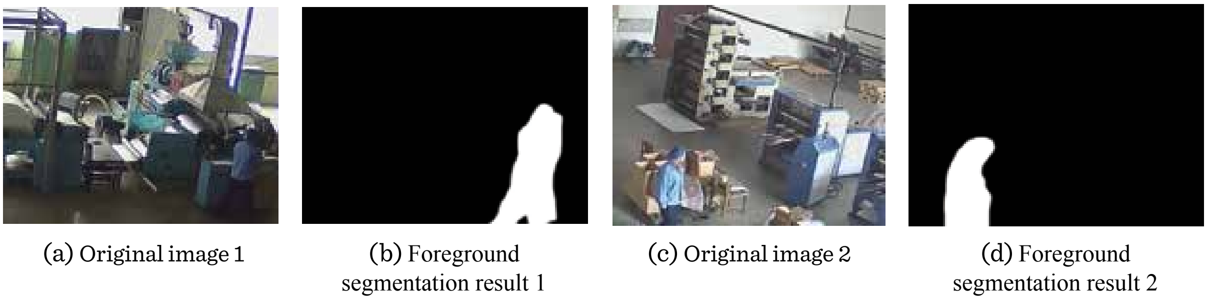

(a) Original image (b) Restored image (c) Preprocessed image

Figure 5

Foreground segmentation results of the proposed method.



(a) Original image 1 (b) Foreground segmentation result 1 (c) Original image 2 (d) Foreground segmentation result 2

3 Comparative validation

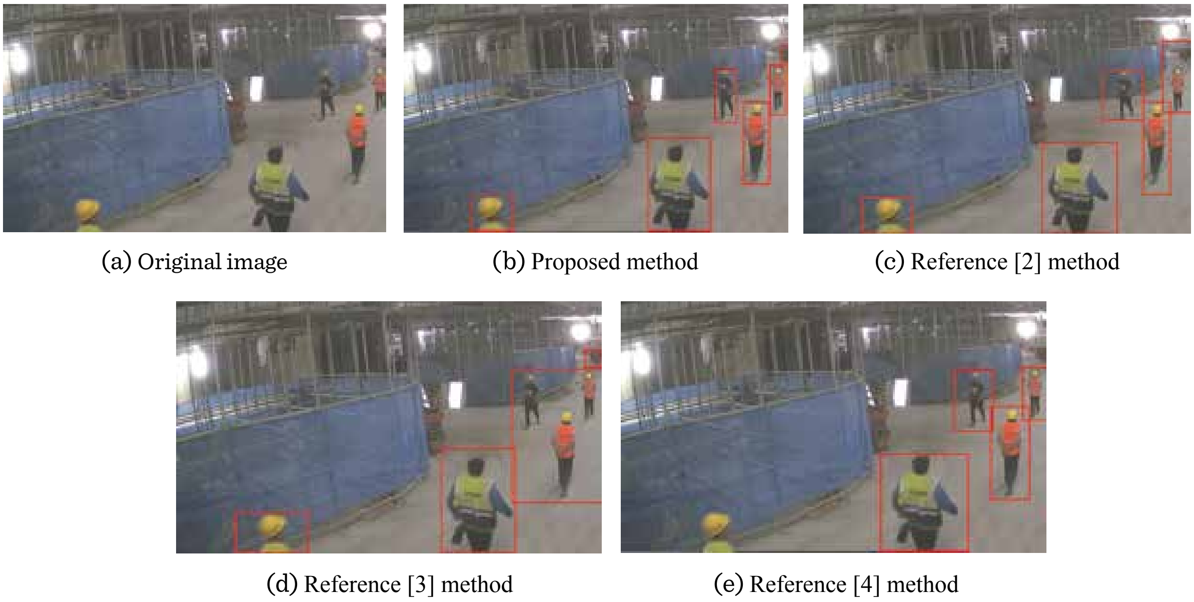
a Visualization contrasts

In order to clearly reflect the effectiveness of the proposed method, the introduction of the reference methods in references [1, 5, 21], respectively, and the proposed method to start the comparison, select a group of factory monitoring images containing multiple employees, using the above four methods to start the employee re-identification, the recognition results are presented in the form of an image, the figure marked with a red box for the identified employees, the specific results are shown in the Figure 6.

Figure 6 clearly shows the significant advantages of the proposed method in employee identification. This method can accurately select different employees in the image, and even employees located at a distance from the image can be effectively identified, which fully proves that the proposed method has strong anti-interference and adaptability, and can flexibly respond to various challenges in the factory environment; The

Figure 6

Recognition results of different methods.



(a) Original image (b) Proposed method (c) Reference [2] method

(d) Reference [3] method (e) Reference [4] method methods proposed in references [1, 5] have made obvious misjudgments in the process of employee identification. Different employees are incorrectly identified in the same frame, which may lead to serious identification errors in practical application; There is an obvious problem of missing recognition in the method in [21], that is, some employees are not correctly identified in the image, which may reduce the accuracy of the entire recognition system.

By comparing the recognition results in Figure 6, it can be clearly seen that the proposed method is superior to the other three methods in the employee recognition task. Not only does it have strong anti-interference and environment adaptability, it can accurately recognize different employees in the image, and it avoids the problems of misjudgment and omission of recognition that occur in other methods. It proves the potential and value of the proposed method in the field of employee re-identification in factory surveillance images.

b Cumulative Matching Curve (CMC) and mAP Value Comparison

Ten sets of factory surveillance images of different types (different lighting, occlusion, time of day, and number of employees) are selected as test images for comparison experiments. The experimental results are shown in Table 2.

Through the experimental test on 10 groups of different types of factory monitoring images, it is found that the proposed method performs well in the recognition task of employees. Compared with the other three comparison methods, The value of the CMC curve of the proposed method rises rapidly at smaller values of k, indicating that it can successfully find the target employee within a few matching times in the library, which proves its efficiency and accuracy. At the same time, the first matching success rate of the proposed method is also significantly higher than that of other methods, and other methods have problems such as unstable identification and low success rate, which further proves its superior performance in re identification tasks of employees. In terms of average precision mean (mAP), the proposed method also performs well, and its mAP value reaches about 0.90, significantly higher than the other three comparison methods. This result fully demonstrates the advantages of the proposed method in recognition accuracy, and shows that the method can stably and accurately identify target employees in a complex factory monitoring environment.

5. Conclusion

In order to ensure effective identification and monitoring of factory employees, the proposed method provides high-quality data input for capsule network recognition through LSGAN image restoration, wavelet-Contourlet transform image enhancement and mixed Gaussian model employee foreground segmentation. Through its unique dynamic routing mechanism, the Attention-free Capsule Network effectively aggregates and classifies image features, and realizes accurate identification of employee identity. The proposed method not only improves the accuracy of employee re identification, but also provides strong technical support for factory safety management.

Although the proposed method has certain advantages, it also has certain limitations. In terms of computational complexity, operations such as dynamic routing mechanisms, wavelet and Contourlet transforms, and multiple iterations of GMM

Table 2

Comparison results of the recognition performance of the methods.

Method	Rank-1/%	Rank-5/%	Rank-10/%	mAP/%
The proposed method	95.8	98.4	99.5	89.9
Method in [1]	92.1	95.2	96.6	85.2
Method in [5]	85.8	90.2	91.1	80.7
Method in [21]	90.8	91.8	93.8	82.2

parameter estimation in capsule networks all increase computational resource consumption and training time costs; Meanwhile, the method may be sensitive to specific types of noise, such as periodic noise or high-intensity random noise that is similar to the true texture features of the image in image data, which may interfere with feature extraction and model judgment. Therefore, future research will continue to focus on two key dimensions: algorithm optimization and cross domain fusion to improve method performance and expand application boundaries. On the one hand, in order to address the computational complexity issues of existing methods, we need to further optimize algorithm design. For example, analyzing the computational bottlenecks of dynamic routing mechanisms in capsule networks, exploring lightweight routing algorithms or approximate calculation methods, and reducing computational overhead while ensuring routing accuracy; For wavelet and Contourlet transform operations, research fast algorithms or hardware acceleration implementations to improve feature extraction efficiency; For multiple iterations of GMM parameter estimation, adaptive iteration termination conditions are adopted to dynamically adjust the number of iterations based on data distribution and convergence, reducing unnecessary calculations. On the other hand, given the latest breakthroughs in the field of multimodal learning, such as TriChronoNet achieving efficient collaborative processing of different modal data through multi module fusion [6], hyper relational interaction modeling mining complex inter modal relationships [13], and modal fusion visual transformers [18] demonstrating excellent performance in multimodal visual tasks. Subsequent research will actively integrate these cutting-edge achievements. Explore the introduction of multimodal fusion strategies and complex relationship modeling techniques from these methods into the current research system, delve into the potential correlations between different modal data, develop more efficient cross modal information integration algorithms, and construct models with stronger scene adaptability.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Code, Data, and Materials Availability

Data sharing is not applicable to this article, as no new data were created or analyzed.

Acknowledgments

Scientific Research Project of Liaoning University of International Business and Economics (2023XJLX-ZD01).

References

1. An, F. P., Liu, J. Pedestrian Re-Identification Algorithm Based on Visual Attention-Positive Sample Generation Network Deep Learning Model. Information Fusion, 2022, 86, 136-145. https://doi.org/10.1016/j.inffus.2022.07.002
2. Ardeshiri, R. R., Razavi-Far, R., Li, T., Wang, X., Ma, C., Liu, M. Gated Recurrent Unit Least-Squares Generative Adversarial Network for Battery Cycle Life Prediction. Measurement, 2022, 196, 111046. https://doi.org/10.1016/j.measurement.2022.111046
3. Bang, J., Park, J., Park, J. GACaps-HTC: Graph Attention Capsule Network for Hierarchical Text Classification. Applied Intelligence, 2023, 53(17), 20577-20594. https://doi.org/10.1007/s10489-023-04585-6
4. Gecili, E., Sivaganesan, S., Asar, O., Clancy, J. P., Ziady, A., Szczesniak, R. D. Bayesian Regularization for a Nonstationary Gaussian Linear Mixed Effects Model. Statistics in Medicine, 2022, 41(4), 681-697. https://doi.org/10.1002/sim.9279
5. Han, H., Zhou, M. C., Shang, X., Cao, W., Abusorrah, A. KISS+ for Rapid and Accurate Pedestrian Re-Identification. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(1), 394-403. https://doi.org/10.1109/TITS.2019.2958741
6. He, M., Jiang, W., Gu, W. TriChronoNet: Advancing Electricity Price Prediction with Multi-Module Fusion. Applied Energy, 2024, 371, 123626. https://doi.org/10.1016/j.apenergy.2024.123626
7. Jung, M. An L0-Norm Based Color Image Deblurring Model Under Mixed Random-Valued Impulse and Gaussian Noise. Applied Mathematical Model-

ling, 2022, 102, 847-866. https://doi.org/10.1016/j.apm.2021.10.027

8. Kim, H., Lee, K., Lee, C., Hwang, S., Youn, C. H. An Alternating Training Method of Attention-Based Adapters for Visual Explanation of Multi-Domain Satellite Images. IEEE Access, 2021, 9, 62332-62346. https://doi.org/10.1109/ACCESS.2021.3074640
9. Kollem, S. A Fast Computational Technique Based on a Novel Tangent Sigmoid Anisotropic Diffusion Function for Image Denoising. Soft Computing, 2024, 28, 7501-7526. https://doi.org/10.1007/s00500-024-09628-9
10. Liu, B., Wang, Y. W. Super Resolution Image Reconstruction Algorithm Based on Generated Countermeasure Network. Computer Simulation, 2023, 40(10), 269-273.
11. Liu, J., Hu, S., Li, H., Liu, Y., Huang, B., Sun, Y. Achieving Threshold Consistency in Three-Way Group Decision Using Optimization Methodology and Expert-Weight-Updating Strategy. International Journal of Approximate Reasoning, 2023, 158, 108922. https://doi.org/10.1016/j.ijar.2023.108922
12. Liu, S., Liu, M., Li, P., Zhao, J., Zhu, Z., Wang, X. SAR Image Denoising via Sparse Representation in Shearlet Domain Based on Continuous Cycle Spinning. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(5), 2985-2992. https://doi.org/10.1109/TGRS.2017.2657602
13. Lu, Y., Wang, W., Bai, R., Zhou, S., Garg, L., Bashir, A. K., Jiang, W., Hu, X. Hyper-Relational Interaction Modeling in Multi-Modal Trajectory Prediction for Intelligent Connected Vehicles in Smart Cities. Information Fusion, 2025, 114, 102682. https://doi.org/10.1016/j.inffus.2024.102682
14. Miyata, T., Aoki, Y. Perceptual JPEG Artifact Removal Using Weighted Sum of IQAs as Loss Function. Nonlinear Theory and Its Applications, 2024, 15(4), 725-736. https://doi.org/10.1587/nolta.15.725
15. Narendiranath Babu, T., Senthilnathan, N., Pancholi, S., Nikhil Kumar, S. P., Rama Prabha, D. Fault Analysis on Continuous Variable Transmission Using DB-06 Wavelet Decomposition and Fault Classification Using ANN. Journal of Intelligent and Fuzzy Systems, 2021, 41(1), 1297-1307. https://doi.org/10.3233/JIFS-210199
16. Peng, G., Liu, D., Lu, J., Shen, T., Wang, S. Analysis of Rock Microseismic Signal Based on Blind Source Wavelet Decomposition Algorithm. AIP Advances, 2022, 12(5), 055316. https://doi.org/10.1063/5.0085361
17. Xing, H. J., He, Z. C. Adaptive Loss Function Based Least Squares One-Class Support Vector Machine. Pattern Recognition Letters, 2022, 156, 174-182. https://doi.org/10.1016/j.patrec.2021.12.014
18. Yang, B., Wang, X., Xing, Y., Cheng, C., Jiang, W., Feng, Q. Modality Fusion Vision Transformer for Hyperspectral and LiDAR Data Collaborative Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17, 17052-17065. https://doi.org/10.1109/JSTARS.2024.3415729
19. Yi, R., Wang, T., Lyu, B. Reconstructing the Dimensions of Consumer-Based Brand Equity: An Empirical Study in the E-Commerce Environment. Frontiers in Psychology, 2022, 13, 925474. https://doi.org/10.3389/fpsyg.2022.925474
20. Zhang, Q., Li, J., Ding, W., Ye, Z., Meng, Z. Mechanical Fault Intelligent Diagnosis Using Attention-Based Dual-Scale Feature Fusion Capsule Network. Measurement, 2023, 207, 112345. https://doi.org/10.1016/j.measurement.2022.112345
21. Zhang, T., Cao, Y., Zhang, L., Li, X. Efficient Feature Fusion Network Based on Center and Scale Prediction for Pedestrian Detection. The Visual Computer, 2023, 39(9), 3865-3872. https://doi.org/10.1007/s00371-022-02528-9
22. Zheng, S., Lan, F., Castellani, M. A Competitive Learning Scheme for Deep Neural Network Pattern Classifier Training. Applied Soft Computing, 2023, 146, 110662. https://doi.org/10.1016/j.asoc.2023.110662
23. Zhong, R., Fu, Y., Song, Y., Han, C. A Fusion Approach to Infrared and Visible Images with Gabor Filter and Sigmoid Function. Infrared Physics and Technology, 2023, 131, 104696. https://doi.org/10.1016/j.infrared.2023.104696