IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 7, 2020, accepted July 15, 2020, date of publication July 20, 2020, date of current version September 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3010435

Feature Pyramid Attention Model and Multi-Label Focal Loss for Pedestrian Attribute Recognition

YE LI[1], **FANGYAN SHI**[1], **SHAOQI HOU**[1], **JIPENG LI**[2], **CHAO LI**[3], **AND GUANGQIANG YIN**[3]

[1]School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Glasgow College, University of Electronic Science and Technology of China, Chengdu 611731, China
[3]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

Corresponding author: Guangqiang Yin (yingq@uestc.edu.cn)

ABSTRACT At present, there are many challenges in the field of pedestrian attribute recognition, such as small targets of some attributes, imbalanced samples, and low recognition accuracy of complex samples. In view of the above problems, we improved the model in two perspectives: 1) We proposed Feature Pyramid Attention Model (FPAM). In order to solve the problem that attributes are distributed in different locations in the pedestrian image, FPAM adopted the attention mechanism on the basis of ResNet-50, by which the model's attention could be focused on key areas of the image. As for the difficulty in small targets attributes, we adopted feature pyramid integration strategy; 2) We proposed Multi Label Focal Loss (MLFL). Referring to Binary Cross Entropy Loss Function (CE) and Weight Binary Cross Entropy Loss Function (WCE), we added the weight parameters of samples which are difficult to classify to improve the recognition accuracy, and the rate of convergence was increased. Results show that our proposed method achieves 84.83% mA, 79.37% Accuracy, 87.47% Precision, 86.09% Recall, and 86.77% F1 on PETA dataset.

INDEX TERMS Pedestrian attribute recognition, feature pyramid attention model (FPAM), multi label focal loss (MLFL).

I. INTRODUCTION

Pedestrian attribute recognition aims to obtain the characteristics, such as age, gender, clothing type and other characteristics of the pedestrian from pedestrian image. By mapping unstructured image information into structured information, pedestrian attribute recognition can greatly compress the space needed for video information, and make it more convenient for human to understand the information of the video. The pedestrian attribute recognition is widely used in image retrieval, image generation and person re-identification. There is always a need to precisely find a particular target pedestrian in a mass of video data, especially in the field of person re-identification. However, the traditional manual retrieval method is time-consuming and laborious, and if the attribute recognition can be used to obtain the information of pedestrians, such as black coat, hat and other characteristics, the search speed can be greatly accelerated by comparing these attributes.

At present, there are mainly two problems in the field of the pedestrian attribute recognition. For one thing, the collected pedestrian pictures have such features as changeable attitude, large change of illumination and large change of video resolution since the location of the monitoring device is uncertain, as a result, the recognition accuracy is low. Fig. 1 intuitively shows the existing problem in the pedestrian dataset.

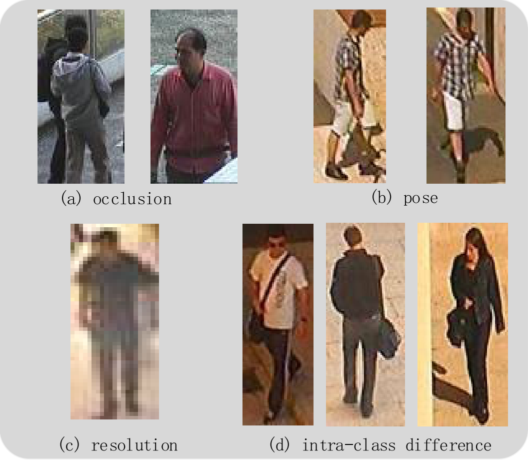

(a) occlusion
(b) pose
(c) resolution
(d) intra-class difference

FIGURE 1. Some existing problems in the pedestrian dataset.

The associate editor coordinating the review of this manuscript and approving it for publication was Lefei Zhang.

For another, the problem in the pedestrian attribute recognition is different from the common classification problems. In attribute recognition, generally dozens of properties of the pedestrian need to be analyzed, and the biggest problem in it is that the attribute labels are unbalanced, for example, most labels of clothing color are white and black, but yellow labels are rare, and this imbalance will lead to the great difficulty in the color learning process of the model. In this paper, we selected 35 kinds of attributes in the PETA dataset and statistically analyzed the data as shown in Fig. 2. It can be clearly concluded that the positive and negative sample ratio of some attributes is about 1:50, and the serious sample imbalance will make it difficult for the model to learn the correct parameters, and eventually lead to a decrease in the accuracy of recognition of severely imbalanced attributes.

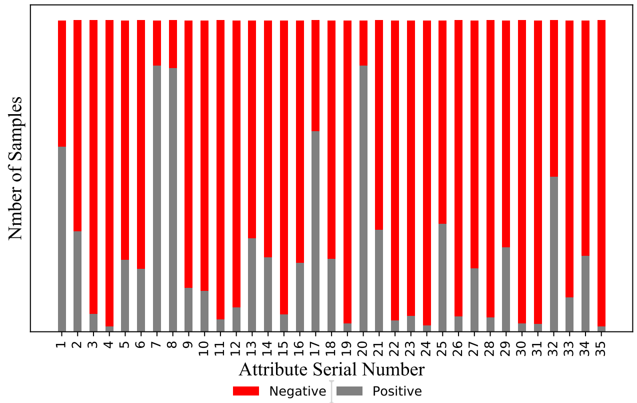


FIGURE 2. Statistical analysis of positive and negative samples of 35 kinds of attributes in the PETA dataset.

Many scholars have delved into the above issues. Li *et al.* [1] proposed a deep learning based multi-attribute recognition model (DeepMAR), but this method treats all the information in the image equally, and does not focus on the regions where the attributes are located, it ignores the recognition of attributes of small target as well. Liu *et al.* [2] introduced the attention mechanism into pedestrian attribute recognition, however this method ignored the influence of the imbalance of samples and the neglected recognition of attributes of small targets.

Previous researches studied some problems in pedestrian attribute recognition, but ignored the problems of great intra-class difference within the attribute and small targets of some attributes, which resulted in the lack of effective improvement of attribute recognition accuracy. For this reason, we proposed Feature Pyramid Attention Model (FPAM), which adds CBAM attention module on the basis of ResNet-50, and integrates multi-level features by taking example by feature pyramid. And in order to solve the problem of multi-attribute classification, we optimized Focal Loss function and proposed Multi Label Focal Loss (MLFL).

In multi-attributes identification of pedestrians, we analyzed the problems of pedestrian attribute recognition. And in order to solve the problem of changeful postures, great intra-class difference within the attribute and small targets of some attributes, we modified the model structure of pedestrian multiple attribute recognition. In addition, we studied the method of feature extraction [3], [4] and unsupervised learning [5]. Then we proposed a new loss function.

The major contributions of this paper are summarized as follows:

- We proposed Feature Pyramid Attention Model (FPAM). In this paper, CBAM attention module was added on the basis of the fundamental network, so that the network could focus on the key parts of the image, which could improve the quality of features extracted by the network and to some extent solve the problem of great intra-class differences within the attribute. And the feature pyramid integration strategy was inserted on the basis of the basic network to integrate the network output feature graph with the superficial feature graph. Through this method, it is easier for the model to identify the attributes of features of small targets.
- We proposed Multi Label Focal Loss (MLFL). We optimized the Focal Loss for multi-attribute recognition task. And the proportion of positive sample of each attribute was used as the hyperparameter to balance the proportion of positive and negative sample loss of each attribute in the loss function. The problems of sample imbalance and problems caused by the varying degree of difficulty of classified samples were solved.

The rest of the paper is organized as follows. Two related work, attribute recognition based on whole body image information and attribute recognition based on information of parts of the body, are introduced in Section II. In Section III, we propose our method, FPAM model and MLFL loss, and describe them in details. The datasets invoked in this paper are presented in Section IV. We implement experiments to prove the advantages of our proposed method in Section V. And a conclusion is drawn in Section VI.

II. RELATED WORK

A. ATTRIBUTE RECOGNITION BASED ON WHOLE BODY IMAGE INFORMATION

Most pedestrian attribute recognition methods are aimed at the whole body image of pedestrians. Layne *et al.* [6], [7] applied manually designed features and SVM [8] to detect pedestrian attributes, and showed the effects of attribute information on pedestrians' reidentification. Li *et al.* [9] detected the appearance attributes of pedestrians' clothes by SVM to help pedestrians reidentify. Deng *et al.* [10] identified pedestrian attributes by cross kernel SVM [11] and used MRF for smooth attribute recognition. Zhu *et al.* [12] introduced Gentle AdaBoost to complete the selection of feature and classifier learning at the same time. Recent neural network methods have also contributed to pedestrian attribute recognition for full graph. Sudowe *et al.* [13] first proposed attribute convolution network (ACN), which learned to identify different attributes through joint training method and could predict multiple attributes at the same time. Li *et al.* [1] also proposed a deep learning based multi-attribute recognition model (DeepMAR), and demonstrated that joint training of multiple attributes can help improve the overall performance of pedestrian attribute recognition. Liu *et al.* [2] introduced the attention mechanism into pedestrian attribute recognition, and demonstrated the improvement of attention mechanism on attribute feature recognition. The HP-net they proposed could capture fine-grained features, and greatly enhanced the feature representation of pedestrian images.

B. ATTRIBUTE RECOGNITION BASED ON INFORMATION OF PARTS OF THE BODY

There are pedestrians attribute recognition based on the information of parts of the body. Bourdev [14] proposed the posture alignment network for depth attribute recognition (PANDA), and they adopted poslets method [15] to obtain the body parts of pedestrians in the images, and at the same time, each body part is separately trained with a non-interfering human attribute classification convolutional neural network to overcome the problem of variable visual angle and occlusion exists in pedestrian attribute recognition. Finally, all the features of the deep network are brought together and an attribute separately trains one SVM classifier. What they focused on is the monitoring scenes rather than natural scenes, and they gave the influence of perspective change, occlusion and body parts on different attributes. Zhu *et al.* [16] introduced a multi-label neural network (MLCNN) based on body parts, which has a predefined attribution-part connection structure to identify pedestrian attributes, but MLCNN cannot solve the difference caused by occlusion and attribute change well.

III. METHOD

We proposed a multi-label attribute recognition model (FPAM) which is on the basis of ResNet-50. The attention module is inserted to strengthen the recognition of area where image attributes locate in. The network applies the feature pyramid strategy, which could integrate the features between superficial and deep layers, to enhance the recognition of attributes of small targets as well. Additionally, MLFL loss function is proposed to solve the problem of imbalance in multi-attribute sample.

A. FEATURE PYRAMID ATTENTION MODEL (FPAM)

1) ATTENTION MODULE

The multi-attribute recognition task is required to accurately identify dozens of properties of pedestrians (PETA data sets marked a total of 65 kinds of attributes) which are located in different locations of pedestrians. Due to the great within-class difference of scale, posture and attributes of pedestrians, the specific position of the attributes on the image is variable. As a result, the key point of pedestrians multiple attribute recognition is how to make the model focus on the information of location where properties exist. If the attached properties labels of strong supervision learning is applied, the difficulty of the generation of datasets will be increased greatly, and hard to be used in practical projects. However, the attention mechanism of weak supervision learning could help the model focus on the interest areas of the image and does not require additional label information, which is a feasible way to improve the performance of attribute recognition without changing dataset. So we embodied CBAM attention module to make the multi-attribute classification model pay attention to the key areas where the attributes exist on the pedestrian image, and to ignore the noise position. By this means the accuracy of multi-attribute recognition could be improved.

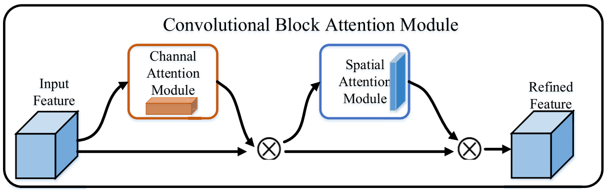
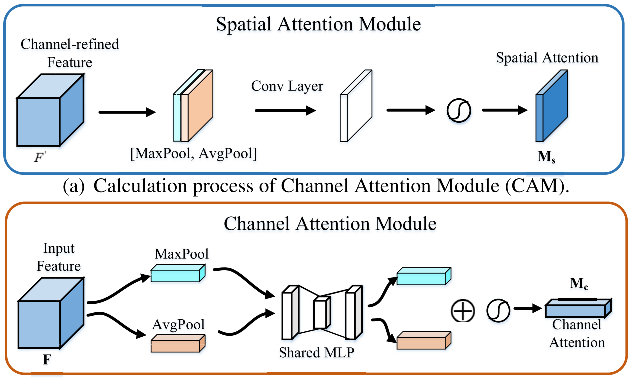


FIGURE 3. Calculation process of Convolutional Block Attention Module (CBAM).



(a) Calculation process of Channel Attention Module (CAM).

(b) Calculation process of Spatial Attention Module (SAM).

FIGURE 4. Detailed calculation process.

Woo *et al.* [17] proposed the convolution block attention module (CBAM), which calculates the feature map from channels and space respectively, then the adaptive feature can be obtained by multiplying the attention feature map with the original feature map. The calculation process is shown in Fig. 3. One feature map of an intermediate layer of the convolution network is defined as $F \in i^{C*H*W}$, and CBAM would calculate the one-dimensional channel feature map $M_C \in i^{C*1*1}$ and the two-dimensional spatial feature map $M_S \in i^{1*H*W}$ respectively, this process is shown in Fig. 4. The calculation formulas of the CBAM are as follows:

$$F' = M_c(F) \otimes F \tag{1}$$

$$F'' = M_s(F') \otimes F' \tag{2}$$

where $\otimes$ stands for element by element multiplication. First multiply the channel feature graph and the input feature map to get F', and then multiply the spatial feature map and F' to get the final CBAM result F''.

The computation of CBAM is so trivial that it can be neglected, and CBAM can be embedded into a variety of convolutional neural networks. Therefore, we embedded the attention module into the ResNet-50 network structure to improve the performance of multi-attribute recognition.

2) FEATURE INTEGRATION

On account of that some small target attributes, such as glasses and masks, take only a small amount of image resolution, so another problem of multi-attribute recognition task is that how to improve the recognition accuracy of small target attributes. Because of the depth of layers, the ResNet-50 network would lose the location information of small targets seriously, making it difficult to obtain the feature information of small target in multi-attribute recognition, for this reason, the multi-attribute recognition model we proposed may have poor recognition performance on small target attributes. To solve the problem, we introduced the feature pyramid strategy, then the integrated feature map was obtained after integrating the position information of small target in superficial layer and the semantic information in deep layer. And the accuracy of identification of small target in the pedestrian attribute recognition could be improved.

Lin *et al.* [18] proposed the feature pyramid networks (FPN) used for target detection. FPN contains both the superincumbent feedforward convolutional neural network approach and bottom-up feature sampling processes, and used the transverse connection to integrate multi-stage feature, details are as Fig. 5. The experiment of [18] proved that multi-scale feature integration enables the target detector to acquire richer multi-scale features, especially for small targets. Consequently, we would introduce the multi-scale feature pyramid strategy.

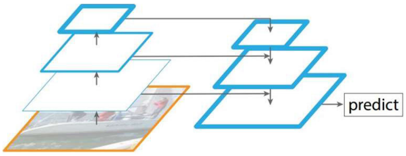


FIGURE 5. Feature integration method used in this paper.

3) FPAM

Combining the advantages of attention module and feature pyramid, we proposed a Feature Pyramid Attention Model (FPAM), as shown in Fig. 6. The original ResNet-50 model consists of four large residual module groups, each of which is composed of 3, 4, 6 and 3 residual modules respectively, and a total 16 residual modules. Every residual module is made up of two 1*1 convolutions and one 3*3 convolution. We first inserted CBAM into each residual module of ResNet-50 to constitute the new Residual CBAM. The detailed structure of Residual CBAM is shown in Fig. 7(b). The original ResNet-50 only adopted the feature map of the forth module group, however, referring to the feature integration strategy of target detection, the feature map in our research should be extracted from the third and the forth Residual CBAM groups to integrate the features and improve the representation of these features. See Fig. 6 for the specific

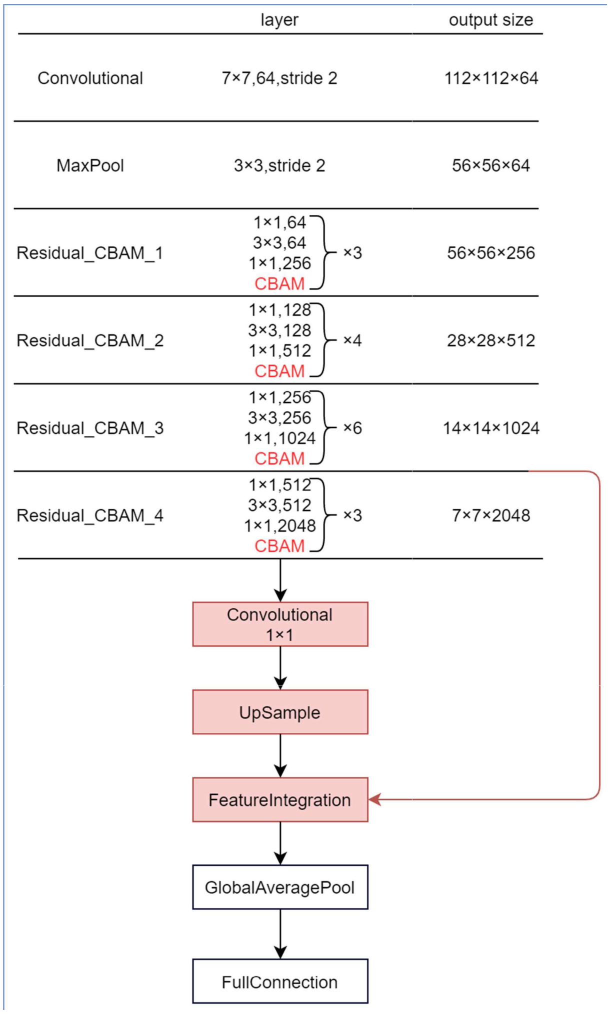


FIGURE 6. The whole structure of Feature Pyramid Attention Model (FPAM). The red sections represent the different sections from ResNet-50.

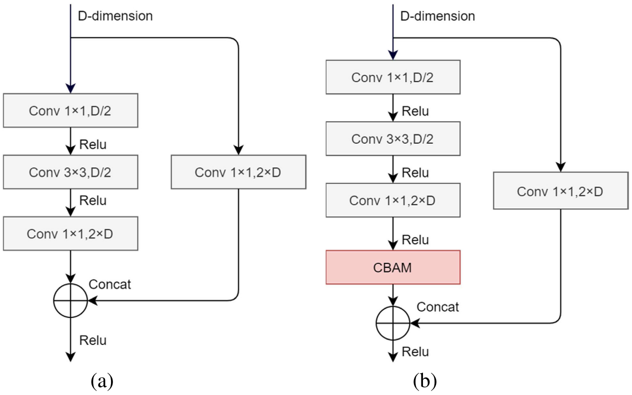


FIGURE 7. Comparison of residual module. (a) The original Residual Module. (b) The Residual CBAM Module proposed in this paper.

structure. Finally, we obtained the outputs and predicted its attributes. Compared with the original ResNet-50 network, there is only a small increase in computation in the network we proposed, which could be neglected.

B. MULTI LABEL FOCAL LOSS (MLFL)

Generally, for multi-attribute classification, the output layer adopts sigmoid function to unify the output between 0 and 1, and the loss function adopts the binary cross entropy loss function, which is defined as follows:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L} y_{ij} * log(p_{ij}) + (1 - y_{ij}) * log(1 - p_{ij}) \tag{3}$$

$$p_{ij} = \frac{1}{1 + e^{-x_{ij}}} \tag{4}$$

Among them, N represents the number of samples, L represents the number of attributes, y_{ij} represents the real label of the j-th attribute of the i-th sample, and p_{ij} represents the predicted value of the network of the j-th attribute of the i-th sample, whose value is distributed between 0 and 1 after applying the sigmoid function.

Formula (3) is regarded as the loss function of multi-attribute classification. Although all attributes are fairly considered, in the multi-attribute classification, the ratio of positive and negative samples of each attribute in the dataset is not balanced in view of the large number of attributes. As the PETA pedestrian attribute dataset in Introduction, it can be seen that there is a large difference in the proportion of positive and negative samples, which will lead to the gradient of the loss function being dominated by the attributes with large proportions, hence making the trained network have poor performance on the recognition of attributes with small proportions. Therefore, Li *et al.* [1] optimized the loss function and adopted the weighted binary cross entropy loss function, which set higher weight of loss function for attribute samples with less proportion and reduce the weight of attribute samples with higher proportion. The weighted binary cross entropy loss function formula is as follows:

$$Loss_w = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L} e^{w} * y_{ij} * log(p_{ij}) + e^{1-w} * (1 - y_{ij}) * log(1 - p_{ij}) \tag{5}$$

$$w = 1 - \frac{N_j}{N_{all}} \tag{6}$$

where w is the proportion of negative samples of the attribute to all samples, N_j is the number of positive samples of the attribute in the training set, and N_{all} represents the number of training attributes in the training set (generally, the number of training set samples).

The weighted loss function considers the proportion imbalance between positive and negative samples, but does not consider the problems caused by the different levels of difficulty to classify the samples. In the later stage of training, most of the simple samples can be classified correctly, and only a small number of the complex samples might be classified incorrectly. However, since simple samples are the majority of the value of the loss function, so that the direction of gradient is dominated by simple samples, but the classification accuracy cannot be improved by simple samples, thus we refers to [19], introducing the weight of simple attributes and complex attributes, increasing the proportion of the loss function value of difficult attributes in the overall loss function, and guiding the network to favor the complex samples to improve the classification accuracy of the network for the complex samples.

The loss function proposed in this paper is as follows:

$$L_{MLF} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L} e^{w} * y_{ij} * (1 - p_{ij})^{\gamma} * log(p_{ij}) + e^{1-w} * (1 - y_{ij}) * p_{ij}^{\gamma} * log(1 - p_{ij}) \tag{7}$$

where N represents the number of samples, L represents the number of attributes, y_{ij} represents the real label of the j-th attribute of the i-th sample, and p_{ij} represents the predicted value of the network of the j-th attribute of the i-th sample, whose value is distributed between 0 and 1 after applying the sigmoid function. w is the proportion of negative samples of the attribute to all samples. γ is the weight coefficient of loss function for the complex samples, and the range is set from 0 to 5. By testing 1, 2, 3, 4 and 5, it can be proved that the best performance is achieved when γ is set as 1. In this paper, it is set as 1. It can be seen that when $\gamma = 1$, if the real label is 1, the loss function of simple samples with a predicted value of 0.95 will be reduced by 20 times, while that of complex samples with a predicted value of 0.2 will only be reduced by 0.8 times. Therefore, this loss function can effectively increase the proportion of complex samples in the loss function.

The loss function proposed in this paper is different from focal loss in [19], which used fixed hyperparameters to balance the positive and negative samples. However, because of the different proportions of positive and negative samples in the multi-attribute classification, it is not suitable to use fixed parameters to balance the ratio of these two kinds of samples. Therefore, we considered the proportion of positive sample of each attribute as the hyperparameters to balance the proportion of positive and negative sample loss to the loss function, and proposed a focal loss function suitable for multi-attribute classification, which is called Multi Label Focal Loss (MLFL).

IV. DATASET

A. PETA DATASET

We used the PETA data set in this paper, which contains a total of 19000 images with resolution ranging from 17*39 to 169*365 pixels. Each pedestrian has 61 dichotomous attributes and 4 multi-category attributes. According to the practice, 35 attributes with balanced positive and negative proportion were selected for training, and 9,500 images were randomly divided into the training set, and 1900 images were used for the verification set and 7600 images for the test set. The experiments in this section combined the training set and the verification set, and a total of 11,400 images were used for the training.

B. EVALUATION PROTOCOLS

Mean accuracy (mA) is the most commonly used standard evaluation of attribute recognition algorithm. For the imbalance of attributes in multi-attribute recognition, the recognition accuracy of positive and negative examples of each attribute will be calculated respectively, and then their average value will be taken as the recognition accuracy of this attribute in order to prevent the model from being biased towards those positive samples with higher proportion. The specific formula is as follows:

$$mA = \frac{1}{2N}\sum_{i=1}^{L}\left[\frac{TP_i}{P_i}+\frac{TN_i}{N_i}\right] \tag{8}$$

N represents the number of samples, and L is the number of attributes. P_i and TP_i represents the number of positive samples of an attribute and the number of correctly identified positive samples respectively, while N_i and TN_i represents the number of negative samples of an attribute and the number of correctly identified negative samples respectively.

However, mA is a label-based evaluation standard, which treats each attribute independently and ignores the correlation between attributes. Therefore, Li *et al.* [20] proposed an example-based evaluation standard, which is more consistent with human's prediction of pedestrian attributes. The example-based evaluation standard includes: accuracy, precision, recall rate and F1 value. The detailed definitions are as follows:

$$Acc_{exam} = \frac{1}{2N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|Y_i \cup f(x_i)|} \tag{9}$$

$$Prec_{exam} = \frac{1}{2N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|f(x_i)|} \tag{10}$$

$$Rec_{exam} = \frac{1}{2N}\sum_{i=1}^{N}\frac{|Y_i \cap f(x_i)|}{|Y_i|} \tag{11}$$

$$F1 = \frac{2 * Prec_{exam} * Rec_{exam}}{Prec_{exam} + Rec_{exam}} \tag{12}$$

where N represents the number of samples, Y_i is the set of real label of the i-th sample, $f(x_i)$ is the predicted attributes set of the i-th sample, and $|\cdot|$ represents the number of attributes in the set.

V. EXPERIMENTS

To evaluate the effectiveness of our approach and study the impact of various factors on person attribute recognition, we conduct several groups of experiments on the PETA dataset. Subsection V-A gives an introduction of basic experiment setting. Subsection V-B compares the effects of attention module and feature integration. Subsection V-C conducts further researches on the loss function. Subsection V-D compares our approaches with existing ones.

A. EXPERIMENT SETTINGS

In this paper, we build the basic network by PyTorch. And we used parameters of ResNet-50 which were pre-trained on the ImageNet to initialize the FPAM. The formula (7) was used as the loss function to optimize the network. Generally, it is necessary to preprocess the data before training the model, which can expand the dataset and improve the generalization ability of the network. In this paper, the pre-processing process of experimental data includes adjusting the size of the image to 224*224 to meet the requirements of the network for input images, and performing horizontal inversion and subtracting average value to expand the dataset. Stochastic weight we used was set to 0.9, the batch size was set to 32, and the initial learning rate was 0.001. A total of 150 epochs were trained, and the learning rate was adjusted to one tenth of the original value after every 50 epochs. Table 1 illustrates the hardware environment and software version of the experiment in this paper.

TABLE 1. Hardware and software version in our experiments.

System Version	Ubuntu 18.04.1
CPU Model and Memory Size	2* Intel(R) E5-2678 v3 @ 2.50GHz 128G
GPU Model and Memory Size	2* Nvidia Titan RTX, 24G
Deep Learning Framework	Pytorch 1.3.1
Driver	440.44
Others	CUDA 10.2

B. EFFECTIVENESS OF FPAM

The FPAM proposed in this paper is based on ResNet-50 with attention module and feature integration structure. In this section, by using the parameters in the above experimental settings, we studied the influence of each module on the accuracy of attribute recognition on the PETA dataset and PA-100K dataset. In order to fully prove the effectiveness of FPAM network, we employed CE, WCE and MLFL loss functions respectively. The experimental comparison results are as follows.

It can be seen from Table 2 that the mA increased by 1% by using CE loss function. By using WCE loss function, the mA increased by 1.28%. And the mA improved by 1.04% by using MLFL loss function. It can be seen from Table 3 that the mA increased by 2.76% by using CE loss function. By using WCE loss function, the mA increased by 0.62%. And the mA improved by 1.59% by using MLFL loss function. According to all kinds of indexes, FPAM network has a significant improvement in pedestrian attribute recognition.

C. EFFECTIVENESS OF MLFL

1) COMPARISON WITH OTHER LOSS FUNCTION

We did not only propose the FPAM network structure, but also optimized the loss function. In this section, the effect

TABLE 2. **Comparisons between different feature extraction model on PETA dataset. FI is feature integration, CBAM is attention module, CE means Cross Entropy loss, WCE means Weighted Cross Entropy loss, and MLFL means Multi Label Focal Loss.**

Network	mA	Accuracy	Precision	Recall	F1
ResNet50+CE	82.93	78.32	87.89	84.26	86.04
ResNet50+CBAM+CE	83.29	78.93	88.35	84.78	86.73
ResNet50+FI+CE	83.69	78.59	87.26	85.16	86.20
ResNet50+CBAM+FI+CE	83.93	79.43	88.03	85.74	86.87
ResNet50+WCE	83.28	78.60	87.41	85.04	86.21
ResNet50+CBAM+WCE	84.33	79.55	87.91	85.99	86.95
ResNet50+FI+WCE	84.40	79.01	86.94	86.07	86.50
ResNet50+CBAM+FI+WCE	84.56	79.26	87.41	86.04	86.72
ResNet50+MLFL	83.79	78.97	87.60	85.39	86.48
ResNet50+CBAM+MLFL	84.09	79.09	87.97	85.42	86.63
ResNet50+FI+MLFL	84.28	78.61	86.88	85.48	86.17
ResNet50+CBAM+FI+MLFL	84.83	79.37	87.47	86.09	86.77

TABLE 3. **Comparisons between different feature extraction model on PA-100K dataset. FI is feature integration, CBAM is attention module, CE means Cross Entropy loss, WCE means Weighted Cross Entropy loss, and MLFL means Multi Label Focal Loss.**

Network	mA	Accuracy	Precision	Recall	F1
ResNet50+CE	75.00	76.44	86.77	83.08	85.62
ResNet50+CBAM+CE	75.89	77.38	87.27	83.75	85.82
ResNet50+FI+CE	76.1	77.81	86.98	83.4	85.53
ResNet50+CBAM+FI+CE	77.76	77.94	89.07	84.33	86.33
ResNet50+WCE	78.91	76.04	85.21	85.55	85.28
ResNet50+CBAM+WCE	78.74	76.51	85.82	85.06	85.58
ResNet50+FI+WCE	79.26	76.95	86.25	85.35	85.92
ResNet50+CBAM+FI+WCE	79.53	76.89	86.89	85.75	85.97
ResNet50+MLFL	78.35	76.36	85.78	85.29	86.05
ResNet50+CBAM+MLFL	78.68	77.01	85.98	85.94	85.51
ResNet50+FI+MLFL	79.42	76.84	85.28	86.12	85.53
ResNet50+CBAM+FI+MLFL	79.94	77.44	86.35	86.14	86.65

of MLFL loss function was studied on the PETA data set by using the parameters in the above experimental settings. We adopted and compared ResNet-50 and FPAM network structures with CE, WCE, and MLFL respectively. The experimental results are as follows.

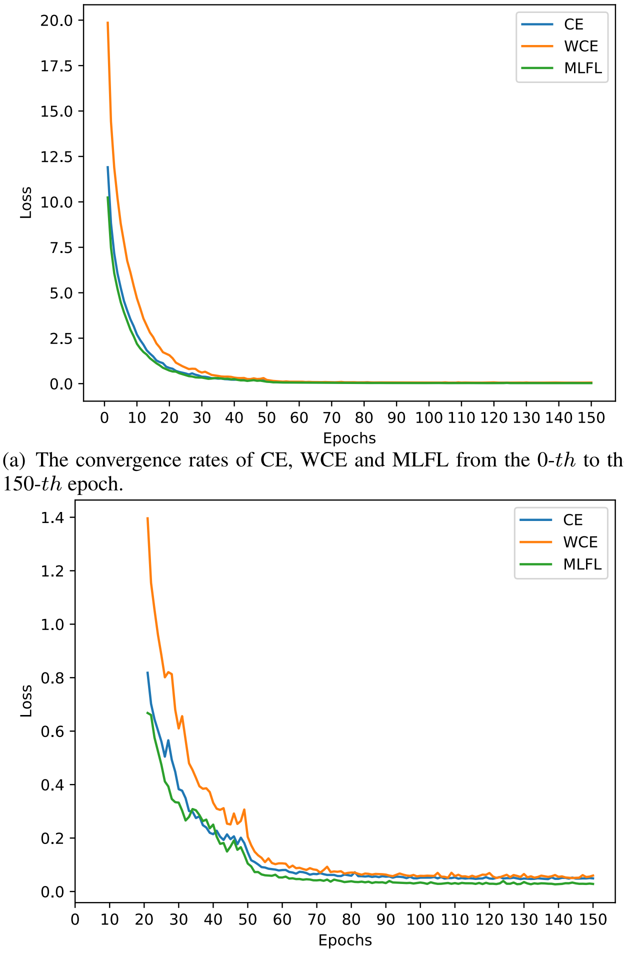
It can be seen from Table 4 that in ResNet-50 network, adopting MLFL loss function can improve mA by 0.86%; In FPAM network, MLFL loss function can improve mA by 0.90%. According to all kinds of indexes, MLFL loss function can significantly improve pedestrian attribute recognition.

TABLE 4. **Comparisons between our loss function and others on ResNet-50 and FPAM.**

Network	mA	Accuracy	Precision	Recall	F1
ResNet50+CE	82.93	78.32	87.89	84.26	86.04
ResNet50+WCE	83.28	78.60	87.41	85.04	86.21
ResNet50+MLFL	**83.79**	**78.97**	**87.60**	**85.39**	**86.48**
FPAM+CE	83.93	79.43	88.03	85.74	86.87
FPAM+WCE	84.56	79.26	87.41	86.04	86.72
FPAM+MLFL	**84.83**	**79.37**	**87.47**	**86.09**	**86.77**

2) CONVERGENCE RATE

Fig. 8 demonstrates the convergence rates of CE, WCE and MLFL respectively. Firstly, we compared the rate of descent. MLFL is the fastest, CE is the second, and WCE is the slowest. The reason for this phenomenon is that the main factor of convergence rates is the complex samples, and this kind of samples require more iterations to allow the model to learn the parameters. The MLFL proposed in this paper increased the proportion of loss function of hard samples compared with CE and WCE loss function, as a result, the model could learn complex samples first, which leads to the acceleration of the convergence rates directly. Secondly, it can be seen from the final loss value of these three loss functions that the loss value of MLFL loss function is the minimum, which proved that the MLFL loss function enables the model to learn better parameters.



(a) The convergence rates of CE, WCE and MLFL from the 0-th to the 150-th epoch.

(b) The convergence rates of CE, WCE and MLFL from the 20-th to the 150-th epoch.

FIGURE 8. **Comparisons of convergence rates of different loss functions on PETA dataset.**

D. COMPARISON WITH STATE-OF-THE-ART METHODS

The experimental results of pedestrian multi-attribute recognition network proposed in this paper on the PETA dataset are as follows: the mean accuracy (mA) based on labels is 84.83%, the accuracy based on examples is 79.37%, the precision is 87.47%, the recall rate is 86.09%, and the F1 value is 86.77%. The comparison results with other algorithms are

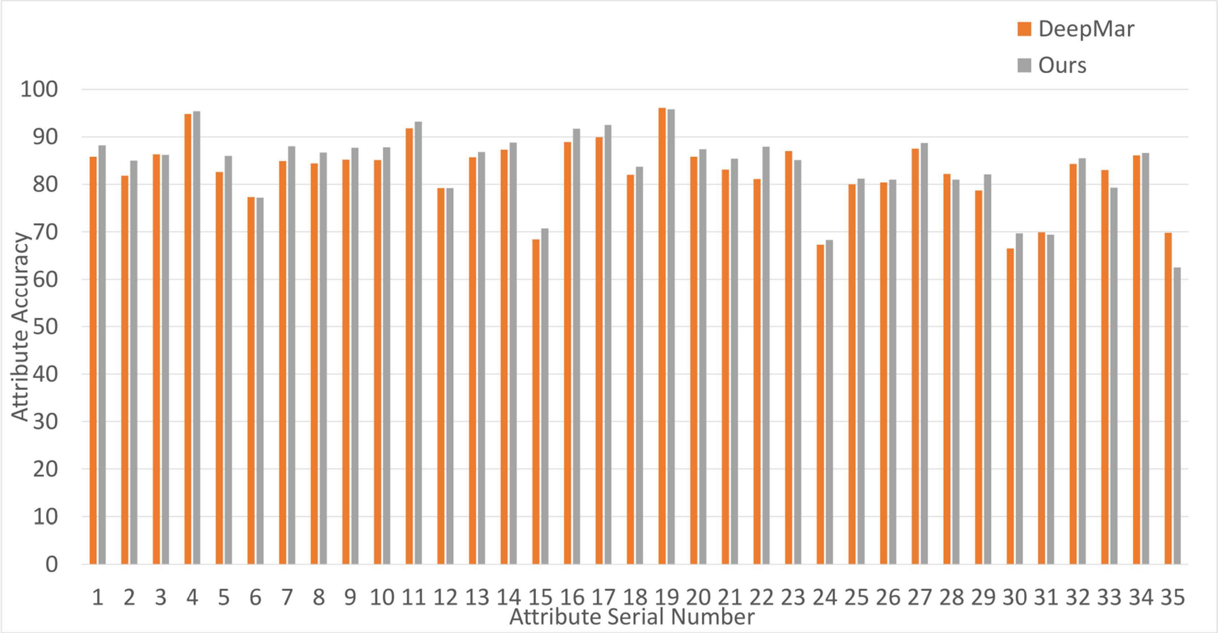


FIGURE 9. Comparison of recognition accuracy of each attribute between the method proposed in this paper and Deep-Mar on PETA dataset.

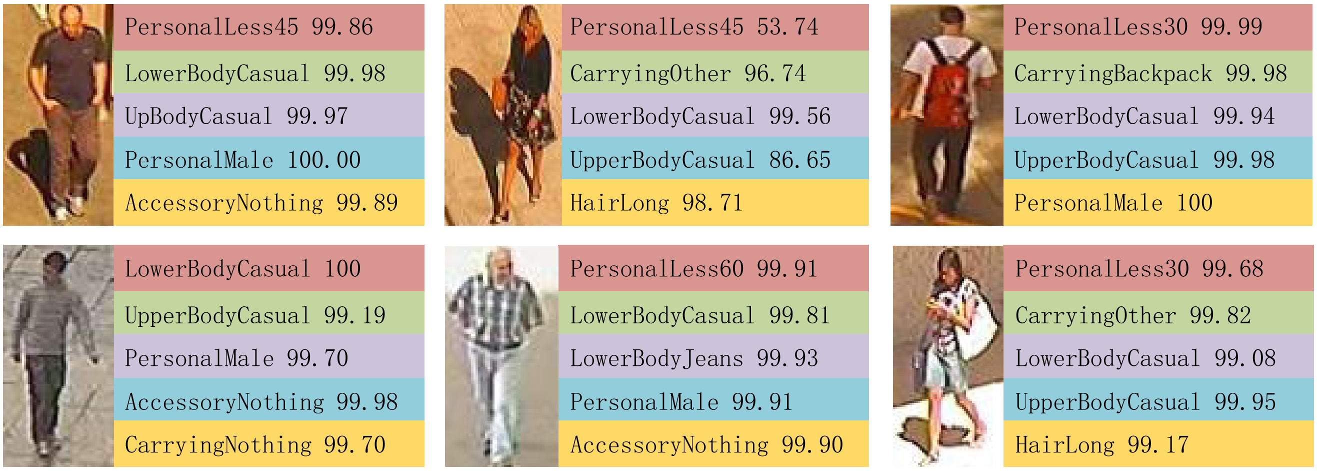


FIGURE 10. Visualization of pedestrian attribute recognition on PETA dataset.

shown in Table 4. The first seven items in the Table are the benchmark experimental results on the PETA dataset listed in [2], and the last item in the Table is the experimental results of the pedestrian multi-attribute recognition model proposed in this paper.

The first three algorithms all used SVM as classifiers, and the difference is that the first algorithm used traditional manual design features, while the second and third algorithms used Caffe Net to extract pedestrian attribute features. Both ACN and Deep-Mar algorithms used convolutional neural network to identify pedestrian attributes. M-net and HP-net are from [2]. The former only used Inception Net as the basic network, while the latter combined with the attention mechanism and inserted the attention mechanism network on the basis of M-net.

TABLE 5. Comparison with the state-of-the-art on PETA dataset.

Method	mA	Accuracy	Precision	Recall	F1
ELF-mm	75.21	43.68	49.45	74.24	59.36
FC7-mm	76.65	45.41	51.33	75.14	61.00
FC6-mm	77.96	48.13	54.06	76.49	63.35
ACN [13]	81.15	73.66	84.06	81.26	82.64
Deep-Mar [1]	82.6	75.07	83.68	83.14	83.41
M-net [2]	80.58	75.68	84.81	82.9	83.85
HP-net [2]	81.77	76.13	84.92	83.24	84.07
Ours	**84.83**	**79.37**	**87.47**	**86.09**	**86.77**

It can be seen from Table 5 that the model proposed in this paper are ahead of other algorithms, which is attributed to the excellent feature extraction ability and generalization ability of ResNet-50 network that enables the algorithm to learn proper feature representation on PETA dataset, and the use

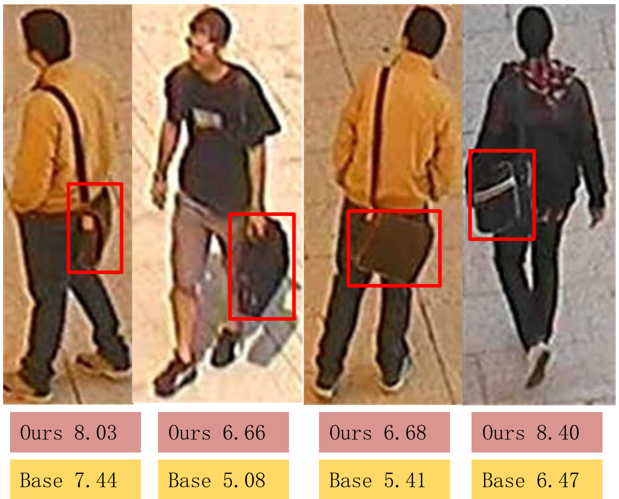


FIGURE 11. **Visualization of recognition accuracy of messenger bag. The positions and shapes of messenger bags varied greatly in the pictures. 'Ours' stands for the proposed method. 'Base' stands for traditional method.**

of CBAM attention module is also a key factor, which makes the multi-attribute recognition network in this paper focus on the image position where the attributes exist and ignore the irrelevant position. Otherwise, through the improvement of the feature integration strategy and loss function, we enabled the model focus on the attributes of small target and the attributes with a significant imbalance proportion of positive and negative samples. Fig. 9 demonstrates the comparison of recognition accuracy of each attribute between the model proposed in this paper and Deep-Mar, it can be seen that the recognition accuracy of most attributes of the model proposed in this paper is higher than that of Deep-Mar model, which could prove that the improvement of the model in this paper is effective. Fig. 10 shows several examples of pedestrian multi-attribute recognition on PETA data set. Fig. 11 shows the comparison between the proposed method and the traditional method in the accuracy of messenger bags recognition.

VI. CONCLUSION

Aiming at problem of pedestrian multi-attribute recognition based on global image, we proposed Feature Pyramid Attention Model (FPAM) which combined attention module and feature integration. And we proposed Multi Label Focal Loss function to ensure that the loss function could be used in multi-attribute recognition. Eventually, we compared and analyzed the experimental results to prove the effectiveness of the model and loss function proposed in the paper. In the future, we plan to study that how to solve the problem of low resolution in pedestrian attribute recognition and how to apply the method of image recognition to fault diagnosis and fault tolerant control [21], [22].

REFERENCES

[1] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 111–115.

[2] X. Liu *et al.*, "HydraPlus-Net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 350–359, doi: 10.1109/ICCV.2017.46.

[3] Z. Wang, B. Du, and Y. Guo, "Domain adaptation with neural embedding matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2387–2397, Jul. 2020.

[4] Z. Liu, J. Wang, G. Liu, and L. Zhang, "Discriminative low-rank preserving projection for dimensionality reduction," *Appl. Soft Comput.*, vol. 85, Dec. 2019, Art. no. 105768.

[5] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognit.*, vol. 102, 2020, Art. no. 107173.

[6] R. Layne, M. T. Hospedales, and S. Gong, "Towards person identification and re-identification with attributes," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 402–412.

[7] R. Layne, T. Hospedales, and S. Gong, "Person re-identification by attributes," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–11.

[8] J. C. Platt, "A fast algorithm for training support vector machines," *J. Inf. Technol.*, vol. 2, no. 5, pp. 1–28, 1998.

[9] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan, "Clothing attributes assisted person reidentification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 5, pp. 869–878, May 2015.

[10] Y. Deng, L. Ping, C. L. Chen, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 789–792.

[11] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[12] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Z. Li, "Pedestrian attribute classification in surveillance: Database and evaluation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 331–338.

[13] P. Sudowe, H. Spitzer, and B. Leibe, "Person attribute recognition with a jointly-trained holistic CNN model," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 87–95.

[14] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. D. Bourdev, "PANDA: Pose aligned networks for deep attributemodeling," *CoRR*, vol. abs/1311.5591, 2013.

[15] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 1543–1550.

[16] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li, "Multi-label CNN based pedestrian attribute learning for soft biometrics," in *Proc. Int. Conf. Biometrics (ICB)*, Phuket, Thailand, May 2015, pp. 535–540.

[17] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Convolutional_block_attention.pdf," in *Proc. ECCV*, 2018.

[18] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[20] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1575–1590, 2019.

[21] Y. Wu, B. Jiang, and N. Lu, "A descriptor system approach for estimation of incipient faults with application to high-speed railway traction devices," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 10, pp. 2108–2118, Oct. 2019.

[22] Y. Wu, B. Jiang, and Y. Wang, "Incipient winding fault detection and diagnosis for squirrel-cage induction motors equipped on CRH trains," *ISA Trans.*, vol. 99, pp. 488–495, Apr. 2020.

[23] Y. Li, G. Yin, S. Hou, J. Cui, and Z. Huang, "Spatiotemporal feature extraction for pedestrian re-identification," in *Proc. Int. Conf. Wireless Algorithm*, 2019, pp. 188–200.

[24] Y. Wu and K. He, "Group normalization," *Int. J. Comput. Vis.*, vol. 128, no. 3, pp. 742–755, Mar. 2020.

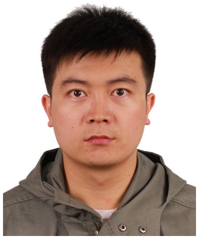
[25] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, no. 3, pp. 333–359, Dec. 2011.

[26] T. Li, C. Zhang, and S. Zhu, "Empirical studies on multi-label classification," in *Proc. 18th IEEE Int. Conf. Tools with Artif. Intell. (ICTAI)*, Nov. 2006, pp. 86–92.

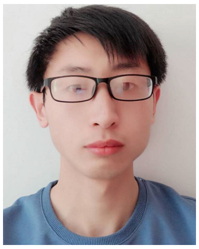
[27] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proc. 6th Int. Conf. Comput. Vis.*, Jan. 1998, pp. 555–562.

[28] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

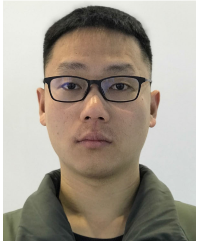
[29] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[31] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[34] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002, doi: 10.1109/TPAMI.2002.1017623.
[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, 2015, pp. 1–11.
[36] Y. Li, H. Chen, C. L. Chen, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 684–700.
[37] P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. Brit. Mach. Vis. Conf.*, London, U.K., 2009, pp. 1–11.



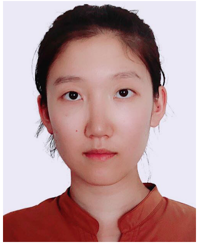
YE LI received the bachelor's degree from the School of Information Science and Technology, Hainan University, in 2014. He is currently pursuing the Ph.D. degree with the Department of Electronic and communication engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan. In 2014, he went to Taiwan Ilan University as an Exchange Student. His research interests include computer vision and deep learning.



FANGYAN SHI received the bachelor's degree from the School of Mechatronics Engineering, University of Electronic Science and Technology of China, in 2017, where he is currently pursuing the degree with the School of Information and Communication Engineering. His research interest includes computer vision.



SHAOQI HOU received the bachelor's degree from the School of Science, Qingdao University of Technology, in 2017. His previous work focused on mathematical modeling and its applications. In 2018, he was recommended for further studies at the University of Electronic Science and Technology of China. His current research interest includes computer vision.



JIPENG LI is currently pursuing the bachelor's degree with the University of Electronic Science and Technology of China, Chengdu, Sichuan.

CHAO LI received the bachelor's degree in financial engineering from the Central University of Economic and Finance of China, in 2017, and the master's degree in applied economics from the University of Bath, U.K., in 2019. His current research interest includes interdiscipline fields, such as person re-identification and financial robot.

GUANGQIANG YIN is currently a Professor with the University of Electronic Science and Technology of China (UESTC). His research interests include computer-vision-related artificial intelligence techniques and applications, and computer modeling of properties of condensed matter.