IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 19, 2020, accepted April 2, 2020, date of publication April 6, 2020, date of current version April 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2985991

A Novel Orthogonality Loss for Deep Hierarchical Multi-Task Learning

GUIQING HE[1], (Member, IEEE), YINCHENG HUO[1], MINGYAO HE[1], HAIXI ZHANG[1], AND JIANPING FAN[2], (Member, IEEE)

[1]School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China
[2]Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

Corresponding author: Guiqing He (guiqing_he@nwpu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61402368, in part by the Aerospace Science and Technology Innovation Foundation of China under Grant 2017ZD53047, in part by the Aerospace Support Foundation of China under Grant 2019-HT-XGD, and in part by the Seed Foundation of Innovation and Creation for Graduate Students in Northwestern Polytechnical University under Grant ZZ2019166.

ABSTRACT In this paper, a novel loss function is proposed to measure the correlation among different learning tasks and select useful feature components for each classification task. Firstly, the knowledge map we proposed is used for organizing the affiliation relationship between objects in natural world. Secondly, a novel loss function–orthogonality loss is proposed to make the deep features more discriminative by removing useless feature components. Furthermore, in order to prevent the extracted feature maps from being too divergent and causing over-fitting which will reduce network performance, this paper also added the orthogonal distribution regularization term to constrain the distribution of network parameters. Finally, the proposed orthogonality loss is applied in a multi-task network structure to learn more discriminative deep feature, and also to evaluate the validity of the proposed loss function.The results show that compared with the traditional deep convolutional neural network and a multi-task network without orthogonality loss, the multi -task based orthogonality loss is significantly better than the other two types of networks on image classification.

INDEX TERMS Orthogonality loss, multi-task learning, orthogonal distribution regularization.

I. INTRODUCTION

In recent years, image classification [1]–[8] has become more and more widely used in field exploration and daily life due to the rapid development of deep learning, and image classification also receives much attention in optical machine learning such as [9] and [10]. Currently, the best tool for feature extraction is the deep convolutional neural network. Deep convolutional neural networks [3]–[8]can not only extract edge information in shallow layers, but also learn more high-level feature representations which become more abstract and closer to human cognitive behavior with the deeper semantic information. The multi-task network [11]–[14] derived from deep learning has gradually entered people's sight. Different tasks of the multi-task network are mutually assigned and trained at the same time, but each task has its own independent loss function. However, in the multi-task network, the joint training of parameters in hidden layer can not make the feature extraction and classification completely matched, and the useless feature components may bring negative influence to the final classification result. Therefore, in order to achieve such an ambitious goal, we must solve the following problems first.

The first problem is how to guide classification task in each level to assist other classification tasks. It is a gradual process from coarse-level to fine-level when identifying thousands of object classes in the real world by imitating the human learning experience. People may only identify coarse classes, such as birds, cars, and plants when they are young. As the brain system matures, the identifiable targets may be refined in the process of learning common sense. For example, one specific type of bird have parrots, sparrows, etc., and there are buses or cars in the types of car. The recognition of the coarse-grained genera is considered as the high level task, and there are many subtasks that identify fine-grained classes under each genera. Therefore, the knowledge [15]–[17] acquired in

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

the high level classification task can also be added to the new task of identifying the fine grained species, which can help separating fine-grained object classes. At the same time, you can better understand the high level classification tasks after learning new tasks. The multi-task scheme constructed in this paper simulates the human learning process, using deep convolutional neural networks as a hidden layer to simulate human brain for feature extraction. Also a hierarchical tree classifier [18]–[20] is leveraged as a task-related output layer for progressive classification, the classification process from easy to difficult and interrelated constitutes different learning tasks. Therefore, based on these progressive relationships, this paper establishes a knowledge map about the database to guide the network learning during the training process. Just as humans need to learn by means of books or predecessors, additional information is used to guide each classification task to assist each other.

The second problem is the classifier and the features extraction network are not completely matched. In the multi-task network structure, the feature representation is shared by multi-level tree classifiers because the weight parameter in hidden layer are shared. But for the different classification tasks of the task-related output layer-hierarchical tree classifier, the features that multi-level classification task need are not exactly the same. For example, when identifying coarse-grained object classes, identifying steering wheel or wheels can help the network accurately identify the car category. But when identifying fine-grained object classes, these common features may have a greater effect on bringing SUVs and vans into the same category, and we expect the network to pay more attention to their unique feature components, such as appearance, shape, which can be used to distinguish them. Therefore, in the process of performing different classification tasks in the multi-task network, we need to distinguish the common feature components and unique feature components. Considering previous description, a novel loss function — orthogonality loss is proposed for feature selection in multi-level hierarchical classifier. Orthogonality loss uses the cosine similarity in metric learning [21]–[23] to measure the similarity of multi-level deep features. When a vector is divided into a projection vector and an orthogonal vector with another vector, the area of the projection vector represents the cosine similarity. If the cosine similarity between the coarse level deep feature vectors and the fine level deep feature vectors approximate to 0, the two vectors are orthogonal, and the overlap features (projection vectors) can be filtered out, leaving their own useful feature components (orthogonal vectors) only. Such method can allow our proposed network structure get more discriminative deep features for different classification tasks.

The last question is how to prevent the network from overfitting. After the feature selection is added to the loss function, the feature is too divergent and overfitting phenomenon is caused. In order to prevent such problem, this paper adds orthogonal distribution regularization term in the orthogonality loss to constrain the feature distribution, so that the partial distribution of deep feature under same learning task is required to be close to the overall distribution. So it is necessary to prevent the network from overfitting by adding regularization, since overfitting may prevent the network from global optimum.

Based on above discussions, the orthogonality loss proposed in this paper is used in multi-task network to select useful feature components for different classification tasks to improve the accuracy of image recognition. The rest of the paper is arranged as follows: section 2 introduces the related work of this paper, section 3 introduces the network structure and orthogonality loss, section 4 introduces the experimental results and explanations, and section 5 draws the conclusions of the full paper.

II. RELATED WORK

In order to extract rich and vivid deep features, we generally train a deep convolutional neural network. With the advancement of many theories [1]–[8] of deep learning and hardware devices from the earliest time-delayed neural network (TDNN) [1] and LeNet-5 [2], convolutional neural networks have developed rapidly and been widely used in the fields of computer vision [24], natural language processing [25] and optical imaging [26], [27]. Deep convolution neural network is a kind of deep network structure with convolution operation and characterization learning ability, which consisting of convolutional layers, pooling layers, nonlinear activating layers and fully connected layers. This kind of network extracts the invariant features through parameter sharing and sparse connections between convolution kernels. Among this, Resnet [7] is widely used because of significant advantages in recognition accuracy and calculation amount. However, the N-way softmax classifier is unable to pay attention to the similarity imbalance between classes.

With the development of classification technology [28]–[36], the application of multi-task network [37]–[40] in recent years has accelerate this problem. The most common way for applying multi-task learning into deep convolutional neural networks is the hidden layer parameter sharing mechanism: the sharing mechanism first proposed in [41], which proposed the hidden layer parameters were shared but the task-related output layers were independent. Such technology has also been rapidly developed and been widely used in natural language processing [37], facial landmark detection [39] and object detection [42]. The multi-task network mentioned in [43] states that the deep convolutional neural network is applied as a hidden layer for parameter sharing, which can be further used to extract the high level representation of object image, and the tree-classifier is used as the task-related output layer for different classifications tasks. Moreover, the construction method of the label tree mentioned in [13] and the ontology tree mentioned in [44]can guide the tree-classifier for multi-task classification.

However, guiding multi-task classification through knowledge map in multi-task network [45] is the closest approach to human behavior. Knowledge map [46], as a novel knowledge

structure and retrieval technology in the era of big data, has gradually revealed its advantages in various aspects and has received extensive attention. The knowledge map was originally proposed by Google in 2012 to improve the capabilities of search engines [46]. However, the powerful semantic processing ability of knowledge map makes it one of the key technologies in the development and application of artificial intelligence [47]. In a multi-task network, the knowledge map is applied to construct a two-layer semantic structure [19] for representing the relationship between objects in the real world, which can be used to guide the transmission of information between different classification tasks of the multi-task network, and guide backpropagation for gradient updates.

The orthogonal transformation of images is widely used in the fields of image feature extraction [48], image enhancement [49], image restoration [50] and image classification [51]. The orthogonal matching pursuit algorithm (OMP) [52] has also been well applied in the field of image fusion. And the Go-CNN network mentioned in [53] can learn the foreground and background of the feature. Therefore, orthogonal transform can be used to extract and distinguish image features. In addition, the projection vectors of the two vectors are common parts and the orthogonal vectors are unique components in geometry. If the two feature vectors are orthogonal, it can be proved that the two feature vectors are completely independent. Therefore, orthogonal transform can also be utilized for feature selection. In order to achieve end-to-end learning, the loss function is usually used to update the network structure to reduce the gap between the predicted value and the true value. The classic losses are Hinge Loss (multiple for SVM) [54], Softmax cross entropy (classification task and feature extraction task) [55], Contrastive Loss (contrast loss function, LeCun proposed in the siamese twin network) [56]. The most fundamental criterion of the loss function is to achieve the defining ultimate goal of the model. Therefore, network performance can be improved by optimizing the loss function.

Based on these observations, this paper proposes a loss function optimization method for multi-task networks. Firstly, the softmax loss of each classification task is preserved. On this basis, the orthogonal part is added to complete feature selection, so that the sub-classifier and the parent node classifier feature vector are orthogonal and the extracted features are relatively independent. Orthogonal distribution regularization is also added to constrain the distribution of features, so that the overall distribution of each sub-task feature vectors is closer to the distribution characteristics of its parent task. Therefore, the orthogonality loss mentioned in this paper obtains a multi-task network with higher classification accuracy and better robustness through feature selection.

III. ALGORITHM

A. KNOWLEDGE MAP

During the training process, completely ignoring the similarity between different classes makes it difficult to achieve the global optimum. The knowledge maps are widely used in large-scale classification task as it can efficiently organize large-scale object classes in a course to fine fashion. In this paper, based on the taxonomic knowledge of object classes in real world, Fashion-60 [43] and Caltech-UCSD Birds-200-2011 [57] are divided into two semantic structures, including coarse-grained genus and fine-grained classes. For Fashion-60 database, there are 60 classes of clothes (including dress, shoes, etc.). The two-layer knowledge map constructed with reference to the functional relationship of each item is shown in Fig.1. According to the function of each item, 60 fine classes are used to represent 60 specific classes of clothing and all of them are assigned into 5 different coarse-grained genus; for Caltech-UCSD Birds-200-2011 database, a knowledge map of 200 species of birds, constructed with reference to the natural system relationships of birds, is shown in Fig.2, which containing 10 coarse-grained genus to represent 10 species of birds and 200 fine classes to represent the specific classes of birds under each genus.

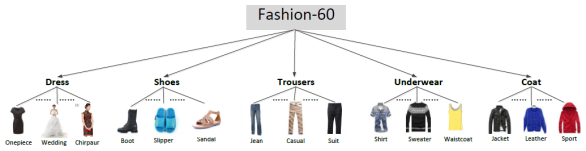


FIGURE 1. Knowledge map of Fashion-60.

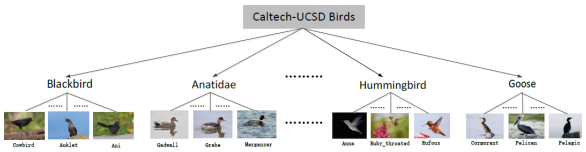


FIGURE 2. Knowledge map of Caltech-UCSD Birds-200-2011.

The knowledge map consists of the two-layer semantic structure is used to guide the hierarchical tree classifier for multi-level classification tasks. Each tree structure constitutes a learning task, and a tree classifier is constructed for considering the inter-species relations between multiple classes, and the knowledge map guide tree classifier replaces the traditional softmax classifier. The objective function we proposed can be used to help to efficiently update weight parameters in both classifier and base deep network to make the gradient distribution under the same task more uniform.

B. ORTHOGONALITY LOSS

In this paper, the multi-task classification is been divided in two different classification tasks. The classifier in each level corresponded for different tasks, so the required deep features should be very different. According to the knowledge map, when dealing with coarse level classification task, the network is expected to pay more attention on the common feature components that all the fine grained classes which under the same learning task, and ignores the unique feature components of their own, then a specific classifier is trained

to recognize those fine grained classes. In addition, it is more difficult when dealing with fine classification tasks, because the similarity between the fine-classes which under the same coarse class is much higher. Therefore, we expect the network focus on the unique feature components of each fine classes and ignore their common feature components to make sure the visual relevance between classes can be ignored, making the network have a higher chance to distinguish images.

Based on the multi-task network for classification, an orthogonality loss is proposed in this paper for feature selection. Orthogonality loss completes feature selection like this: it randomly selects an image among the whole training set for feature extraction, and then the features are input to different task-related output layers. Under the guidance of the loss function in different classification tasks, the task-related output layer features are divided into coarse level deep features and fine level deep features. The coarse level deep features are used for the coarse classification task which aims to find out which coarse class the image belonging to; the fine level deep features are used for the fine classification task to determine which fine class under the coarse class the image belonging to. As shown in Fig.3, the spatial projection of the coarse classifier feature and the fine classifier feature is the overlap feature components. The overlap feature components contain the overlap components of the coarse level deep features and the fine level deep features: in the coarse level classification task, the overlap feature components of the coarse level deep feature will contain some unique feature components of each fine classes, which restrain to classify the same coarse class images together. Similarly, when performing a fine level classification task, the overlap feature of the fine classifier feature will have some common feature of the same coarse class, which will restrain to separate the fine classes belonging to a same coarse class.

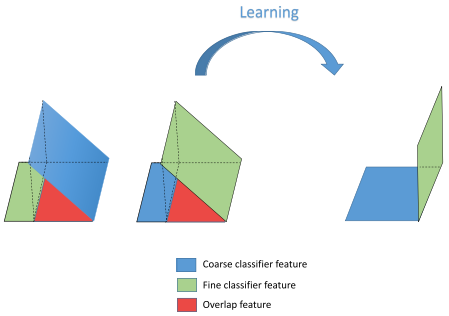


FIGURE 3. Learning with orthogonality loss.

Therefore, we hope the coarse level deep feature and the fine level deep feature are orthogonal in space, which makes the overlap feature approximate to 0. We used the loss function to achieve the ultimate goal by adding the target of the feature selection to the loss function and measuring the size of the overlap feature with orthogonality loss. So, the network can automatically complete feature selection and distinguish the two feature vectors spatially to help different classifiers remove features that are useless to their own classification tasks.

C. THE STRUCTURE OF THE MULTI-TASK NETWORK

Based on the above understanding of the orthogonality loss, the structure of the network is first described in this section, as shown in Fig.4. Considering an image as input, one classifier is used to complete the coarse classification task to determine which coarse class the image belongs to; the corresponding fine-grained classifier is selected to determine which fine class under a coarse class the object image belong to. The hidden layers of the two classification tasks are the deep convolutional neural networks, whose parameters are shared. In this paper, the hidden layer of the network is used for feature extraction. After the features are passed to a fully connected layer FC6, they enter the task-related output layer, which also called the tree classifier. The tree classifier consists of different sub-classifiers, each of them has the same structure but the weight parameters are not shared, where the same structural means that they both contain two fully connected layers and a softmax layer for classification tasks. Therefore, the feature vectors through the task-related output layer – coarse classifier and fine classifier can be differentiated.

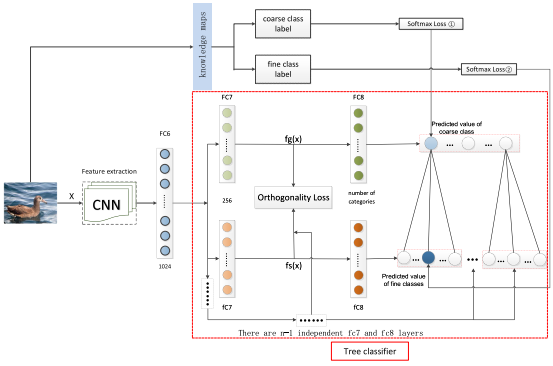


FIGURE 4. The structure of the multi-task network.

As shown in Fig.4, suppose the input image is X, CNNs with shared weight are used for two different tasks, and then features are extracted through a same FC6 layer, so that all processing is the same and the extracted features is the same. But there are two independent fully connected layer of the sub-classifiers and their parameters are not shared, so the FC7 and fc7 extracted features are different. So when performing a coarse level classification task, the coarse classifier feature is$f_g(x)$; and when performing a fine classification task, the fine classifier feature is $f_s(x)$. So $f_g(x)$ and $f_s(x)$ can be used to calculate the orthogonality loss of the network. Assuming that N training images are input, the orthogonality loss function can be formulated in following (1):

$$L_1(x) = \frac{1}{N} \min_{f_1, f_2, \ldots, f_k} \sum \alpha Tr[f_s(x) f_g^T(x)] \tag{1}$$

where k represents the number of coarse classes, $f_1, f_2, \ldots, f_k$ represents k fine classification tasks (only one fine classifier structure is drawn in Fig.4). $f_g(x)$ represents the coarse classifier feature of N images and $f_s(x)$ represents the fine classifier feature of N images.The trace of $f_s(x)f_g^T(x)$ represents the sum of dot products of the coarse classifier features and fine classifier features with the N images. When $Tr[f_s(x)f_g^T(x)]$ approaches to 0, it means that the corresponding vectors of the coarse grained deep feature and the fine grained deep feature tend to be orthogonal. α is a hyper parameter, and the magnitude of α represents the influence of orthogonality loss of the entire network parameters during backpropagation.

Under the guidance of orthogonality loss function, $f_g(x)$ and $f_s(x)$ tend to be orthogonal, so that the network's derived classifier features and fine classifier features may become more discriminative. Then $f_g(x)$ and $f_s(x)$ are transferred to the softmax layers for classification, the coarse classification result and the fine classification result are respectively obtained. Combining the two results determines which fine class under the coarse class the object image should belonging to. Different tasks have different losses, so the loss by the classification is composed of the gap between the predicted value and the coarse class label in the coarse classification task and the difference between the predicted value and the fine class label in the fine classification task. In Fig.4, softmax loss ① maens the gap between the predicted value and the coarse class label in the coarse classification task; and softmax loss ② means the difference between the predicted value and the fine class label in the fine classification task. which can be measured by the following loss function (2):

$$L_2(x) = \min_{g\in G, s\in S_g} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k_g} l(y(i)=j)log\frac{e^{\theta_{gj}^T x^{(i)}}}{\sum_{l=1}^{k_g} e^{\theta_{gl}^T x^{(i)}}} + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k_s} l(y(i)=j)log\frac{e^{\theta_{sj}^T x^{(i)}}}{\sum_{l=1}^{k_s} e^{\theta_{sl}^T x^{(i)}}} \quad (2)$$

where g represents the coarse class and s represents the fine class. $l(y(i)=j)$ represents characteristic function, if $y(i)=j$, $l(y(i)=j)=1$. X represents the depth features of the input image obtained, θ_g and θ_s represents the model parameters in the coarse and fine classifiers respectively, k_g and k_s the number of categories of the coarse-grained class and the fine-grained class respectively.When the loss function (2) is infinitely close to 0, the predicted value is infinitely close to the true value.

D. ORTHOGONAL DISTRIBUTION REGULARIZATION

Considering large-scale training data, the choice of features using orthogonality loss distinguishes $f_g(x)$ and $f_s(x)$. However, the features obtained after several training iterations may over-fit the requirements by different learning tasks, making $f_g(x)$ and $f_s(x)$ too divergent and may cause overfitting. Therefore, orthogonal distribution regularization term is constructed in this paper to limit the distribution of parameters. According to the laws of natural systems, there is a fixed relationship between things. Therefore, the knowledge map constructed in this paper is a fixed tree structure, and the corresponding tree classifier also has a fixed composition. There is a fixed number of fine-grained classes for each coarse-grained class. As shown in Fig.5, there are k parent nodes (coarse genus) and N fine gained child nodes (object classes). Each parent node contains S leaf nodes (S_1, S_2, ……, S_k are different). Therefore, there is a correspondence relation between parent nodes and leaf nodes. We assign leaf nodes into the same parent node according to the commonality of each leaf node. Therefore, we construct the following distribution model to limit the deep feature from being too divergent:

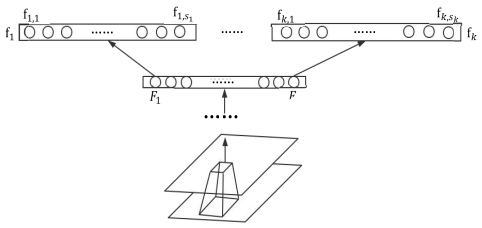


FIGURE 5. The structure of the task-related output layers.

$$F \sim N(0, \frac{1}{\gamma_1}D_g) \qquad f_s \sim N(F_{parent(k)}, \frac{1}{\gamma_2}D_s) \quad (3)$$

where F represents the parameter distribution of the coarse classifier features, which conforms to the normal distribution with a mean of 0 and a variance of $\frac{1}{\gamma_1}D_g$. f_s represents the parameter distribution of the fine classifier feature, and its mean value is the parameter of its parent node. Just like the feature points of $f_{slipper}$ and f_{boot} will be closer F_{shoes} to and the variance is $\frac{1}{\gamma_2}D_s$.

When the feature selection is implemented, the limitation for model can make the partial distribution tend to become whole distribution no matter how to distinguish the fine classifier features from the coarse classifier features, and also make the distribution not too divergent due to over-fitting classification tasks. Similarly, in this paper, like (4), orthogonal distribution regularization is added to the loss function to make the network self-learning:

$$L_3(x) = \min_{w_1, w_2, \ldots, w_s} \beta[(\frac{1}{n}\sum f_s) - F_{parent(k)}] \quad (4)$$

where f_s represents the fine classifier feature of the network; F_{parent} represents the network classifier feature of the network; and β represents the influence factor of the orthogonal distribution regularization on the multi-task network.

IV. EXPERIMENT

Datasets: In this paper, there are two image databases been used to validate the orthogonality loss function and the knowledge map are constructed for each database: (1) Fashion-60, containing 60 costume classes and 5 coarse grained classes. (2) Caltech-UCSD Birds-200-2011, containing 200 fine grained classes and 10 coarse grained classes.

Experiment Environment In this paper, those experiments were performed on a GeForce GTX 1080 GPU. The learning rate was set to 0.01 and multiplied by 0.1 every 40 epochs.

The Basic Architecture of Hierarchical Deep Network In this paper, we used Resnet-18 as the feature extraction network and tree classifier as the task-related output layers to build the multi-task network. We used the Resnet-18 for feature extraction and tree classifier for multi-classification. The loss function is the combination of softmax loss function and orthogonality loss function. When using the multi-label classification on the database, we fuse the softmax losses of two layers.

Compared Baseline Models In this paper, we propose an orthogonality loss function to improve the classification performance of the multi-task network. There are a few baseline models that we can compare with. One is the traditional deep learning network like Alexnet [3], VGG-19 [5] or Resnet-18 [7]. The other one is the standard multi-task network model with Resnet-18. We simply make the standard multi-task network as the baseline. In the experiment, we trained the standard multi-task network added with orthogonality loss function to verify the effectiveness of our proposed method.Compared with the network without adding the loss function, there will be extra computational cost at the same time, but the extra computational cost at the same time are very low. When using the proposed loss function, the network training will be slightly slower, but the testing time will not change.

A. EXPERIMENT WITH FASHION-60

In this section, we apply our proposed method with multiple baseline method on the Fashion-60 database which contains 60 fine grained classes and 5 coarse grained classes. First, the influence of the influence factor α is observed on the experimental results when we add the orthogonality loss function. Then the appropriate value of α is selected to compare the experimental results with the baseline. Finally, we validate whether there is any improvement after the orthogonal distribution regularization is added.

1) THE VALUE OF α

First, we need to select appropriate α for training. Therefore, we have selected 14 different values for α from 0.001-6. The experimental results are shown in Fig.6. As can be seen from the figure, when α is 2.5, the network performs best. When the impact factor is small, the effect of the orthogonality loss function on the network is not obvious. When the value is gradually increased, the performance of the network will gradually decrease, which indicates that the role of the orthogonality loss function will increase to affect the original performance when the value is too large. Therefore, we choose $\alpha = 2.5$ to train the network and compare it with the baseline network.

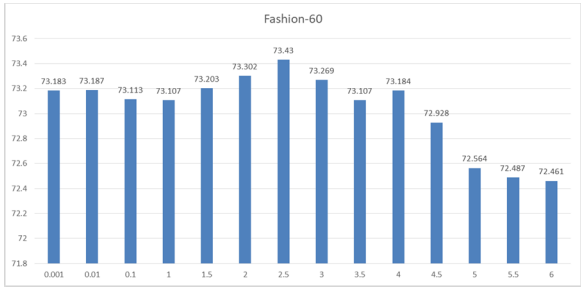


FIGURE 6. The accuracy of different α with Fashion-60.

TABLE 1. The accuracy of some methods with Fashion-60.

Methods	Basic architecture	Fine-classes	Coarse-classes
CNN	Alexnet	68.770	91.162
CNN	VGG-19	70.094	92.259
CNN	Resnet-18	71.517	93.289
Multi-task network	Resnet-18+tree classifier (baseline)	72.410	94.757
Multi-task network	Resnet-18+tree classifier + Orthogonality Loss(α=2.5)	73.430	94.910

2) COMPARISON WITH STATE-OF-ART METHODS

The accuracy of some methods with Fashion-60 are shown in Table 1, the results show that compared with the traditional deep convolutional neural network and a multi-task network without hierarchical orthogonality loss, the multi-task network based on orthogonality loss is better than the other two types of networks in classification. The result proves that the orthogonality loss function proposed in this paper effectively completes the feature selection, making the features obtained in the multi-task network more in line with the task requirements.

Subsequently, we compare the accuracy of each class between baseline and the multi-task network based on orthogonality loss where $\alpha = 2.5$. The comparison of accuracy on the coarse level classification task is shown in Fig.7. The red bars represent the accuracy of the multi-task network based on orthogonality loss, and the blue bars represent the accuracy of the baseline methods. It can be seen from the data in the Fig.7 that in the coarse classification task, the recognition accuracies of each coarse class after adding the orthogonality loss are improved. The experimental results show that the

FIGURE 7. The accuracy of 5 coarse classes on two methods. (The red represents the multi -task network based on orthogonality loss, the blue represents the baseline.)

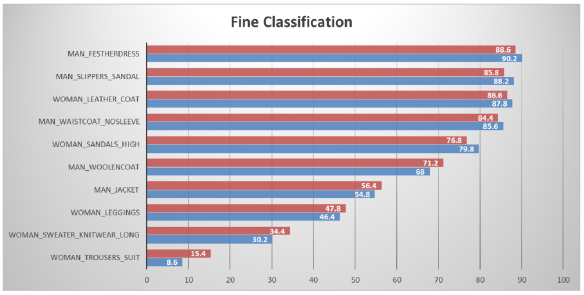


FIGURE 8. The accuracy of 10 fine classes with more obvious changes on two methods. (The red represents the multi -task network based on orthogonality loss, the blue represents the baseline.)

orthogonality loss guides the multi-task network to eliminate the useless feature components in the coarse grained deep feature. We extract the fine-grained classes with more obvious changes in the fine classification, as shown in Fig.8. As can be seen from Table 1, the overall recognition accuracy of the multi-task network has been improved when adding the orthogonality loss. However, from the data with obvious changes extracted in Fig.8, it can be easily found out that in the fine-grained classification task, the recognition accuracy of most fine-grained classes after adding the orthogonality loss is improved, but there are also some classesÍŕ recognition performance is degraded. It can be seen that the recognition accuracy shows a huge differences, but after adding the orthogonality loss, the lower accuracy of some fine classifications has been improved. In comparison, the higher accuracy of some fine classifications has been declined. The reason may be that the orthogonality loss reduces the gap between the accuracies of fine classifications task and makes the network global optimal to improve the overall performance of fine classification tasks by increasing lower fine classification accuracy.

3) ORTHOGONAL DISTRIBUTION Regularization(ODP)

Finally, we observe whether there is any improvement after the orthogonal distribution regularization is added. The parameter β in (4) takes the same value as α in (1). (4) (orthogonal distribution regularization term) is a regular term added to (1) (orthogonal loss function) to limit the distribution of feature parameters and prevent overfitting. Therefore, (4) and (1) need to have the same influence factor on the multi-task network in order to balance the restriction and discrimination. As shown in Table 2, we compared three methods. And compared with the multi-task network added with center loss, our proposed method achieves some improvement in recognition accuracy. According to the data in this table, after adding orthogonal distribution regularization, the performance of the multi-task network has improved on Fashion-60. The experimental results show that not only the orthogonality loss function effectively completes the feature selection, making the features obtained in the multi-task network more in line with the task requirements. And the orthogonal distribution regularization added can also effectively limit the distribution of parameters to avoid overfitting of the network.

TABLE 2. The accuracy of some advanced methods with Fashion-60.

Methods	Accuracy
Triplet loss [58]	66.947
Center Loss [59]	73.010
Orthogonality Loss	73.430
Orthogonality Loss + ODP	73.527

B. FURTHER EXPERIMENT ON CALTECH-UCSD BIRDS-200-2011

In order to verify the effectiveness of the algorithm, we conducted further experiments on the Caltech-UCSD Birds-200-2011, which contains 200 fine classes and 10 coarse classes.

1) THE VALUE OF α

Similarly, we select 14 different values for α from 0.001-6. The experimental results are shown in Fig.9. As can be seen from the figure, the network performs best when α is 2. When the impact factor α is small, the orthogonality loss function is not stable to the network. And when the value gradually increases beyond a certain range, the performance of the network will gradually decrease. Compared with the experimental results on Fashion-60, the optimal value of α is different, but for both types of databases, the overall trend of the effect of α on network performance is the same. Further experimental results show that the value of α may be different in different databases, but the influence of α on network performance is regular. Therefore, it will not work when the value of α is too small, and counteraction appears when the value is too large. We need to find a balance point between orthogonality loss and softmax loss. Therefore, when give for a new dataset, our suggestion for the value of α is to first select a number between 1 and 4.5 for training and check whether the network performance is improved. However, because of the diversity of the database, experiments with values between 0.001 and 10 can further ensure the optimality of the value of α.

2) COMPARE WITH THE BASELINE

In addition, we combine the algorithm mentioned in this paper with the traditional deep convolutional neural

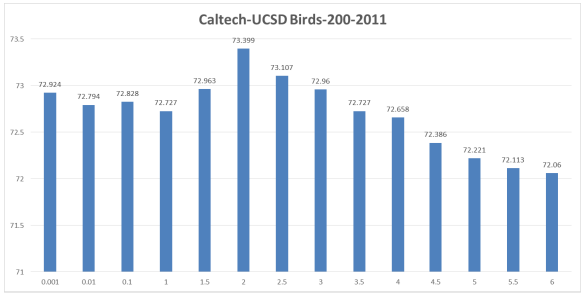


FIGURE 9. The accuracy of different α with Caltech-UCSD Birds-200-2011.

TABLE 3. The accuracy of some methods with Caltech-UCSD Birds-200-2011.

Methods	Basic architecture	Fine-classes	Coarse-classes
CNN	Alexnet	67.683	92.969
CNN	VGG-19	68.816	93.614
CNN	Resnet-18	70.094	94.531
Multi-task network	Resnet-18+tree classifier (baseline)	72.491	96.303
Multi-task network	Resnet-18+tree classifier + Orthogonality Loss(α=2.5)	73.399	96.842

TABLE 4. The accuracy of some advanced methods with Caltech-UCSD Birds-200-2011.

Methods	Accuracy
Triplet loss [58]	65.014
Center Loss [59]	72.879
Orthogonality Loss	73.399
Orthogonality Loss + ODP	73.682

network and a multi-task network without orthogonality loss, the experimental results are shown in Table 3. As can be seen from the data in the Table 3, the multi-task network based on orthogonality loss is better than the other two types of networks in classification on Caltech-UCSD Birds-200-2011. This further proves that the orthogonality loss function proposed in this paper effectively completes the feature selection, making the features obtained in the multi-task network more discriminative.

3) ORTHOGONAL DISTRIBUTION Regularization(ODP)

After adding orthogonal distribution regularization, the performance of the multi-task network has also improved. The parameter β also takes the same value as α. As shown in Table 4,the experimental results on Caltech-UCSD Birds-200-2011 further prove that not only the orthogonality loss function effectively completes the feature selection, making the features obtained in the multi-task network more in line with the task requirements,and also the orthogonal distribution regularization added can effectively limit the distribution of parameters to avoid over-fitting of the network.

V. CONCLUSION

In this paper, a novel loss function-orthogonality loss is proposed to achieve feature selection in multi-task network structure, which helps achieving improvements on image classification. The orthogonality loss can guide the multi-task network extract more specific deep features for different classification tasks and improve the overall classification performance of the whole network. And the orthogonal distribution regularization term is also added to limit the distribution of parameters to reduce the risk of over-fitting. Finally, the results of the classification experiment on Fashion-60 and Caltech-UCSD Birds-200-2011 prove the effectiveness of the proposed algorithm in this paper. It is worth noting that when a new dataset is given, using this method needs to build a knowledge map consisted of two-layers semantic structure firstly, which is used to guide the hierarchical tree classifier to perform multi-level classification tasks.

REFERENCES

[1] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 3, pp. 328–339, Mar. 1989, doi: 10.1109/29.21701.
[2] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.
[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[4] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: https://arxiv.org/abs/1312.4400
[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent*, 2015, pp. 1–14.
[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9, doi: 10.1109/cvpr.2015.7298594.
[7] 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, doi: 10.1109/cvpr.2016.90.
[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708, doi: 10.1109/cvpr.2017.243.
[9] J. Li, D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, "Class-specific differential detection in diffractive optical neural networks improves inference accuracy," *Adv. Photon.*, vol. 1, no. 04, p. 1, Aug. 2019, doi: 10.1117/1.AP.1.4.046001.
[10] S. Jiao, J. Feng, J. Feng, Y. Gao, T. Lei, Z. Xie, and X. Yuan, "Optical machine learning with incoherent light and a single-pixel detector," *Opt. Lett.*, vol. 44, no. 21, pp. 5186–5189, 2019, doi: 10.1364/OL.44.005186.
[11] Z. Kuang, J. Yu, Z. Li, B. Zhang, and J. Fan, "Integrating multi-level deep learning and concept ontology for large-scale visual recognition," *Pattern Recognit.*, vol. 78, pp. 198–214, Jun. 2018, doi: 10.1016/j.patcog.2018.01.027.
[12] S. Bengio, J. Weston, and D. Grangier, "Label embedding trees for large multi-class tasks," in *Proc. Neural Inf. Process. Syst.*, 2010, pp. 163–171.
[13] J. Fan, T. Zhao, Z. Kuang, Y. Zheng, J. Zhang, J. Yu, and J. Peng, "HD-MTL: Hierarchical deep multi-task learning for large-scale visual recognition," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1923–1938, Apr. 2017, doi: 10.1109/TIP.2017.2667405.
[14] J. Yu, Z. Kuang, B. Zhang, W. Zhang, D. Lin, and J. Fan, "Leveraging content sensitiveness and user trustworthiness to recommend fine-grained privacy settings for social image sharing," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 5, pp. 1317–1332, May 2018, doi: 10.1109/TIFS.2017.2787986.
[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Mach. Learn.*, Mar. 2015. [Online]. Available: http://arxiv.org/abs/1503.02531
[16] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 20–28, doi: 10.1109/CVPR.2017.10.
[17] A. Li, T. Luo, Z. Lu, T. Xiang, and L. Wang, "Large-scale few-shot learning: Knowledge transfer with class hierarchy," in *Proc. CVPR*, 2019, pp. 7212–7220.
[18] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping," *Ann. Appl. Statist.*, vol. 6, no. 3, pp. 1095–1117, Sep. 2012, doi: 10.1214/12-aoas549.
[19] H. Zhang, G. He, J. Peng, Z. Kuang, and J. Fan, "Deep learning of path-based tree classifiers for large-scale plant species identification," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Apr. 2018, pp. 25–30, doi: 10.1109/mipr.2018.00013.
[20] H. Zhang, Z. Kuang, X. Peng, G. He, J. Peng, and J. Fan, "Aggregating diverse deep attention networks for large-scale plant species identification," *Neurocomputing*, vol. 378, pp. 283–294, Feb. 2020, doi: 10.1016/j.neucom.2019.10.077.
[21] N. M. Seel, "Metric learning," in *Encyclopedia of the Sciences of Learning*. New York, NY, USA: Springer, 2012, doi: 10.1007/978-1-4419-1428-6_4949.

[22] J. Wang and F. Zhou, "Deep metric learning with angular loss," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2593–2601, doi: 10.1109/ICCV.2017.283.

[23] H. U. Zhengping and G. Zengjie, "Neighborhood repulsed metric learning for kinship verification based on local feature fusion," *Pattern Recognit. Artif. Intell.*, vol. 30, no. 6, pp. 530–537, 2017, doi: 10.16451/j.cnki.issn1003-6059.201706006.

[24] G. C. Stockman, *Computer Vision*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[25] J. Zhang and C. Zong, "Deep learning for natural language processing," in *Deep Learning: Fundamentals, Theory and Applications*. Cham, Switzerland: Springer, 2019, pp. 111–138, doi: 10.1007/978-3-030-06073-2_5.

[26] Y. Zhang, H. Ceylan Koydemir, M. M. Shimogawa, S. Yalcin, A. Guziak, T. Liu, I. Oguz, Y. Huang, B. Bai, Y. Luo, Y. Luo, Z. Wei, H. Wang, V. Bianco, B. Zhang, R. Nadkarni, K. Hill, and A. Ozcan, "Motility-based label-free detection of parasites in bodily fluids using holographic speckle analysis and deep learning," *Light, Sci. Appl.*, vol. 7, no. 1, pp. 1–18, Dec. 2018, doi: 10.1038/s41377-018-0110-1.

[27] S. Jiao, Z. Jin, C. Chang, C. Zhou, W. Zou, and X. Li, "Compression of phase-only holograms with JPEG standard and deep learning," *Appl. Sci.*, vol. 8, no. 8, p. 1258, 2018, doi: 10.3390/app8081258.

[28] J. Yu, Y. Rui, and D. Tao, "Click prediction for Web image reranking using multimodal sparse coding," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2019–2032, May 2014, doi: 10.1109/TIP.2014.2311377.

[29] J. Yu, D. Tao, M. Wang, and Y. Rui, "Learning to rank using user clicks and visual features for image retrieval," *IEEE Trans. Cybern.*, vol. 45, no. 4, pp. 767–779, Apr. 2015, doi: 10.1109/TCYB.2014.2336697.

[30] J. Yu, Y. Rui, and B. Chen, "Exploiting click constraints and multi-view features for image re-ranking," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 159–168, Jan. 2014, doi: 10.1109/tmm.2013.2284755.

[31] Q. Wang, J. Wan, and X. Li, "Robust hierarchical deep learning for vehicular management," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4148–4156, May 2019, doi: 10.1109/tvt.2018.2883046.

[32] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 46–58, Jan. 2020, doi: 10.1109/TPAMI.2018.2875002.

[33] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019, doi: 10.1109/tgrs.2018.2864987.

[34] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 3–13, Jan. 2019, doi: 10.1109/TGRS.2018.2849692.

[35] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 626–640, Mar. 2020, doi: 10.1109/TMM.2019.2931351.

[36] Z. Xia, X. Feng, J. Lin, and A. Hadid, "Deep convolutional hashing using pairwise multi-label supervision for large-scale visual search," *Signal Process., Image Commun.*, vol. 59, pp. 109–116, Nov. 2017, doi: 10.1016/j.image.2017.06.008.

[37] B. Ahn, D.-G. Choi, J. Park, and I. S. Kweon, "Real-time head pose estimation using multi-task deep neural network," *Robot. Auto. Syst.*, vol. 103, pp. 1–12, May 2018.

[38] S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard, "Latent multi-task architecture learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 4822–4829.

[39] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 94–108.

[40] S. Chennupati, G. Sistu, S. Yogamani, and S. A. Rawashdeh, "MultiNet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning," in *Proc. CVPR Workshop*, Ithaca, NY, USA, 2019.

[41] R. Caruana, "Multitask learning: A knowledge based source of inductive bias," in *Proc. 10th Int. Conf. Mach. Learn.*, 1993, pp. 41–48.

[42] B. R. Girshick, "Fast R-CNN," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.

[43] Z. Kuang, Z. Li, T. Zhao, and J. Fan, "Deep multi-task learning for large-scale image classification," in *Proc. IEEE 3rd Int. Conf. Multimedia Big Data (BigMM)*, Apr. 2017, pp. 310–317, doi: 10.1109/BigMM.2017.72.

[44] M. Marszalek and C. Schmid, "Constructing category hierarchies for visual recognition," *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2008, pp. 479–491.

[45] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2D knowledge graph embeddings," in *Proc. AAAI*, San Francisco, CA, USA, 2017.

[46] J. Pujara, H. Miao, L. Getoor, and W. Cohen, "Knowledge graph identification," in *Proc. Int. Semantic Web Conf.* Berlin, Germany: Springer, 2013, pp. 542–557.

[47] W. Xiong, T. H. William, and Y. Wang, "DeepPath: A reinforcement learning method for knowledge graph reasoning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 564–573, doi: 10.18653/v1/d17-1060.

[48] H. Li, H. Li, and L. Zhang, "Quaternion-based multiscale analysis for feature extraction of hyperspectral images," *IEEE Trans. Signal Process.*, vol. 67, no. 6, pp. 1418–1430, Mar. 2019.

[49] Z. Wang, S. Li, Y. Lv, and K. Yang, "Remote sensing image enhancement based on orthogonal wavelet transformation analysis and pseudo-color processing," *Int. J. Comput. Intell. Syst.*, vol. 3, no. 6, pp. 745–753, 2010.

[50] H. Duan and X. Wang, "Echo state networks with orthogonal pigeon-inspired optimization for image restoration," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2413–2425, Nov. 2016.

[51] J. Wen, Z. Tian, X. Liu, and W. Lin, "Neighborhood preserving orthogonal PNMF feature extraction for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 2, pp. 759–768, Apr. 2013.

[52] M. Yang, N.-B. Liu, and W. Liu, "Image 1D OMP sparse decomposition with modified fruit-fly optimization algorithm," *Cluster Comput.*, vol. 20, no. 4, pp. 3015–3022, Dec. 2017.

[53] Y. Chen, X. Jin, J. Feng, and S. Yan, "Training group orthogonal neural networks with privileged information," in *Proc. IJCAI*, 2017, pp. 1532–1538.

[54] C.-P. Lee and C.-J. Lin, "A study on L2-loss (squared hinge-loss) multi-class SVM," *Neural Comput.*, vol. 25, no. 5, pp. 1302–1323, May 2013.

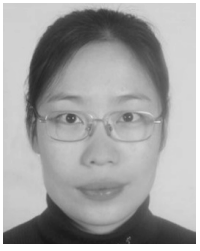
[55] T. Pang, K. Xu, Y. Dong, C. Du, N. Chen, and J. Zhu, "Rethinking softmax cross-entropy loss for adversarial robustness," in *Proc. ICLR*, Feb. 2020. [Online]. Available: http://arxiv.org/abs/1905.10626

[56] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a siamese time delay neural network," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 1993, pp. 737–744.

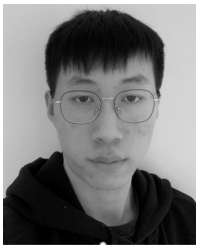
[57] (2011). *Caltech-UCSD Birds-200-2011 Bird Database*. [Online]. Available: http://www.vision.caltech.edu/visipedia/CUB-200-2011.html

[58] E. Hoffer and N. Ailon "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognit*. Cham, Switzerland: Springer, 2015.

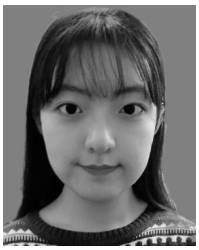
[59] R. Zhang, Q. Wang, and Y. Lu, "Combination of ResNet and center loss based metric learning for handwritten Chinese character recognition," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, 2017, pp. 25–29.



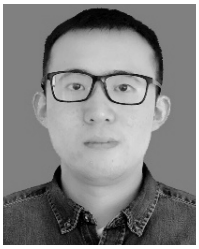
GUIQING HE (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in signal and information processing from Northwestern Polytechnical University, Xi'an, China, in 2000, 2005, and 2008, respectively. She is currently an Associate Professor with the School of Electronics and Information, Northwestern Polytechnical University. Her current research interests include data fusion, analyzing and processing of remote sensing image, machine learning, and computer vision.



YINCHENG HUO received the bachelor's degree in electronic information engineering from Northwestern Polytechnical University, Xi'an, China, in 2018, where he is currently pursuing the master's degree in signal and information processing. His current research interests include speech signal processing and deep learning about object detection.



MINGYAO HE is currently pursuing the B.E. degree with the School of Electronic and Information, Northwestern Polytechnical University, Shaanxi, China. Her research interests include deep learning and image recognition.



HAIXI ZHANG received the B.E. and M.S. degrees from Northwestern Polytechnical University, Xi'an, China, in 2011 and 2013, respectively, where he is currently pursuing the Ph.D. degree with the School of Electronic and Information. His research interests include machine learning, computer vision, and image processing.

JIANPING FAN (Member, IEEE) received the M.S. degree in theory physics from Northwestern University, Xi'an, China, in 1994, and the Ph.D. degree in optical storage and computer science from the Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China, in 1997. From 1997 to 1998, he was a Researcher with Fudan University, Shanghai. From 1998 to 1999, he was a Researcher with the Japan Society of Promotion of Science, Osaka University, Japan. From 1999 to 2001, he was a Postdoctoral Researcher with the Department of Computer Science, Purdue University, West Lafayette, IN, USA. He is currently a Professor with the University of North Carolina at Charlotte. His research interests include image/video privacy protection, automatic image/video understanding, and large-scale deep learning.