Cycle GAN-MF: A Cycle-consistent Generative Adversarial Network Based on Multifeature Fusion for Pedestrian Re-recognition

Yongqi Fan[1], Li Hang[1], and Botong Sun[2]

[1] Software College of Shenyang Normal University
No. 253, Huanghe North Street, Huanggu District, Shenyang 110034 China
vanyongqi@gmail.com;lihangsoft@163.com
Corresponding author: Hang Li
[2] College of Journalism and Communication, Shenyang Normal University
No. 253, Huanghe North Street, Huanggu District, Shenyang 110034 China
1114670787@qq.com

Abstract. In pedestrian re-recognition, the traditional pedestrian re-recognition method will be affected by the changes of background, veil, clothing and so on, which will make the recognition effect decline. In order to reduce the impact of background, veil, clothing and other changes on the recognition effect, this paper proposes a pedestrian re-recognition method based on the cycle-consistent generative adversarial network and multifeature fusion. By comparing the measured distance between two pedestrians, pedestrian re-recognition is accomplished. Firstly, this paper uses Cycle GAN to transform and expand the data set, so as to reduce the influence of pedestrian posture changes as much as possible. The method consists of two branches: global feature extraction and local feature extraction. Then the global feature and local feature are fused. The fused features are used for comparison measurement learning, and the similarity scores are calculated to sort the samples. A large number of experimental results on large data sets CUHK03 and VIPER show that this new method reduces the influence of background, veil, clothing and other changes on the recognition effect.

Keywords: Pedestrian re-recognition, Cycle-consistent generative adversarial network, Multifeature fusion, Global feature extraction, Local feature extraction.

1. Introduction

Pedestrian re-recognition refers to the verification method of whether pedestrians passing by two cameras belong to the same pedestrian within the non-crossing coverage range of two cameras, as shown in Figure 1. Pedestrian re-recognition technology has a very wide application prospect in the field of criminal investigation and home security management. Pedestrian re-identification not only plays an important role in criminal investigation, but also plays a vital role in life, such as looking for the lost elderly and children [1-3].

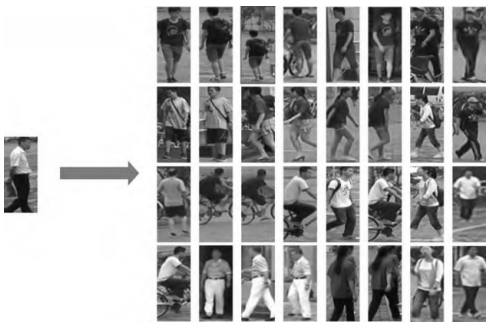

Fig. 1. Pedestrian re-recognition diagram

In recent years, advances have been made in computer vision to understand behavior using visual sensor networks. Pedestrian re-recognition has been paid more and more attention by researchers. Most traditional research methods complete pedestrian re-recognition based on low-level features, such as color, texture, clothing and other features [4,5]. In the earliest research on color feature-based recognition methods, a method of learning color features from pixels of two camera views was proposed to achieve the purpose of pedestrian re-recognition. These studies and similar ones are based on a limited number of low-level features. Therefore, all classification results

and models are differentiated in one dimension. This means that different pedestrians can only be classified by one feature. In addition, researchers have recently introduced an advanced feature representation method (sparse representation [6]) to represent the content and model behavior of pixels. Sparse representation has good pedestrian representation ability, but it can not provide enough distinguishing information for classification purposes. Some studies combine various features to classify and model pedestrian recognition [7,8]. However, its efficiency is limited by computational complexity, as combinations produce high-dimensional representations.

Due to many problems such as low resolution frame number, illumination and attitude change, occlusion, camera Angle and appearance similarity, it is difficult for traditional methods to achieve the ideal effect. In order to achieve better results in pedestrian re-recognition, researchers study the use of metric learning for pedestrian re-recognition. At present, many pedestrian re-recognition methods based on metric learning have been studied, such as KISS metric (KISSME) [9], Regularized Local Metric Learning (RLML) [10], etc. In order to reduce the influence of background, some researchers focus on the local characteristics of pedestrians. In order to extract local features, pedestrians are divided into several small blocks for local block metric learning. The traditional block segmentation includes transverse block segmentation and longitudinal block segmentation. The experimental results show that the effect of transverse block segmentation is better than that of longitudinal block segmentation.

In order to solve the above problems, this paper proposes a pedestrian re-recognition method based on the cycle-consistent generative adversarial network and multifeature fusion. The method includes two parts: global feature extraction and local feature extraction. After global feature extraction and local feature extraction, the two extracted features are fused. Local feature extraction in this paper is different from traditional local feature extraction. In this paper, in order to effectively extract local features of different pedestrians in different scenes, we use local attention model to segment and extract features of pedestrian's obvious local features.

2. Related Works

2.1. Distance Metric Learning

Traditional distance measurement methods, such as Euclidean distance, Minkowski distance, Manhattan distance [11], Chebyshev distance [12], etc., are non-learning, fixed measurement methods, these methods are directly calculated from the untransformed feature difference. In contrast, Mahalanobis distance can effectively measure the global transformation of feature space. In the feature space, the correlation and feature dimensions are emphasized, while the irrelevant feature dimensions are suppressed. This method has been popularized in many subsequent methods such as LMNN, ITML, LADF and KISSME. The measurement learning method based on Mahalanobis distance function class has been widely concerned in human re-recognition. In general, Mahalanobis distance measures the square of the distance between two data points with the uniform metric M.

The purpose of metric learning is to learn Mahalanobis distance function to measure the distance between image pairs. So the distance between positive samples is smaller than the distance between negative samples, and the metric function $d_M(x_i, y_i)$ is defined as:

$$d_M(x_i, y_i) = ||x_i - y_i||_M^2 = (x_i - y_i)^T M (x_i - y_i). \tag{1}$$

Where x_i and y_i are the characteristic values of images taken by different cameras. is a positive semi-definite matrix, which guarantees that the learned satisfies both the triangle inequality and non-negative. The purpose of metric learning is to reduce or limit the distance between similar samples and increase the distance between different categories of samples through training and learning.

2.2. Cycle GAN Network Principle

By constructing generators and discriminators, GAN network [13,14] enables two different styles of images to learn from each other and play games with each other, constantly improving their generation ability and discrimination ability. Finally, one style of image is generated into another style of image, and the generation effect is very realistic. Although there are many variant models in the GAN network, the Cycle GAN network adopts convolution and deconvolution strategies and does not require images to be input in pairs, resulting in better quality of generated images. Therefore, this paper uses the Cycle GAN network to achieve conversion and expansion between RGB images and IR images.

The principle of Cycle GAN network to achieve RGB image and IR image conversion and expansion is as follows. RGB images are passed through generator G(AB) to generate IR images, and the original IR images are passed through discriminator D(A). For example, the original IR image is judged as Real, and the generated IR image is judged as Fake. The Fake-IR image is once again generated by the generator G(BA) to generate Cycle-RGB images, which are almost consistent with the original RGB images. Similarly, IR images will also be

converted to Fake-RGB images and Cycle-IR images. Fake-RGB images and fake-IR images can be almost Fake. The conversion and extension of Cycle GAN to RGB image and IR image are shown in Figure 2.

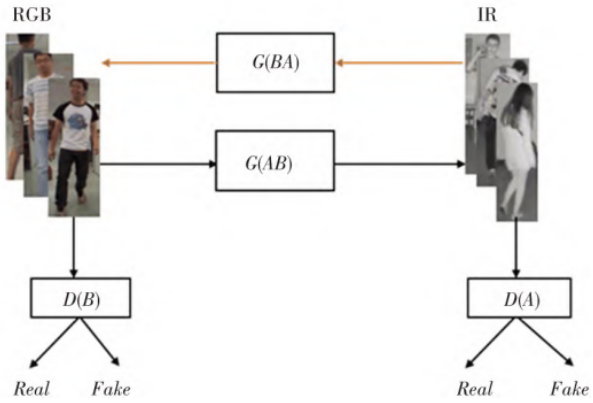

Fig. 2. Schematic diagram of Cycle GAN network for RGB image and IR image conversion and expansion

3. Proposed Pedestrian Re-recognition Method

Pedestrian re-recognition method based on local and global integration of significant areas of concern includes two parts: global feature extraction and local feature extraction. After global feature extraction and local feature extraction, the two extracted features are fused. Finally, measure learning and ranking of the fused features were carried out to screen out the samples with the highest similarity score, as shown in Figure 3. We improve the triplet loss and apply it to train the pedestrian re-recognition model.

3.1. Cycle GAN

The generator network part of the original Cycle GAN uses the deep residual network. In the improved Cycle GAN network, U-net network is used to replace the original depth residual network, and L1 distance function was added to Cycle GAN to minimize the average error of pixel level and improve the accuracy of the synthesized image.

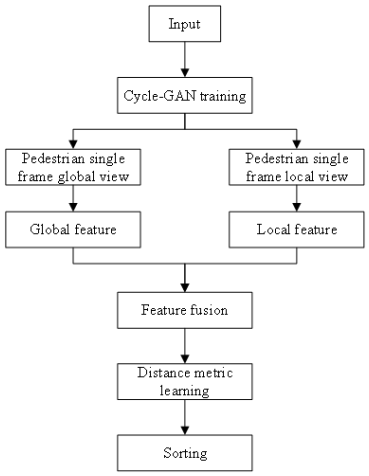

Fig. 3. Proposed pedestrian re-recognition process

Cycle GAN consists of two kinds of loss functions, namely, generative antagonistic loss function and cyclic consistency loss function. Generators G and F generate anti-loss functions as shown in formulas (2) and (3). The

forward cyclic consistency loss function $L_{forward-cyc}$ and backward cyclic consistency loss function $L_{backward-cyc}$ form the cyclic consistency loss function $L_{Cycle-consistency}$, as shown in formulas (4)-(6). The complete loss function of Cycle GAN is shown in Formula (7).

$$L_{GAN-G} = E_{Tdata}[||\log(D_{Tdata})||_1] + E_{Fdata}[||\log(1 - D_{Fdata})||_1]. \tag{2}$$

$$L_{GAN-F} = E_{Fdata}[||\log(D_{Fdata})||_1] + E_{Tdata}[||\log(1 - D_{Tdata})||_1]. \tag{3}$$

$$L_{forward-cyc} = E_{Tdata}[||(D_{Tdata})||_1]. \tag{4}$$

$$L_{backward-cyc} = E_{Tdata}[||\log(D_{Tdata})||_1]. \tag{5}$$

$$L_{Cycle-consistency} = L_{forward-cyc} + L_{backward-cyc}. \tag{6}$$

$$L_{CycleGAN} = L_{GAN-G} + L_{GAN-F} + \lambda L_{Cycle-consistency}. \tag{7}$$

Where λ is the weight of the cyclic consistency loss function.

3.2. Area of Concern Model

The area of concern refers to the local area with significant features of pedestrians, which can make pedestrians significantly different from other pedestrians such as pedestrian backpack, obvious color of the top, etc. At present, there are models based on LSTM neural network and RNN neural network. In this paper, the deep recurrent attention model is chosen as the observation area model to extract the pedestrian's concern area. The deep loop focus model consists of a Recurrent network, Glimpse network, Emission network, Context network and Classification network as shown in Figure 4.

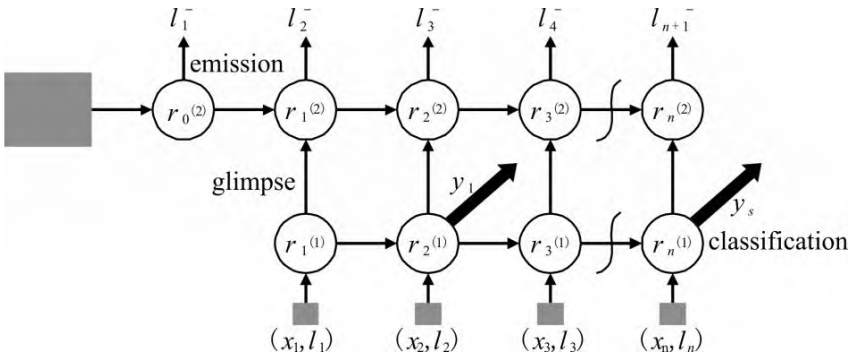

Fig. 4. Deep recurrent focus model

The glimpse network [15] is a nonlinear function that receives the current input of the image patch or glimpse. The job of the glimpse network is to extract a useful set of features from the location l_n of the original visual input. x_n is a feature position, l_n is a position tuple and $l_n = (x_n, y_n)$, then the feature vector g_n is:

$$g_n = G_{image}(x_n|W_{image})G_{loc}(x_n|W_{loc})). \tag{8}$$

Where $G_{image}(x_n|W_{image})$ denotes the output vector of function $G_{image}()$. W_{image} is the weight.

Recurrent networks aggregate information extracted from individual glimpses and combining information in a way that preserves spatial information. At each time step, the glimpsed feature vector g_n from the glimpsed network is provided to the recurrent network as input. The recurrent network consists of two recurrent layers with the nonlinear function R_{re}. r^1 and r^2 are the two outputs of the recurrent layer:

$$r_n^1 = R_{re}(g_n, r_{n-1}^1|W_{r1}). \tag{9}$$

$$r_n^2 = R_{re}(r_n^1, r_{n-1}^2 | W_{r2}). \tag{10}$$

The transmitting network [16] takes the current state of the recurrent network as input to predict where to extract the next image batch of the glimpse network. It acts as a controller, directing attention based on the current internal state of the recurrent network. It consists of a fully connected hidden layer mapping feature vector r_n^2 that repeatedly coordinates the tuple l'_{n+1} layer from the top.

$$l'_{n+1} = E(r_n^2 | W_e). \tag{11}$$

The context network provides the initial state for the recursive network, and its output is used by the transmitting network to predict the first glimpse of the location.

The classification network predicts the tags according to the final eigenvector r_N^1 of the lower recurrent layer. The classification network has a full connection hiding layer and a soft-max output layer.

3.3. Feature Fusion

Feature fusion is to amplify the effect of local salient areas while preserving the overall features. This can not only reduce the influence of the background but also improve the generalization of the model. We discuss the following fusion methods and select the best method as the fusion method in this paper. The selection of fusion method will be discussed in the experimental section.

A. Sum fusion.

y^{sum} represents the result of sum fusion, and $x_{i,j}^G$ and $x_{i,j}^L$ respectively represent the eigenvalues of the global and local features at position (i, j):

$$y^{sum} = x_{i,j}^G + x_{i,j}^L. \tag{12}$$

B. Maximum fusion.

It is same as the sum fusion. y^{max} represents the result of maximum fusion. $x_{i,j}^G$ and $x_{i,j}^L$ respectively represent the eigenvalues of the global and local features at position (i, j):

$$y^{max} = x_{i,j}^G, x_{i,j}^L. \tag{13}$$

C. Concatenation fusion.

This fusion operation superimposes two feature maps on the same spatial position (i, j):

$$y_{i,j,2d}^{cat} = x_{i,j,d}^G; y_{i,j,2d-1}^{cat} = x_{i,j,d}^L. \tag{14}$$

y^{cat} represents the result of series fusion, and d represents the number of channels.

3.4. Contrast Metric Distance Learning

After the fusion of the global features and the local features is completed, Formula (15) is used to execute contrast metric distance learning of the integrated features of different pedestrians.

$$d_M = (x_1^R - x_2^R)^T M (x_1^R - x_2^R). \tag{15}$$

x_1^R and x_2^R represent the feature values after the fusion of pedestrian 1 and pedestrian 2 respectively. A metric distance will be generated after the contrast metric distance learning, and the similarity score can be calculated by using this metric distance and Equation (16):

$$S_{(x_1^R, x_2^R)} = \frac{x_1^R, x_2^R}{|x_1^R| \cdot |x_2^R|} \cdot \frac{1}{d_M}. \tag{16}$$

$S_{(x_1^R, x_2^R)}$ represents the similarity score of x_1^R and x_2^R. After the similarity score is calculated, all similarity scores are sorted, and the one with the highest similarity score is the same pedestrian.

4. Experiments and Analysis

All experiments in this paper are run under the PyTorch framework. Batch size is set to 64. The epoch is set to 200. The initial learning rate is set at 0.001, and decay is performed from the 50th epoch with a decay coefficient of 0.1.

4.1. Dataset

VIPER [17] contains 632 pairs of pedestrian images taken from arbitrary angles under different lighting conditions. The test protocol randomly divides the data set into two halves, 316 pairs for training and the remaining 316 pairs for testing.

CUHK03 [18] is the first personal identification data set large enough for deep learning. It is a challenging data set collected on a university campus and contains 13,164 images of 1360 identities from two camera perspectives.

4.2. Result Analysis

In the experimental part, the results of pedestrian re-recognition based on multi-feature fusion and existing advanced methods are compared and analyzed, and the best fusion method is discussed. Different fusion methods will have different effects, in order to achieve the best fusion effect, this paper conducts experiments on different fusion methods. Based on CUHK03 pedestrian database, this paper conducts experiments on sum fusion, maximum fusion and concatenation fusion respectively, and the experimental results are shown in Table 1.

Table 1. Results using different fusion methods on CUHK03 pedestrian database/%

Fusion method	Rank1	Rank5	Rank10	Rank20
Sum fusion	48.3	80.2	83.9	91.2
Max fusion	49.5	79.3	84.2	89.8
Concatenation fusion	52.2	80.5	89.7	96.1

Table 1 shows the accuracy of sum fusion, maximum fusion and concatenation fusion in CUHK03 pedestrian database. As can be seen from the table, concatenation fusion achieves 52.2%, 80.5%, 89.7% and 96.1% accuracy in Rank1, Rank5, Rank10 and Rank20, respectively, which is 2.7%, 0.3%, 5.5% and 4.9% higher than the other two fusion methods.

By comparing with other methods on the CUHKO3 and VIPER pedestrian databases, this paper further confirms the performance of the pedestrian re-recognition method based on the Cycle GAN-MF. On the CUHK03 pedestrian database, the Cycle GAN-MF method achieves better results than other methods. As shown in Table 2, the Cycle GAN-MF method achieves 52.2%, 80.5%, 89.7% and 96.1% accuracy in Rank1, Rank5, Rank10 and Rank20, respectively, which is 0.8%, -0.8%, 1.2% and 0.2% higher than CMCL.

Table 2. Results using different fusion methods on CUHK03 pedestrian database/%

Fusion method	Rank1	Rank5	Rank10	Rank20
CMCL[19]	51.4	81.3	88.5	95.9
LCNN[20]	49.9	75.8	88.4	93.1
AMR[21]	41.4	72.5	85.1	93.5
Cycle GAN-MF	52.2	80.5	89.7	96.1

On the VIPER pedestrian database, the Cycle GAN-MF achieves higher results than other methods. As shown in Table 3, pedestrian re-recognition method based on Cycle GAN-MF achieves 45.2%, 75.1%, 90.3% and 96.3% accuracy rates in Rank1, Rank5, Rank10 and Rank20, respectively, which is 0.9% higher in Rank5, 4.3% in Rank10 and 2.2% in Rank20 than CMCL.

Table 3. Results using different fusion methods on VIPER pedestrian database/%

Fusion method	Rank1	Rank5	Rank10	Rank20
CMCL[19]	44.3	74.2	86.0	94.1
LCNN[20]	40.7	72.4	85.4	91.4
AMR[21]	34.4	70.3	77.4	91.2
Cycle GAN-MF	45.2	75.1	90.3	96.3

5. Conclusion

Pedestrian re-recognition is a complicated problem, which will be affected by the changes of background, veil, clothing and so on. In order to reduce the impact of background, veil, clothing and other changes on the recognition effect, this paper proposes a pedestrian re-recognition method based on Cycle GAN-MF. The pedestrian re-recognition method proposed in this paper mainly includes two parts: local feature and global feature. In this paper, the significant concern area of pedestrians is proposed through the deep circulation attention model, and then the local features of this area are proposed and integrated with the global features to form a new fusion feature, so as to increase the role of local significant features and reduce the influence of background, veil, clothing and other changes. Through experiments, the pedestrian re-recognition method based on the Cycle GAN-MF has achieved good results on both CUHK03 and VIPER pedestrian databases, and has also played a role in reducing the impact of changes in background, veil, clothing, etc.

6. Conflict of Interest

The authors declare that there are no conflict of interests, we do not have any possible conflicts of interest.

Acknowledgments. None.

References

1. Han K, Zhang N, Xie H, et al. Application of Multi-Feature Fusion Based on Deep Learning in Pedestrian Re-Recognition Method[J]. Mobile Information Systems, 2022.
2. Meng, X., Wang, X., Yin, S. et al. Few-shot image classification algorithm based on attention mechanism and weight fusion. Journal of Engineering and Applied Science. 70, 14 (2023). https://doi.org/10.1186/s44147-023-00186-9.
3. Hong F, Lu C, Tao W, et al. OMNet: Object-Perception Multi-Branch Network for Pedestrian Re-Identification[J]. Big Data Research, 2022, 27: 100302.
4. Karim, Shahid, Geng Tong, Jinyang Li, Akeel Qadir, Umar Farooq, and Yiting Yu. "Current Advances and Future Perspectives of Image Fusion: A Comprehensive Review." Information Fusion, Vol. 90, pp.185-217, February 2023.
5. Yin S. Object Detection Based on Deep Learning: A Brief Review[J]. IJLAI Transactions on Science and Engineering, 2023, 1(02): 1-6.
6. An F P. Pedestrian re-recognition algorithm based on optimization deep learning-sequence memory model[J]. Complexity, 2019, 2019: 1-16.
7. Zhang J, Sun D, Li X, et al. Pedestrian Attitude Estimation and Recognition Algorithm Based on RF Data[J]. Wireless Communications and Mobile Computing, 2022, 2022.
8. S. Yin, L. Wang, M. Shafiq, L. Teng, A. A. Laghari and M. F. Khan, "G2Grad-CAMRL: An Object Detection and Interpretation Model Based on Gradient-weighted Class Activation Mapping and Reinforcement Learning in Remote Sensing Images," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023. doi: 10.1109/JSTARS.2023.3241405.
9. Tao D, Guo Y, Song M, et al. Person re-identification by dual-regularized kiss metric learning[J]. IEEE Transactions on Image Processing, 2016, 25(6): 2726-2738.
10. Liong V E, Lu J, Ge Y. Regularized local metric learning for person re-identification[J]. Pattern Recognition Letters, 2015, 68: 288-296.
11. Shen Y, Zhang F, Liu D, et al. Manhattan-distance IOU loss for fast and accurate bounding box regression and object detection[J]. Neurocomputing, 2022, 500: 99-114.
12. Amali A, Pranoto G T. Manhattan, Euclidean And Chebyshev Methods In K-Means Algorithm For Village Status Grouping In Aceh Province[J]. Journal of Applied Intelligent System, 2022, 7(3): 211-222.
13. Peng Li Asif Ali Laghari, Mamoon Rashid, Jing Gao, Thippa Reddy Gadekallu, Abdul Rehman Javed, Shoulin Yin, "A Deep Multimodal Adversarial Cycle-Consistent Network for Smart Enterprise System," in IEEE Transactions on Industrial Informatics, 19(1), pp. 693-702, 2023. doi: 10.1109/TII.2022.3197201.
14. Wu S, Dong C, Qiao Y. Blind image restoration based on cycle-consistent network[J]. IEEE Transactions on Multimedia, 2022.
15. Sheik S A, Muniyandi A P. Secure authentication schemes in cloud computing with glimpse of artificial neural networks: A review[J]. Cyber Security and Applications, 2023, 1: 100002.
16. Wu J, Xia J, Gou F. Information transmission mode and IoT community reconstruction based on user influence in opportunistic social networks[J]. Peer-to-Peer Networking and Applications, 2022, 15(3): 1398-1416.
17. M*ü*ller P, Schwerhoff M, Summers A J. Viper: A verification infrastructure for permission-based reasoning[C]//Verification, Model Checking, and Abstract Interpretation: 17th International Conference, VMCAI 2016, St. Petersburg, FL, USA, January 17-19, 2016. Proceedings 17. Springer Berlin Heidelberg, 2016: 41-62.
18. Zhang S, Zhang Q, Wei X, et al. Person re-identification with triplet focal loss[J]. IEEE Access, 2018, 6: 78092-78099.

19. Wen X, Feng X, Li P, et al. Cross-modality collaborative learning identified pedestrian[J]. The Visual Computer, 2022: 1-16.
20. Ke X, Lin X, Qin L. Lightweight convolutional neural network-based pedestrian detection and re-identification in multiple scenarios[J]. Machine Vision and Applications, 2021, 32: 1-23.
21. Li C, Yang X, Yin K, et al. Pedestrian re-identification based on attribute mining and reasoning[J]. IET Image Processing, 2021, 15(11): 2399-2411.

Biography

Yongqi Fan is with the Software College of Shenyang Normal University. Research direction is computer application and AI.

Hang Li is with the Software College of Shenyang Normal University. Research direction is computer application and AI.

Botong Sun is with the College of Journalism and Communication, Shenyang Normal University. Research direction is computer application and AI.