A Transfer Learning Method with Multi-feature Calibration for Building Identification

Jiafa Mao
School of Computer Science and Technology
Zhejiang University of Technology
Hangzhou, China
maojiafa@zjut.edu.cn

Linlin Yu*
**Corresponding author*
School of Computer Science and Technology
Zhejiang University of Technology
Hangzhou, China
1311776752@qq.com

Hui Yu
School of Computer Science and Technology
Zhejiang University of Technology
Hangzhou, China
hui_yu@zjut.edu.cn

Yahong Hu
School of Computer Science and Technology
Zhejiang University of Technology
Hangzhou, China
huyahong@zjut.edu.cn

Weiguo Sheng
Department of Computer Science
Hangzhou Normal Uni.
Hangzhou, China
w.sheng@ieee.org

***Abstract*—Traditional building identification methods are difficult for extracting the specific information of various buildings. In this paper, A transfer learning method with multi-feature calibration is proposed for building identification. Our model is based on the pre-training and fine-tuning framework of transfer learning. First, a CNN-based feature extractor, pre-trained by ImageNet, is adopted to extract features, then flatten the feature maps and feed it to a fully-connected network for image classification. This basic transfer learning model can correctly identify 81.2% of test samples. Further, a multi-feature calibration method is proposed. By defining the features of multi-functional buildings artificially, the feature vectors via the extractor are more representative and it can be efficiently applied on some small-sample data sets. We use a self-made building data set to test our methods. The experimental results show that the recognition accurate rate of the model with multi-feature calibration attains to 91.9%.**

Keywords—Building Identification, Transfer Learning, Bottleneck Layer, Multi-Feature Calibration

I. Introduction

A. Motivation

With the increasing usage of information and the advent of the digital age, images become more valuable and widely used. At the same time, some challenges occur in the field of image processing, such as how to use computers to automatically divide images into different categories in a way that people can understand, and how to classify and manage a large number of images quickly and effectively. Building classification is an important branch under the scene classification, and it is a key step in solving the problems of building images retrieval. In recent years, with the rapid development of computer hardware, software and the Internet, people can easily upload and share photos taken with social tools. How to enable computers to automatically understand images and effectively classify, manage and efficiently use these image resources has become a major problem for academics and industry.

Currently, significance of building classification are in two aspects.1) By improving the classification accuracy, image browsing and image retrieval based on image content can be realized. By using building classification, computers can classify image into specific building categories according to the recognized picture semantics, and realize image management and retrieval easily. 2) Through the classification of images, the workload of architects can be relived. Accurate classification of design models and complete design dataset can improve the output of the building design effectively.

B. Related Works

Building identification is one of the research hotspots in the field of computer vision and pattern recognition. It enables people to quickly obtain the location, name, description and other related information of the building according to the image, and has important application value in the fields of building positioning, building marking, etc. How to calibrate effectively is a key issue in building identification. Building identification is very similar to object identification, and we can start with color texture [1,2] and point of interest [3].

Yang [4] uses the SIFT feature extraction method to obtain key information points, and then integrates the key points of each building into a code word. These code words are used to build the codebook which is used to identify the building.

Guo Z [5] proposed a village building recognition based on HRRS (satellite remote sensing image) images, which integrates part of the feature extractor of each individually optimized model, and connects them into a combined network based on multi-scale feature learning methods. Its model has improved accuracy compared to the CNN optimized by individuals, but it has strict requirements on the resolution of the data set, and the recognition content is limited to the division of village buildings. Meanwhile in the field of remote sensing image building recognition. Tremblay-Gosselin J [6] proposed a method combining automatic double-seed region expansion algorithm and binary support vector machine for building detection. This method requires users to manually label interest points (buildings in the image) Non-interest points (such as streets or vegetation). At the same time, the entire image is radiated by the region growth algorithm, and the resulting regions are merged.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE I. NEURAL NETWORK MODEL COMPARISION

Network model	Network depth	Additional technology
Alex-Net [13]	5	-
ZF-Net [14]	5	Increased network width.
VGG_CNN_M/S/F[15]	5	Increased network width.
VGG [16]	16-19	3x3 small convolution kernel parameter reduction.
Inception v1[17]	22	Inception module and auxiliary classifier.
Inception v2/v3[18]	46	Improved Inception module and convolution kernel decomposition
Residual-Net [19]	50/101/152	Residual connection
Inception v4[20]	175	Combined with Residual connection and Inception module

Finally, a support vector machine is used to build a training machine to distinguish between interesting and uninteresting parts. However due to the uniform sampling method, It leads to large memory and time costs.

Different from simply calibrating buildings from remote sensing images. Moun *et al*. [7,8,9] take pictures of buildings through handheld devices. GPS location and compass information are collected to reduce search area. The flood filling algorithm is used to cut the outline of the building by analyzing the areas with the same color in the field. Finally, the SIFT features of the building outline are extracted and matched with the data in the building picture database. The most similar building image and its associated description are returned according to the match rank. The whole process of building positioning and recognition is reasonable, but the matching accuracy is influenced by the amounts of images in the matching database since it is difficult to make judgments on building pictures not in the database. Moreover, during image matching, a large number of feature points are inspected. Because it took a lot of time to calculate the similarity, the time complexity is beyond our acceptance. Although its great recognition accuracy, the storage cost and time cost are very large because the uniform sampling seeds are used instead of the automatic selection method.

In the previous image classification algorithms, most of the feature extraction is artificially designed. The basic features of the image are obtained through gradient learning, but there are still large representation differences among the advanced features of the image. The convolutional neural network can extract features from the special direction to achieve end-to-end image feature extraction and classification. It can achieve more abstract features, but at the same time, training complex neural networks requires a lot of annotation data. Meanwhile, it's difficult to collect millions of annotation data, and it costs a lot of time to train a complex neural network. Based on the influence of the number of features on the recognition performance of buildings, a feature extraction based on convolutional neural network with transfer learning is proposed in this paper. We propose a multi-feature calibration technology for feature segmentation, and the network model is fine-tuned. The input of the last layer defined as Bottlenecks is obtained through the Inception network model. Then use Bottlenecks to retrain the final classifier and perform simulation experiments on the Spyder visualization platform. Experiments show that the technical scheme has higher recognition accuracy and faster recognition speed, which provides reference for quickly and effectively calibrating buildings from building images.

II. TRANSFER LEARNING

A. Convolutional Neural Network

The convolutional neural network learns the characteristics of the original image through the cooperation of the convolutional layer and the downsampling layer. Combined with the classical BP algorithm [10] to adjust the parameters and complete the weight update, the BP network update weight formula as:

$$\omega(t+1)=\omega(t)+\eta\delta(t)x(t) \tag{1}$$

Where *x(t)* is the output of the neuron, $\delta(t)$ indicates the error term of the neuron, η representing learning rate. The network structure of the convolutional layer in the convolutional neural network is a discrete type of convolution, expressed as：

$$\mathrm{X}_{\beta}^{\gamma}=f(\textstyle\sum_{\alpha\in M\beta}x_{\alpha}^{\gamma-1}k_{\alpha\beta}^{\gamma}+b_{\beta}^{\gamma}) \tag{2}$$

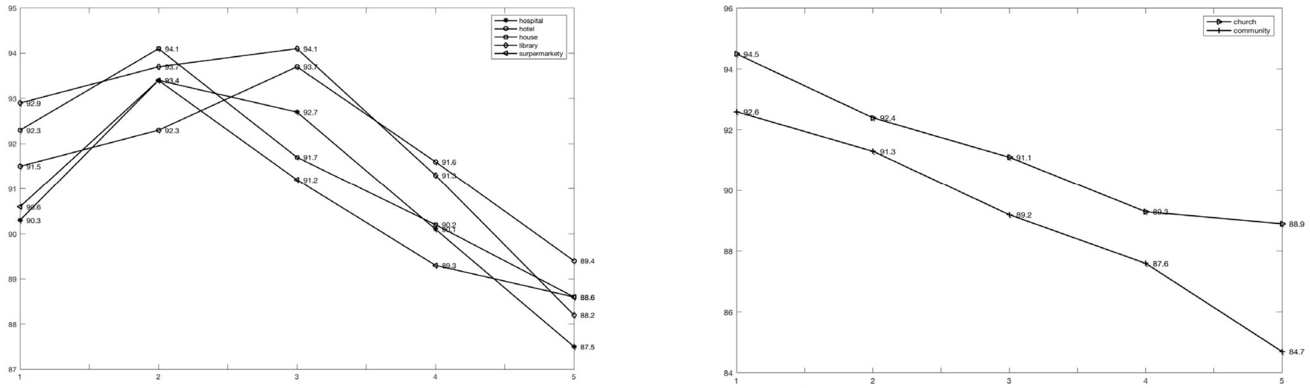
M_{β} represents a choice of input characteristics, *k* is the convolution kernel,   indicates the number of layers in the network, b represents the offset added by each input feature map. For a particular output map, the input mapping features can be obtained using different convolution kernel convolutions. *f* represents the activation function used by convolutional neurons. The most commonly used one is the sigmoid function [11] and hyperbolic tangent function. The difference between them is that the sigmoid function maps [-∞, +∞] to [0,1] and the hyperbolic tangent to [-1,1].

The function of the downsampling layer is to sample the characteristics of the input to achieve dimensionality reduction of the feature data. The number of input features and output features is the same, but the size of the output features is significantly reduced compared to the input features, and the downsampling layer can be expressed as：

$$x_{\beta}^{\gamma}=f(B_{\beta}^{\gamma}sub(x_{\beta}^{\gamma-1})+b_{\beta}^{\gamma}) \tag{3}$$

Where *sub(*)* represents the function used for downsampling, and *B* and *b* are the offsets of the output features. The meaning of *f* is similar to that of the convolutional layer, indicating that the activation function of the downsampled layer neurons may be the same as or different from the convolutional layer.

Currently, convolutional neural networks have several popular network structures, as shown in Table I. In 2012,



(a)

(b)

Fig. 2. Comparison of Building Label Number and Recognition Rate

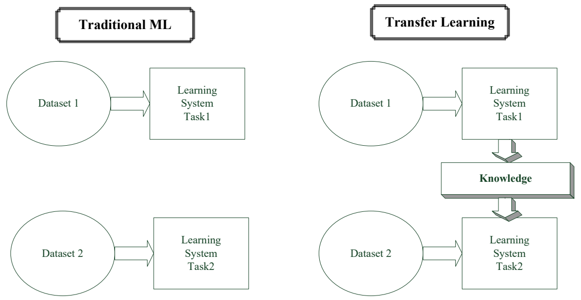


Fig. 1. Traditional Machine Learning and Transfer Learning

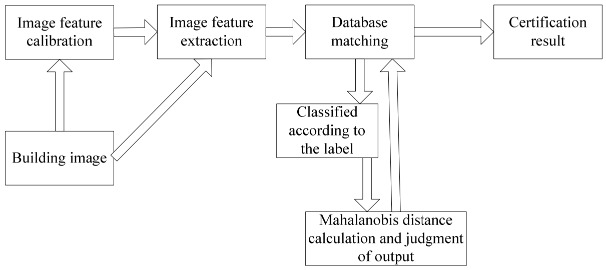


Image feature calibration
Image feature extraction
Database matching
Certification result
Building image
Classified according to the label
Mahalanobis distance calculation and judgment of output

Fig. 3. Comparison Recog-net Certification Process

AlexNet won the ILSVRC2012 classification championship. Convolutional neural networks are widely used in object detection, object classification, video analysis and other fields. After extensive training set learning, CNN convolutional layers are layer-by-layer abstract features from bottom to top, from corner lines to components to overall features. The convolutional neural networks based on end-to-end learning gradually replace traditional manual feature engineering in feature extraction and classification, and they become the most efficient and robust feature extraction algorithm.

B. Transfer Learning Overview

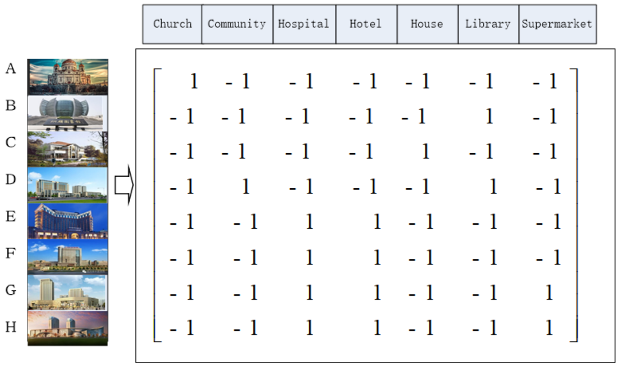
After supervised learning, transfer learning [12] will lead the next wave of commercialization of machine learning technology. And will become the next driving force for machine learning to achieve commercial success, but Today's deep learning algorithms are still lacking which has weak generalization ability in new situations (different from the training set), and the ability to transfer the knowledge learned elsewhere to the new scene is the so called transfer learning.

Traditional machine learning is independent of each other, just purely based on specific data sets and tasks to train their respective data models. And there is no transfer part in the learning independence (as shown in Figure left).But in the transfer learning (right in Figure 1), the user can use the knowledge in the previous training model (features, weights, etc.)) training new models, It can even solve problems such as new tasks with small amount of data. When facing the problem that a structurally complete sample size is too small, Chen [26] *at al* design suitable transfer mechanisms, which can extract additional useful fault features based on the sample with the complete structure. Thus, in the case of a small sample size, it can fully improve the accuracy of fault diagnosis. On the other hand, by transferring the model, we can train an image recognition system with tens of thousands of images. In this way, when we face a new image field, we don't need so much images to train. but transfer our original image recognition system to a new field, and then get the same effect with just a few thousand images in a new field. Its advantage is that it can be combined with deep learning to distinguish the degree of transferring at different levels, and those with higher similarity are more likely to be transferred.

III. BUILDING IDENTIFICATION VIA TRANSFER LEARNING

A. The Effect of Multi-Feature Images on Classification Accuracy

There are many types of buildings. Most of buildings have distinctive features which can be extracted easily. However, some buildings with similar features but different functions are easily misidentified. For example, a Multi-functional building may look like a hotel and a mall, or even like a hospital. This type of building is obviously no suitable while it is only with a single tag identification. For this reason, and before our experiment, we give an experiment on the effect of the number of tags on the recognition accuracy. We have given different amounts of label calibration for different types of buildings, and



$$\begin{array}{c|ccccccc} & \text{Church} & \text{Community} & \text{Hospital} & \text{Hotel} & \text{House} & \text{Library} & \text{Supermarket} \\ \hline A & 1 & -1 & -1 & -1 & -1 & -1 & -1 \\ B & -1 & -1 & -1 & -1 & -1 & 1 & -1 \\ C & -1 & -1 & -1 & -1 & 1 & -1 & -1 \\ D & -1 & 1 & -1 & -1 & -1 & 1 & -1 \\ E & -1 & -1 & 1 & 1 & -1 & -1 & -1 \\ F & -1 & -1 & 1 & 1 & -1 & -1 & -1 \\ G & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ H & -1 & -1 & 1 & 1 & -1 & -1 & 1 \end{array}$$

Fig. 4. Building Feature Matrix

have undergone extensive testing on the dataset to obtain a comparison of the classification accuracy and number of labels for each building, as Figure 2.

In the above figure, the abscissa is the number of labels added to the building picture, and the ordinate is the recognition accuracy of various buildings. In order to make the experimental results clearer, we divided it into two graphs. Figure 2.a shows the buildings with obvious architectural features such as church and community. It can be found that the number of labels is increased with the number of labels. Recognition accuracy will decrease. The accuracy of the five types of building identification in Figure 2.b increases with the increase of the number of calibrations. The identification of buildings such as hospital, house and supermarket have the highest recognition accuracy when two labels are given. For hotel and library, the best recognition rate is obtained when they are tagged with three labels.

B. Image Authentication Model Recog-Net

This paper proposes an end-to-end image authentication technology model Recog-net under transfer learning. The model consists of two parts, the first part completes the image feature extraction meanwhile the second part completes the image authentication. Considering the influence of multi-feature images on the classification accuracy rate, we propose a scheme to add image feature calibration technology module between building image and image feature extraction. This model is described as Recog-net-with-label, that is, when a building has two or more style functions at the same time, the network model we set in the identification process will identify it and a relatively reasonable degree of confidence is obtained. Figure 3 shows the specific certification process of Recog-net.

C. Image Feature Calibration

In the paper we propose an image feature calibration technique to ensure the characteristics of the image are more detailed and artificially identified. In order to make the network express the morphological features of the building more optimally, the building is given the label without the need to cut and calibrate the building in the picture. The characteristic matrix of several types of buildings are shown in Figure 4.

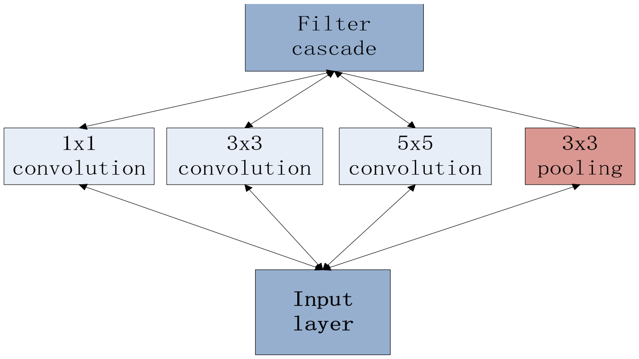


Fig. 5. Inception Module Layer

Figure 4 is a label diagram of the building classification, the left letter indicates the picture of each building and the top is the name of each label, meanwhile the picture category of each building will be calibrated by digital label. The number 1 means the label is assigned to the building, and -1 means the building doesn't have the label. As shown in the figure, if the classification information of a building classification image is very clear (such as Image A, B and C, it can be seen clearly that they belong to churches, libraries and villas respectively), for which they are given a single label. Similarly, for an input image of a building, the building classification may exist under various conditions (such as Image E and F, that is, the appearance of the building looks like both hospital style and hotel style, and the relevant attributes are also added as 1). At the same time, more effective labels for buildings that shown in Image G and H are provide, and the label files are generated corresponding to each image during pre-training. In the process of network training, the feature extractor is pre-trained by imageNet, and the new corresponding tag file image can be regarded as the process of image feature extraction through the trained convolutional neural network until the bottleneck layer process. In the trained network model, the output of the bottleneck layer can be well divided into various types of images through a single layer of fully connected layers. Therefore, it is reasonable to assume that the node vector output from the bottleneck layer can be used as a more streamlined and more expressive feature vector of any image. Also on the new data set, the image feature can be extracted directly using the trained neural network. Then the extracted feature vector is used as input to train a new single-layer fully-connected network to deal with the building identification problem. It is also a guarantee to calibrate it reasonably for the improvement of the classification accuracy.

D. Network Structure

The image extraction network Recog-net based on GoogleNet deep convolutional network proposed in this paper solves the problem of low efficiency of traditional manual feature. The input of this network is 229×229×3. At the same time the convolutional layer in the block all uses 3x3 small convolution kernel for parameter reduction. Figure 6 below shows the Recog-net network structure. Its core is the inception module. The whole inception structure is connected by multiple inception modules. There are two main contributions to the inception structure. The first is using 1x1 convolution for lifting and lowering, the other is to perform convolution re-aggregation

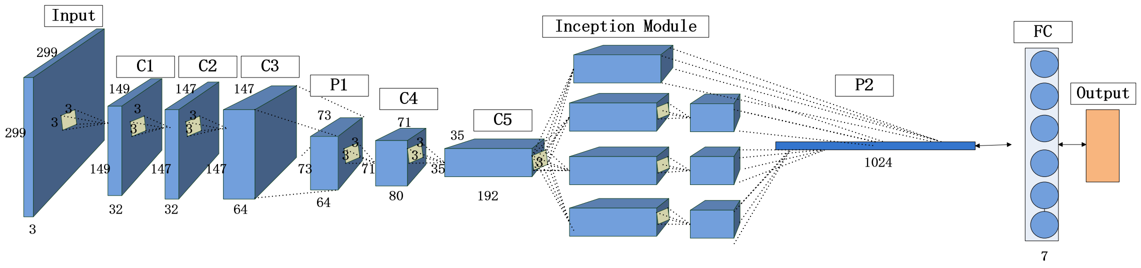


Fig. 6. Comparison of Rec-net Network Structure

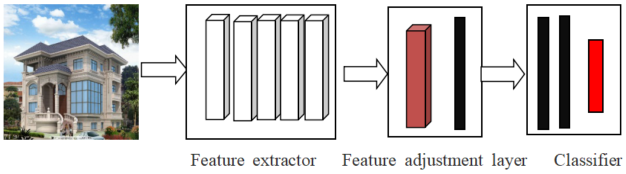


Fig. 7. Network Topology

simultaneously on multiple sizes. Figure 5 above shows the basic connector of the inception module, which is a convolution commonly used in CNN (1x1, 3x3, 5x5). The pooling operations (3x3) are stacked together (the same size after convolution, pooling and adding channels). On one hand, it increases the width of the network, and on the other hand, it increases the adaptability of the network to the scale. The network in the network convolutional layer is able to extract every detail of the input, while the 5x5 filter can also cover the input of most of the receiving layers. A pooling operation can be performed to reduce the size of the space and reduce overfitting. Above these layers, a RELU operation is performed after each convolutional layer to increase the nonlinear characteristics of the network. For convolutional layers, hundreds of convolution kernels are often used to generate a large number of feature maps to capture various features of the image. Therefore, the local activation values of these feature maps can be aggregated as feature vectors to construct more discriminative expressions than manually extracted descriptors and features directly extracted from the CNN fully connected layer.

Figure 7 shows the Rec-net network topology, which is divided into feature extractor, feature adjustment layer and classifier. In order to improve the feature quality, Inception-net is chosen as the feature extractor. The feature adjustment layer consists of a layer of convolutional layer (p1) and a layer of fully connected layer (p2). Layer P1 uses a 1×1 convolution kernel to abstract the features that are learned on the self-made dataset. The characteristics of the target dataset, P2 full connection layer re-enhance the P1 feature, and perform end-to-end feature nonlinear dimensionality reduction, which is beneficial to feature storage, and finally classified by classifier.

E. Image Authentication

After the image feature extraction module completes the feature extraction, two output results are obtained, one is the feature vector of the image, and the other is the label after the image classification. The image authentication module first classifies the image to be authenticated into the corresponding category of the image library, and then performs the Mahalanobis distance calculation with other images of the category. The result is counted as D_i $(i = 1, 2, 3..., m)$, and D_{min} was taken out. Determining, according to the calculation result, whether the image to be authenticated is from the database or the image of the database is from the same source image. In this paper, the threshold method is used to judge, and the threshold *Thm* is set according to a large number of experimental results.

If *Dmin>Thm*, means the distances between this image and all other images are too large, then it doesn't belong to the database. Otherwise, the image belongs to the database, and either this image is the same image or is derived from the same image which corresponds to *Dmin.*

Definition of Mahalanobis distance: *M* sample vectors *X1~Xm* are known, S represents the covariance matrix, μ is the mean value. Then the Mahalanobis distance from the sample vector *Xi* to μ is expressed as:

$$D(X) = \sqrt{(X-\mu)^T S^{-1} (X-\mu)} \tag{4}$$

And the Mahalanobis distance between the vectors X_i and X_j is defined as:

$$D(X_i,X_j) = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)} \tag{5}$$

This process mainly tests the accuracy of the image authentication module. Small-scale rotation, affine, gradation transformation, and other operations are carried out on each image in the self-made dataset Build-7 (as shown in Figure 7) to synthesizing a new Build-7 dataset to simulate tampering images. The accuracy of image authentication depends mainly on the results of previous image classification and the selection of thresholds, so the purpose of this experiment is to find a suitable threshold. The initial value of the threshold is calculated by the distance between the transformed image and the original image, and is set after statistical analysis. Through a large number of experimental analysis, the Mahalanobis distance threshold between the original image and the pre-processed transformation is about 1.2. The Mahalanobis distances of the two completely different images are basically more than 3.0. Because the threshold setting can be rough, we can set the threshold to about 2.0 to experiment.

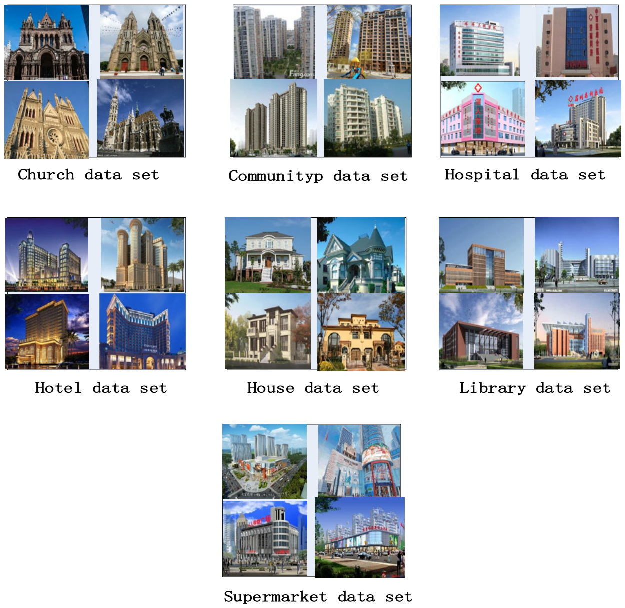
IV. Experiment

A. Datasets

Since there is no related building classification dataset on ImageNet, the datasets used in this paper are all self-made

TABLE II. BUILD-7 DATASET ALGORITHM PERFORMANCE COMPARISON

works	Feature extraction algorithm	Feature dimension	Classification correctness （%)
[21]	HSV	256	27. 6
[22]	SIFT-PCA+BOW (1000)	1000	28. 7
[23]	ScSPM(500)+PCA	1024	34. 3
[24]	HOG+PCA	128	33. 2
[25]	GIST	960	38. 9
[13]	Alex-Net (FC7)	4096	82. 7
[15]	VGG_CNN_F(FC7)	4096	79. 6
[15]	VGG_CNN_M(FC7)	4096	83. 4
[15]	VGG_CNN_S(FC7)	4096	82. 5
[17]	Google-Net	1024	82. 7
[19]	Res-Net	1024	81. 4
Algorithm 1	Rec-net	512	81. 2
Algorithm 2	Rec-net -with-lable	512	91. 9



Church data set

Communityp data set

Hospital data set

Hotel data set

House data set

Library data set

Supermarket data set

Fig. 8. Build-7 Data Set

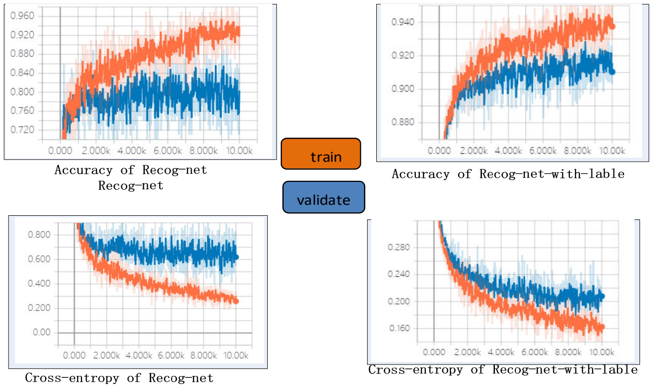
dataset Build-7, as shown in Figure 8. The images are from Google and Baidu, and they are divided into seven types of common buildings. The total number of the images is 2,975, including 421 churches, 432 communities, 421 hospitals, 483 hotels, 405 houses and 418 supermarkets. Each image is pre-processed to the same size of 299 × 299 × 3 pixels.

B. Experimental Environment

The hardware configuration of this experiment is Intel core i7 6700k, 8 core 16 lines, 3.4GHZ frequency, 16GB memory, NVIDIA GTX 1070 6GBRAM. we develop the method via Tensorflow, which is a very famous deep learning framework based on Python.

C. Data Training Process

We train the fully-connected network after extracting the feature vector by he CNN-based feature extractor. In this paper, the number of iterations is set to 10,000, each batch-size is 100, and one round of prediction is performed every 10 times. At the same time, 80% of the total data is used as the main training sample, another 10% is used as the cross-validation sample in the training process, and the remaining 10% is used as the test dataset for predicting the performance of the classifier. The optimizer is SGD (Random Gradient Descent Algorithm). The training comparison is mainly for the comparison of Recog-net and Recog-net-with-label.



Accuracy of Recog-net
Recog-net

Accuracy of Recog-net-with-lable

Cross-entropy of Recog-net

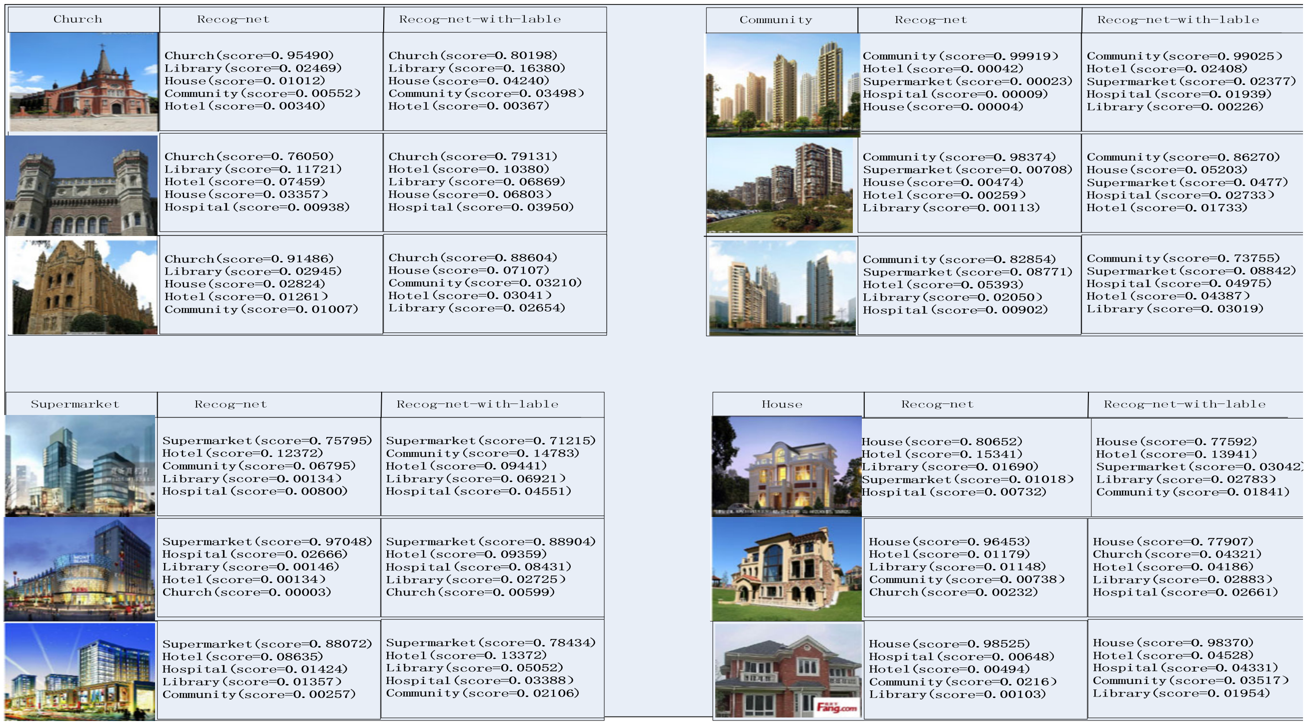
Cross-entropy of Recog-net-with-lable

Fig. 9. Network Training Process

From Figure 9, it can be seen that as the number of training increases, the loss value gradually decreases. For Recog-net, the loss value at the 9kth is 0.21, which shows the feasibility of transfer learning. The accuracy is also improved. For the change in the loss value from 0.35 to 0.14 for Recog-net-with-label, the superiority of the transfer learning is demonstrated. The accuracy of 91.9% verification is also greatly improved compared with 81.2% of the former.

D. Exploring the Proposed Method

The purpose of this experiment is to verify the feature quality of Recog-net and the accuracy of classification. The experimental dataset uses the self-made dataset Build-7, the comparison algorithms are HSV [21], SIFT+BOW [22],



Church	Recog-net	Recog-net-with-lable

	Church(score=0. 95490)
	Library(score=0. 02469)
	House(score=0. 01012)
	Community(score=0. 00552)
	Hotel(score=0. 00340)
		Church(score=0. 80198)
		Library(score=0. 16380)
		House(score=0. 04240)
		Community(score=0. 03498)
		Hotel(score=0. 00367)

	Church(score=0. 76050)
	Library(score=0. 11721)
	Hotel(score=0. 07459)
	House(score=0. 03357)
	Hospital(score=0. 00938)
		Church(score=0. 79131)
		Hotel(score=0. 10380)
		Library(score=0. 06869)
		House(score=0. 06803)
		Hospital(score=0. 03950)

	Church(score=0. 91486)
	Library(score=0. 02945)
	House(score=0. 02824)
	Hotel(score=0. 01261)
	Community(score=0. 01007)
		Church(score=0. 88604)
		House(score=0. 07107)
		Community(score=0. 03210)
		Hotel(score=0. 03041)
		Library(score=0. 02654)

Community	Recog-net	Recog-net-with-lable

	Community(score=0. 99919)
	Hotel(score=0. 00042)
	Supermarket(score=0. 00023)
	Hospital(score=0. 00009)
	House(score=0. 00004)
		Community(score=0. 99025)
		Hotel(score=0. 02408)
		Supermarket(score=0. 02377)
		Hospital(score=0. 01939)
		Library(score=0. 00226)

	Community(score=0. 98374)
	Supermarket(score=0. 00708)
	House(score=0. 00474)
	Hotel(score=0. 00259)
	Library(score=0. 00113)
		Community(score=0. 86270)
		House(score=0. 05203)
		Supermarket(score=0. 0477)
		Hospital(score=0. 02733)
		Hotel(score=0. 01733)

	Community(score=0. 82854)
	Supermarket(score=0. 08771)
	Hotel(score=0. 05393)
	Library(score=0. 02050)
	Hospital(score=0. 00902)
		Community(score=0. 73755)
		Supermarket(score=0. 08842)
		Hospital(score=0. 04975)
		Hotel(score=0. 04387)
		Library(score=0. 03019)

Supermarket	Recog-net	Recog-net-with-lable

	Supermarket(score=0. 75795)
	Hotel(score=0. 12372)
	Community(score=0. 06795)
	Library(score=0. 00134)
	Hospital(score=0. 00800)
		Supermarket(score=0. 71215)
		Community(score=0. 14783)
		Hotel(score=0. 09441)
		Library(score=0. 06921)
		Hospital(score=0. 04551)

	Supermarket(score=0. 97048)
	Hospital(score=0. 02666)
	Library(score=0. 00146)
	Hotel(score=0. 00134)
	Church(score=0. 00003)
		Supermarket(score=0. 88904)
		Hotel(score=0. 09359)
		Hospital(score=0. 08431)
		Library(score=0. 02725)
		Church(score=0. 00599)

	Supermarket(score=0. 88072)
	Hotel(score=0. 08635)
	Hospital(score=0. 01424)
	Library(score=0. 01357)
	Community(score=0. 00257)
		Supermarket(score=0. 78434)
		Hotel(score=0. 13372)
		Library(score=0. 05052)
		Hospital(score=0. 03388)
		Community(score=0. 02106)

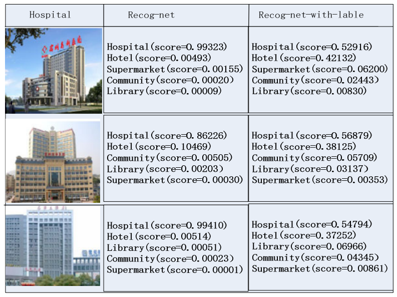
House	Recog-net	Recog-net-with-lable

	House(score=0. 80652)
	Hotel(score=0. 15341)
	Library(score=0. 01690)
	Supermarket(score=0. 01018)
	Hospital(score=0. 00732)
		House(score=0. 77592)
		Hotel(score=0. 13941)
		Supermarket(score=0. 03042
		Library(score=0. 02783)
		Community(score=0. 01841)

	House(score=0. 96453)
	Hotel(score=0. 01179)
	Library(score=0. 01148)
	Community(score=0. 00738)
	Church(score=0. 00232)
		House(score=0. 77907)
		Church(score=0. 04321)
		Hotel(score=0. 04186)
		Library(score=0. 02883)
		Hospital(score=0. 02661)

	House(score=0. 98525)
	Hospital(score=0. 00648)
	Hotel(score=0. 00494)
	Community(score=0. 0216)
	Library(score=0. 00103)
		House(score=0. 98370)
		Hotel(score=0. 04528)
		Hospital(score=0. 04331)
		Community(score=0. 03517)
		Library(score=0. 01954)

Fig. 10. Network Comparison of Four Types of Building Data Test



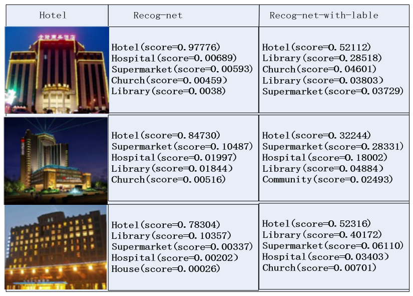
Hospital	Recog-net	Recog-net-with-lable

	Hospital(score=0. 99323)
	Hotel(score=0. 00493)
	Supermarket(score=0. 00155)
	Community(score=0. 00020)
	Library(score=0. 00009)
		Hospital(score=0. 52916)
		Hotel(score=0. 42132)
		Supermarket(score=0. 06200)
		Community(score=0. 02443)
		Library(score=0. 00830)

	Hospital(score=0. 86226)
	Hotel(score=0. 10469)
	Community(score=0. 00505)
	Library(score=0. 00203)
	Supermarket(score=0. 00030)
		Hospital(score=0. 56879)
		Hotel(score=0. 38125)
		Community(score=0. 05709)
		Library(score=0. 03137)
		Supermarket(score=0. 00353)

	Hospital(score=0. 99410)
	Hotel(score=0. 00514)
	Library(score=0. 00051)
	Community(score=0. 00023)
	Supermarket(score=0. 00001)
		Hospital(score=0. 54794)
		Hotel(score=0. 37252)
		Library(score=0. 06966)
		Community(score=0. 04345)
		Supermarket(score=0. 00861)

Fig. 11. Hospital Test Comparison



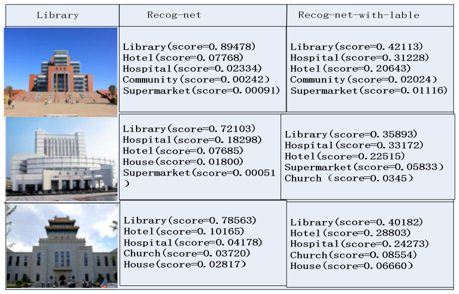
Hotel	Recog-net	Recog-net-with-lable

	Hotel(score=0. 97776)
	Hospital(score=0. 00689)
	Supermarket(score=0. 00593)
	Church(score=0. 00459)
	Library(score=0. 0038)
		Hotel(score=0. 52112)
		Library(score=0. 28518)
		Church(score=0. 04601)
		Library(score=0. 03803)
		Supermarket(score=0. 03729)

	Hotel(score=0. 84730)
	Supermarket(score=0. 10487)
	Hospital(score=0. 01997)
	Library(score=0. 01844)
	Church(score=0. 00516)
		Hotel(score=0. 32244)
		Supermarket(score=0. 28331)
		Hospital(score=0. 18002)
		Library(score=0. 04884)
		Community(score=0. 02493)

	Hotel(score=0. 78304)
	Library(score=0. 10357)
	Supermarket(score=0. 00337)
	Hospital(score=0. 00202)
	House(score=0. 00026)
		Hotel(score=0. 52316)
		Library(score=0. 40172)
		Supermarket(score=0. 06110)
		Hospital(score=0. 03403)
		Church(score=0. 00701)

Fig. 12. Hotel Test Comparison



Library	Recog-net	Recog-net-with-lable

	Library(score=0. 89478)
	Hotel(score=0. 07768)
	Hospital(score=0. 02334)
	Community(score=0. 00242)
	Supermarket(score=0. 00091)
		Library(score=0. 42113)
		Hospital(score=0. 31228)
		Hotel(score=0. 20643)
		Community(score=0. 02024)
		Supermarket(score=0. 01116)

	Library(score=0. 72103)
	Hospital(score=0. 18298)
	Hotel(score=0. 07685)
	House(score=0. 01800)
	Supermarket(score=0. 00051
	)
		Library(score=0. 35893)
		Hospital(score=0. 33172)
		Hotel(score=0. 22515)
		Supermarket(score=0. 05833)
		Church (score=0. 0345)

	Library(score=0. 78563)
	Hotel(score=0. 10165)
	Hospital(score=0. 04178)
	Church(score=0. 03720)
	House(score=0. 02817)
		Library(score=0. 40182)
		Hotel(score=0. 28803)
		Hospital(score=0. 24273)
		Church(score=0. 08554)
		House(score=0. 06660)

Fig. 13. Library Test Comparison

ScSPM [23], HOG [24], GIST [25], Alex Net, VGGNet, Recog-net and other neural network models obtained by transferring and fine-tuning. In this paper, SIFT was used for comparative experiments, but SIFTs' process of feature point extraction was slow. The dimension of each feature point is 128 dimensions, and because the number of feature points in each picture is different, the feature vector dimensions of each graph are inconsistent. Therefore, in this experiment, the SIFT features are first extracted and then combined with the word bag model (BOW). And the spatial pyramid matching (SPM) algorithm, which respectively represents the feature vector.

In the experimental simulation, it is known from Table II that in terms of dimensions, the traditional algorithms have better performance than the neural network. On the one hand, they benefit from the PCA dimension reduction and on the other hand the coding methods, such as BOW, IFV, SPM. Considering the feature quality, it can be seen from the classification correctness data that the neural network feature extraction algorithm is much higher in efficiency and performance than the traditional feature engineering. The main reason is that traditional algorithms require artificially designed features and rely on prior knowledge. This leads to poor feature versatility, and it is not easy to capture the essential features of objectives in complex scenes.

On the contrary, based on end-to-end learning the neural network structures rely on the advantages of big data and high-dimensional parameter space, and they can synthesize advanced features from the bottom to the top gradually. The data-driven self-learning method ensures that the convolutional neural network has excellent feature extraction ability. At the same time, the comparison of the neural network model reveals that the error of Recog-net-with-lable is 8 to 10 percentage points lower than other networks. The label calibration can greatly improve the quality of network feature extraction and improve its classification accuracy.

As shown in Fig. 10 to Fig. 13, Recog-net based on inception network migration performs well in building recognition. For buildings with obvious outlines as shown in Figure 10 (church, community, house, supermarket), the confidence in the top-5 accuracy rate can clearly identify the building information. This recognition method is not identified according to the text information in the figure. The text information in the picture is only intended to provide a classification reference. However, due to the intra-class variability and inter-class similarity among buildings, the Recog-net network is prone to errors in identifying pictures with similarity in different classes. For this reason, the calibrated network is able to assign a reasonable confidence level to the image of the building where the problem exists (show as Figures 11, 12, 13). That is, the building cannot be a library only, and it may be a hospital from its morphological characteristics. The corresponding proportions are also reflected in the difference of mutual confidence. It makes the testing process more practical, not just a single (Recog-net) building classification. When this identification method is applied to buildings with multiple classification possibilities, these building can be distinguished more accurately and the results will be more convincing.

V. Conclusions

From the experimental data presented in this paper, it can be clearly seen that the versatility and stability of the traditional features are not very good and there are some limitations of their usage. Convolutional neural networks are highly versatile and stable, therefore, they can play a vital role in feature extraction.

Compared with some existing convolutional neural networks, the proposed Recog-net structure has higher accuracy, and is superior to the existing ones in feature learning. Based on this network, label calibration is performed. Recog-net-with-lable improves the accuracy by nearly 10% in the recognition process, and has a good recognition effect on the image recognition of similarity between different building classes. In general, the proposed image identification model has the advantages of low storage requirement, high calculation speed. At the same time, it is easy to train small datasets by using transfer learning. It is promising to apply this model to enterprises or individual having the requirement but lack of data.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 61573316, 61873082), the ZheJiang province Natural Science Foundation of China (No. LY20F020022), the National Key R&D Program of China (No. 2018YFB0204003).

References

[1] Urai T, Okunaka D, Tokumaru M. Clothing image retrieval based on a similarity evaluation method for Kansei retrieval system[M]. 2012..

[2] Xiaoou Tang, Ke Liu, Jingyu Cui, Fang Wen and Xiaogang Wang, IntentSearch: Capturing User Intention for One-Click Internet Image Search, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 7, pp.1342-1353, Jul. 2012

[3] Jinyong Wu, Yong Zhao, Xing Zhang, Jun Wang, and Yike Wang, A Cascaded Retrieval Method of Specified Object Based on Fusing Multiple Features, Third Global Congress on Intelligent Systems (GCIS), pp. 117-121, Nov. 2012

[4] Yang D K, Lin Y W, Chiu Y I, *et al*. Vision based campus guide system on intelligent mobile phone. Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on. IEEE, 2013.

[5] Guo Z, Chen Q, Wu G, *et al*. Village Building Identification Based on Ensemble Convolutional Neural Networks[J]. Sensors, 17(11),2017

[6] Tremblay-Gosselin J, Cretu A M. A supervised training and learning method for building identification in remotely sensed imaging IEEE International Symposium on Robotic & Sensors Environments. 2014.

[7] Moun C, Netramai C. Localization and building identification in outdoor environment for smartphone using integrated GPS and camera Fourth International Conference on Digital Information & Communication Technology & Its Applications. IEEE,2014.

[8] Woodley R S, Noll W, Barker.*et al*. Automatic building identification using gps and machine learning. IEEE Geoscience & Remote Sensing Symposium.2010.

[9] Chen K H, Wu C R, Yang Y L, *et al*. Efficient building identification using structural and spatial information on mobile devices. In IEEE International Conference on Multimedia & Expo Workshops. 2014.

[10] Jinbin H, Xuqing T. BP Algorithm of Artificial Neural Network and Its Application. Information Technology, 28(4):1-4,2004.

[11] Ruizhong Z, Cibing L. Application of Sigmoid Function in Neural Network, Thermal Conduction and Earth Pressure Calculation. Journal of Fuzhou University (Natural Science), 29(3): 79-83. 2001.

[12] Song T. Research on several problems of target ecognition based on migration learning. University of Electronic Science and Technology of China, 2017.

[13] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks NIPS. Curran Associates Inc.2012.

[14] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks European Conference on Computer Vision. Springer International Publishing, 818-833,2014.

[15] Chatfield K, Simonyan K, Vedaldi A, *et al*. Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531, 2014.

[16] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Computer Science, 2014.

[17] Szegedy C, Liu W, Jia Y, *et al*. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 1-9.2015.

[18] Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision. In IEEE International Conference on Computer Vision.2015

[19] He K, Zhang X, Ren S, *et al*. Deep residual learning for, image recognition. In IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[20] Szegedy C, Ioffe S, Vanhoucke V. Inception-v4, inception-resnet and the impact of residual connections on learning. In IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[21] Horprasert T, Harwood D, Davis L S. A statistical approach for real-time robust background subtraction and shadow detection. IEEE ICCV, 99: 1-19.1999.

[22] Banerji S, Sinha A, Liu C. A New Bag of Words LBP(BoWL) Descriptor for Scene Image Classification. International Conferences on Computer Analysis of Image & Patterns Spring Berlin Heidelberg,2013.

[23] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In IEEE Conference on Computer Vision and Pattern Recognition, 2:2169-2178,2006.

[24] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In IEEE Conference on Computer Vision and Pattern Recognition, vol.1,886-893,2005.

[25] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision, 42(3):145-175,2001.

[26] Chen, D M, Yang S, Zhou F N. Transfer Learning Based Fault Diagnosis with Missing Data Due to Multi-Rate Sampling[J]. Sensors, 2019, 19(8).