IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 16, 2020, accepted November 19, 2020, date of publication November 23, 2020,
date of current version December 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3039862

Efficient Transfer Learning Combined Skip-Connected Structure for Masked Face Poses Classification

SENQIU CHEN[1,3], WENBO LIU[1,3], AND GONG ZHANG[2], (Member, IEEE)
[1]College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211006, China
[2]College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211006, China
[3]Key Laboratory of Ministry of Industry and Information Technology of Non-destructive Testing and Monitoring Technology of High-speed Carrying Facilities, Nanjing 211006, China

Corresponding author: Senqiu Chen (senqiuchen@nuaa.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB2003304, in part by the National Natural Science Foundation of China under Grant 61871218, and in part by the Fundamental Research Funds for the Central University, China under Grant 3082019NC2019002.

ABSTRACT Aiming at the new requirements of masked face poses classification during the epidemic outbreak, this paper proposes an efficient transfer learning approach combined skip-connected structure to improve the accuracy of masked face poses classification in the absence of masked face poses data. We have worked on the following two aspects: 1) According to the features transition of the convolutional neural networks, we propose an efficient transfer learning approach and opt for a more appropriate source domain to solve the problem that the specificity of features in the pre-trained deep networks will damage the performance when transferring to the target domain. First, a semisynthetic masked face poses dataset is constructed to replace ImageNet as the source domain, which can reduce the span of transfer and improve the pertinence of transfer learning. Second, the shallow networks which contain the general features are frozen while the deep networks which contain the specific features are retrained and the entire networks are fine-tuned afterwards. It optimizes the specific features in the source domain when transferring, and promoted transfer learning more effectively; 2) To further improve the overall accuracy by improving the accuracy of masked face pose classes with subtle differences, a skip-connected structure is proposed to fuse general features containing rich detailed information in the shallow networks into the classifier. Experiments on AlexNet and VGG16 show that the proposed method has certain advantages, and the overall accuracy can reach 96.43% and 99.29% at the final respectively.

INDEX TERMS Masked face pose classification, transfer learning, skip-connected structure, detailed feature and semantic feature, deep learning.

I. INTRODUCTION

Since the worldwide outbreak of the novel coronavirus (COVID-19), mankind is suffering from a serious disaster and all fields are making their efforts to fight against the epidemic [1]. AI (artificial intelligence) technologies based on deep learning make many contributions [2], [3], such as face mask detection [4], COVID-19 diagnosis [5]–[8], outbreak forecast [9], and drug research [10], etc. AI provides support with its strong capabilities of information processing and is expected to contribute to epidemic analysis and control, medical care, and vaccine research, etc.

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

In recent years, face pose estimation has become one of the important topics in the field of face information research [11]–[13] with the continuous development of computer vision and intelligent analysis technology. Face pose estimation is a key technology in the field of human behavior analysis, human-computer interaction, and motivation detection, etc. And it has a wide range of application prospects. Because of the COVID-19 epidemic, wearing a mask in public areas has become a common phenomenon and this trend is rising. Therefore, the study of masked face pose estimation has become a new challenge and has important practical significance.

The basic approaches for face pose estimation were comprehensively summarized in the literature [14], and there are

many novel proposed methods recently. Wang *et al.* proposed a novel tree-based neural network architecture which embeds the relationship of the continuity in pose intervals [15]; Li *et al.* combined the task-simplification mechanism and anchor-guided estimation method into one unified learning framework to estimate the face poses [16]; Lee *et al.* proposed a fast and accurate estimation algorithm based on the convolutional random projection forest [17], etc. Although these methods achieve great results, they will be affected more or less if the object is a masked face because there lost a lot of face information when wearing a face mask. In general, the appearance-based methods and the model-based methods are the main categories. Among them, the feature regression method in appearance-based methods has certain advantages and has little demand for face key points or auxiliary information. The key to this type of method is to construct the mapping relationship between image space and pose space. Especially, the face pose estimation methods based on CNNs (convolutional neural networks) [18] have achieved success in low-resolution images with noise, occlusion, and motion blur [17], [19]–[24]. But the lack of sufficient labeled data is a major obstacle that restricts deep learning methods. Due to the impact of COVID-19 epidemic, a large amount of face data collection cannot be achieved temporarily. How to effectively achieve the masked face pose estimation with a small amount of data has become the focus of this paper.

Transfer learning is an outstanding method in the field of few-shot learning. Transfer learning, which studies how to transfer the knowledge learned from the source domain to the target domain, can solve new problems faster and better with a small amount of data and low cost [25]. Transfer learning is an effective method when lacking data, and it has achieved many results in the research of face pose estimation [26]–[29]. These successes motivate us to use transfer learning to solve the problem in this paper.

As the CNNs gradually go deeper, the acquired information transitions from detailed features, such as color blobs and texture feature to complex semantic features. And the semantic features in deep networks show specificity [30], which have particular contributions to the specified tasks in the specific data domain. Therefore, the performance of the pre-trained model in the source domain will be impaired when transferring to the target domain which has different data distribution from the source domain. So we make improvements to solve this problem in two aspects: the source domain and the transfer learning strategy.

ImageNet is a commonly used source domain which has a wide data distribution and can provide rich features to be learned. The networks pre-trained on ImageNet are transferred to the target domain to solve new problems, and there are many successful research results in cancer diagnosis [31], defect detection [32], SAR target detection [33], and other fields lacking massive data. But ImageNet is not the best source domain if considering the data distribution between the source domain and the target domain. Transferring from a source domain with similar data distribution and the same task is more efficient because it can reduce the span between the two domains and improve the specificity of transfer [27], [34], [35].

In addition to opting for a more appropriate source domain, an efficient transfer approach can make transfer learning more effective. The traditional transfer learning approach is widely used, but academics have put forward many improved methods on this basic method recently. Wang *et al.* [36] increased the width and depth of the penultimate classification layer; Tasfia *et al.* [37] added an enhancement layer before the classification layer; Zhao *et al.* [38] proposed a method that transfer learning with the fully pre-trained model; Shermin *et al.* [39] retained the pre-trained classification layer and appended a layer after it, then fine-tuned the additional layer and the classifier. Although many transfer methods are proposed, few academics consider improving the transfer learning effect from the characteristic feature transition of CNNs.

Besides, due to existing the subtle differences of inter-class in the face pose images, we find that the accuracy of masked face pose classes with minor differences is lower. The valid information, which can distinguish such minor differences, exists in the local details of the image [40]. So the detailed information is more important than the semantic information to the classes with subtle differences [41]. But a large amount of detailed information is lost in the deep networks with the transition. To make full use of the image details lost in the deep networks, a skip-connected structure is proposed to fuse the features in the shallow and deep networks.

In this paper we make several contributions:

1) We analyze how features in the shallow and deep networks influence the accuracy of masked face pose classification in the transfer learning process. And we optimize the specific features in deep networks from two aspects: improving the pertinence of the source domain and retraining the specific features. Besides, we analyze the sensitivity of the classes with subtle differences to the detailed information, and the overall accuracy was further improved by increasing the use of detailed information.
2) We construct a semisynthetic masked face pose dataset that has the similar data distribution and the same task with the target domain to replace ImageNet as the source domain. The accuracy and generalization ability of the model are improved by reducing the transfer span and improving the transfer pertinence.
3) According to the features transition from general to specific, we propose a transfer learning approach that is to freeze the shallow convolution parts while retraining the deep convolution parts and fine-tuning the entire networks to recover the co-adaptability between the layers afterwards. It achieves transfer learning more effectively by retaining the general features and optimizing the specific features.
4) A skip-connected structure is proposed to send the general features in shallow networks into the classifier.

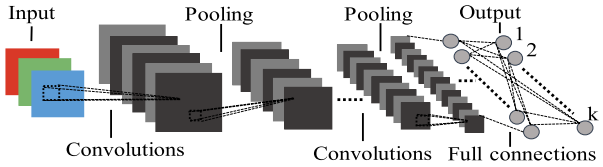

FIGURE 1. Structure of a CNNs used for classification of K classes.

The detailed features lost in the deeper networks are fully used through the proposed structure. The accuracy of masked face pose classes with subtle differences is effectively improved, and the overall accuracy is further improved.

II. METHODOLOGY

A. CONVOLUTIONAL NEURAL NETWORKS

CNNs, a well-known deep learning framework, is widely used in the field of computer vision. The research about CNNs can be traced back to the 1990s [42], but it has attracted much attention when Krizhevsky *et al.* [43] achieved amazing success in the 2012 LSVRC (the ImageNet Large Scale Visual Recognition Challenge). It is constructed by convolutional layers, pooling layers, fully connected layers, and activation function. Fig. 1. gives the structure of CNNs. The features are extracted by a series of alternately stacked convolutional layers and pooling layers, which contain low-level features, such as color blobs and texture features, and high-level semantic features from shallow to deep. And the fully connected layers is a classifier which integrates the features extracted from the last convolutional layer.

With the development of CNNs, many outstanding architectures have been proposed. AlexNet [43], which only contains five convolutional layers, three pooling layers, and three fully connected layers, was proposed by Krizhevsky et al in 2012. AlexNet has excellent classification ability and is still one of the classic and practical architectures so far. Besides, VGG16, which contains five convolution modules, five pooling layers, and three fully connected layers, was proposed by the computer vision group of Oxford University in 2015. AlexNet and VGG16 are commonly used as the base architecture of transfer learning due to their efficient network structure and excellent performance. They have similar structures (five layers/blocks) but one is shallower and another is deeper, so we can observe the effectiveness of proposed methods in two different depths. Besides, the proposed strategy discussed on the typical architecture can get more general results.

B. TRANSFER LEARNING DEFINITION

Transfer learning is an outstanding method in the few-shot learning field, which uses the knowledge learned from the source domain to solve new problems in the target domain [44]. Here we give some definitions, the domain is represented as $D = \{\chi, P(X)\}$, where χ represents the feature space, $P(X)$ represents the marginal distribution of $X = \{x_1, x_2, \cdots x_n\} \in \chi$, $x_i \in R^d$, $i = 1, 2, \ldots n$. Given a specific domain, a task is represented as $\Gamma = \{Y, f(\cdot)\}$,

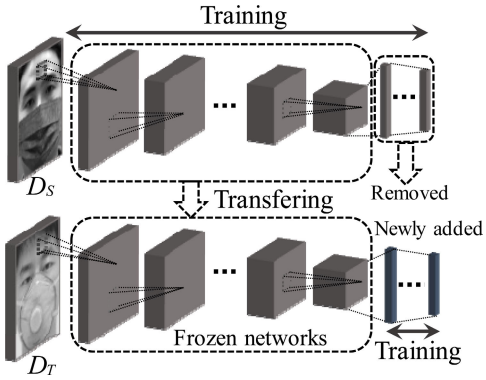

FIGURE 2. Schematic diagram of the traditional transfer learning.

Where Y represents the label space and $f(\cdot)$ represents a target prediction function. A task Γ is learned from the training data (x_i, y_i), $x_i \in X$, $y_i \in Y$. The transfer learning is defined as: Given a source domain D_S and source task Γ_S, a target domain D_T and target task Γ_T, the knowledge acquired from the source task Γ_S on the source domain D_S is used to solve the new task Γ_T on the target domain D_T, where $D_S \neq D_T$ and/or $\Gamma_S \neq \Gamma_T$.

C. PROPOSED TRANSFER LEARNING STRATEGY

1) PROBLEMS OF TRADITIONAL TRANSFER LEARNING METHOD AND SOLUTIONS IN THIS PAPER

The traditional method of transfer learning is to freeze the pre-trained convolutional base as a feature extractor and remove the existing classifier firstly, then add and train a new classifier on the target domain, Fig. 2. shows the traditional transfer learning process. Such a transfer learning method is clear and easy to implement, but it ignores the feature transition mechanism of CNNs. With the continuous deepening of networks, the extracted features gradually transition from detailed features, such as color blobs and texture features to abstract and semantic features. The features in the shallow networks are general but the semantic features in the deep networks are specific when transferring [30]. The semantic features make particular contributions to the specified tasks in the specific data domain, which means they are more suitable for solving specific tasks than general features. The semantic features that are specific to the source domain will inevitably hurt the transfer effect if the entire convolutional base is transferred to the target domain without any improvements.

Optimizing the features in deep convolution parts is one of the important improvements to enhance the transfer effect and there are three solutions. The first solution is not to transfer the deep convolution parts of the pre-trained model and train these parts from scratch on the target domain; the second solution is to opt for a more appropriate source domain which has a similar data distribution and the same task as the target domain; the third solution is to retrain the deep convolution parts of the pre-trained model on the target domain. Among them, the first solution cannot achieve a great effect, because the ability of CNNs' feature extraction is severely damaged

when the amount of training data is too small. In this paper, we combine the second and the third solutions to propose a new solution, which is opting for a more pertinent source domain and retraining the deep convolution parts on the target domain. The following will be discussed in detail.

2) IMPROVEMENT OF PERTINENCE OF THE SOURCE DOMAIN

ImageNet, which contains about 15 million pictures and 22,000 classes, is commonly used as the source domain. ImageNet can provide sufficient and extensive data. The subset of ImageNet used in the ISLVRC contains about 1.33 million images in 1,000 classes, covering common objects such as faces, cars, and pedestrians, as well as a large number of rare objects [45]. Although ImageNet can provide rich data, its data distribution and tasks are not consistent with the target domain. ImageNet is relatively extensive whereas the target domain is specific. Therefore, ImageNet lacks pertinence during transferring, and the task in the target domain cannot be solved specifically. Besides, Kornblith pointed out in the literature [46] that transferring a model pre-trained on ImageNet to a small dataset lacks adaptability and cannot achieve a good effect, especially in fine-grained image classification tasks. So the transfer effect will be not efficient if the model is pre-trained on ImageNet in the task of masked face poses classification.

In order to improve the pertinence of transfer learning, we construct a semisynthetic masked face pose dataset as the source domain. The semisynthetic dataset has a highly similar data distribution and the same task to the real dataset which is the target domain. The semisynthetic dataset reduces the span between the source and target domain, and overcomes the inadequate performance of the model pre-trained on ImageNet in fine-grained image classification task. By improving the pertinence of the source domain and reducing the span of transfer, the effect is improved.

3) A NEW TRANSFER LEARNING METHOD

A more appropriate source domain is one aspect of improving the transfer effect, and an efficient transfer learning method is another. An efficient transfer learning method can further improve the model effect. As discussed above, the features in shallow networks are general and the features in deep networks are specific. The general features can be easily transferred to the target domain, and adapt to the target domain without damaging the model performance. However, the specific features only promote model performance in the specific data domain with specific tasks. When transferring between two different domains, the features specific to the source domain can inevitably damage the performance of the model that is transferred to the target domain. In this paper, the specificity of features in the deep networks is optimized by retraining these specific features on the target domain. Due to the similarity of data distribution between the semisynthetic and real masked face pose dataset, retraining in the target domain can be more easily complicated and the specificity of features in deep networks can be optimized to a large extent. Besides, there is a close co-adaptability between the convolutional layers. The freeze operation breaks the co-adaptability between the frozen part and unfrozen part. So the entire networks need to be unfrozen and fine-tuned to recover the co-adaptability after retraining.

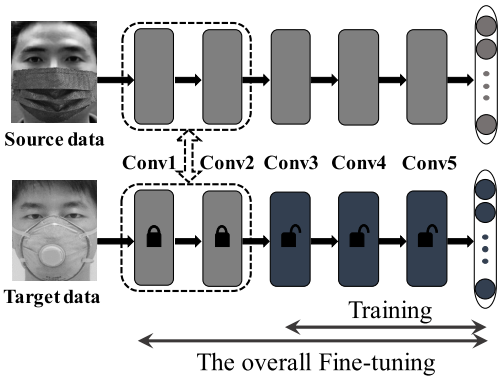

FIGURE 3. Schematic diagram of the proposed transfer learning method.

Our experiments are on AlexNet/VGG16 architecture, which has five convolutional layers/blocks. We hierarchically freeze the convolution layers/blocks from shallow to deep for transfer learning. The experiment results show that the effect of freezing the first two convolutional layers/blocks is the best. Fig. 3. shows the schematic diagram of the proposed method. First, we transfer the entire convolutional base which is pre-trained on the semisynthetic dataset, and freeze the shallow convolutional parts (Conv1 and Conv2 layers/blocks). Then, we retrain the deep convolutional parts (Conv3, Conv4, and Conv5 layers/blocks) on the basis of pre-trained parameters. Finally, we unfreeze and fine-tune the entire networks to recover the co-adaptability between layers.

D. PROPOSED SKIP-CONNECTED STRUCTURE

Effective transfer learning strategy is an important means to improve the accuracy, but we find that the further improvement of the overall accuracy is subject to the accuracy of the masked face pose classes with subtle differences. There are existing subtle differences in inter-class, resulting in difficult distinction and low accuracy. Therefore, improving the accuracy of pose classes with subtle differences is another key point to improve the overall accuracy.

For fine-grained image classification, the local detailed information has an excellent ability to distinguish such subtle differences of the inter-class. So detailed information is more important than semantic information to distinguish the minor differences, and classification can be effectively performed with the help of local detailed information. Making full use of the local detailed information of CNNs is an important means to improve the accuracy. With the deepening of the networks, the features extracted by CNNs gradually transition from the spatial details, such as color blobs and textures features to the high-level semantic features. Fig. 4. shows a part of features visualization results, which are extracted by VGG16. It can be seen intuitively from the results that the features in

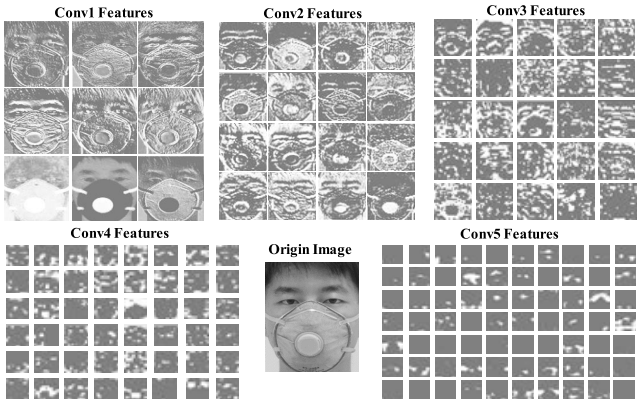

FIGURE 4. Features visualization of partial convolutional layers.

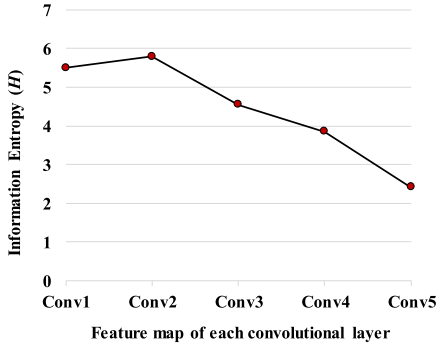

FIGURE 5. Information entropy of features in each CNNs layer.

shallow networks contain more detailed information, whereas the features in deep networks are more abstract, especially the last convolutional layer has the strongest and the most abstract semantic information. Based on the information theory, we know that the amount of information that an image conveys can be expressed by the image entropy. Given $f(x, y)$ as an image with k grayscale, the probability of i grayscale is p_i, the entropy of an image is:

$$H = -\sum_{i=0}^{k-1} p_i \cdot \log(p_i) \tag{1}$$

where H is the entropy, p_i is the probability with i ($i = 1 \sim k$) grayscale, $k = 255$.

The entropy is larger, the more detailed information is conveyed, and vice versa. An image with a larger entropy contains more detailed information, whereas a smaller entropy means more detailed information is lost and the semantic information is more highlighted. Fig. 5. is the information entropy curve of the features extracted by VGG16. It can be seen that as the convolutional layers/blocks are gradually deepened, the entropy is continuously decreased. So the deeper the network, the more detailed information is lost. A large amount of spatial information is lost, resulting in the lack of detailed features in the MLP (Multi-Layer Perceptron, the fully connected classifier). Besides, it can be seen from the curve that the entropy of Conv2 is the largest, indicating that it contains the most detailed information. A large amount

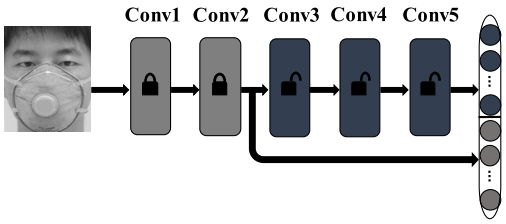

FIGURE 6. Schematic diagram of the proposed skip-connected structure.

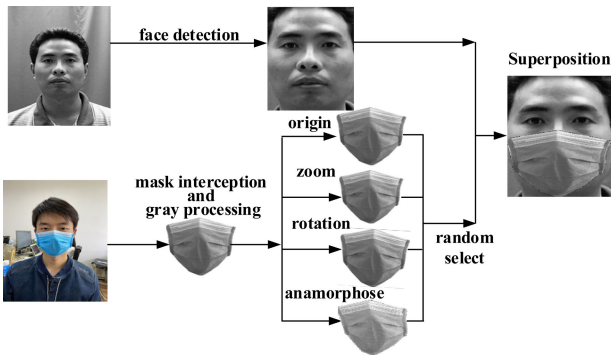

FIGURE 7. Flow chart of the synthetic images generation.

of detailed information is acquired and retained by multiple convolution kernels in the shallow networks. To compensate for the local detail information lost in the deep networks and make it full use, we design a skip-connected structure to solve this problem. The network structure diagram is shown in Fig. 6. The features in the shallow networks Conv2 and the features of the last layer of Conv5 are flattened and sent into the MLP together. The detailed features and the semantic features are merged. Therefore, the problem of low accuracy in the classes with subtle differences is solved by supplementing detailed information, and the overall model accuracy rate is further improved.

III. DATASET GENERATION

To simulate the masked face pose images more realistically, we make the semisynthetic data by superimposing the general face pose images and the mask images. The general face poses dataset comes from the CAS-PEAL-R1 dataset [47] created by the Institute of Computer Technology, Chinese Academy of Sciences. And the subset of this dataset in the yaw direction is used in this paper. Fig. 7. shows the production process. There is a large amount of background and non-face parts in the general face pose images. Considering these factors may affect the performance, we use a face object detector to preprocess the general face pose images to filter out the background and non-face parts firstly; then we appropriately perform scaling, rotation, and deformation operations on the mask images to enrich the diversity of mask images; finally, the images of each general face pose and the mask images in the corresponding pose are superimposed to complete the generation of the semisynthetic data. The semisynthetic dataset constructed in this paper includes images of 1040 people wearing masks in seven poses in the yaw direction ($-67°$, $-45°$, $-22°$, $0°$, $+22°$, $+45°$, $+67°$),

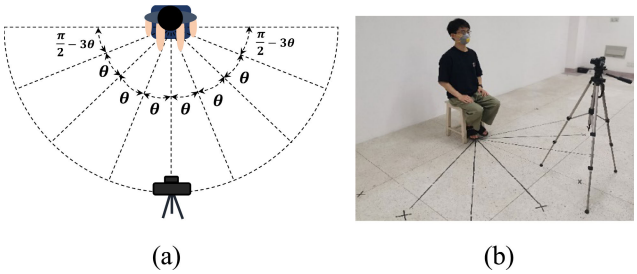

FIGURE 8. The display of face pose images collection. (a) Illustration of the acquisition configuration, θ is about 22.5°. (b) Setup of the photographic room.

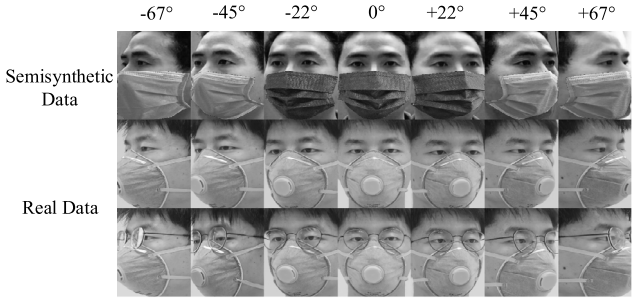

FIGURE 9. The display of the semisynthetic and real masked face pose samples. the first row is semisynthetic samples, the second and third rows are real samples, there are 7 pose classes in the yaw direction.

each pose has 1040 samples, a total of 7280 pictures, and the size is 200 × 240px.

Due to the impact of the COVID-19 epidemic, we cannot collect a large amount of face data temporarily. In order to keep the pose consistent with the semisynthetic data, we build an image acquisition environment as shown in Fig. 8. And we collect images of 57 people wearing masks in corresponding poses.

The collected face images still include a large amount of background and non-face parts, so the face target detector is also required to preprocess the images firstly. Secondly, considering people wearing glasses, the constructed dataset includes faces with glasses and faces without glasses. The real face pose dataset constructed in this paper includes 114 samples in each class, the image size is 200 × 240px, and there are 798 samples in total. The sample examples of the constructed semisynthetic dataset and the real dataset are shown in Fig. 9.

IV. EXPERIMENTS AND ANALYSIS

A. DATA PREPARATION AND EXPERIMENT SETTINGS

1) DATA PREPARATION

The semisynthetic dataset includes the masked face pose images of 1040 people in 7 different poses, a total of 7280 images. 5180 images are used as the training samples, and the remaining 2100 images are used as the testing samples. The real dataset includes the masked face pose images of 57 people in the same pose, a total of 798 images. And 658 images are used as the training samples, and the remaining 140 images are used as the testing samples. The image size is uniformly resized to 128 × 128px to meet the input requirements of the networks, and the data enhancement operations such as brightness transformation, noise addition, and blurring are randomly performed on the images to enhance the generalization ability of the model. The added noise is salt and pepper noise and Gaussian noise with a mean value of 0 and a variance of 0.002; the brightness is transformed to 0.5 times and 1 times as the original; the mean blur filter is used for the blur operation.

2) EXPERIMENT SETTINGS

In addition, training skills and methods are also important. The experiment adopts the stochastic gradient descent method (SGD) as the training algorithm, the momentum is set to 0.9, the weight decay is set to 0.0005; we use the variable learning rate method to make the error convergence more stable. The learning rate is reduced to 1/10 of the original when the iteration reaches 2/3 of the training epoch; batch size is set to 16; and the loss function is the cross entropy. The epochs of training from scratch on the semisynthetic dataset are set to 50, and the initial learning rate is 0.005. The epochs of transfer training on the real dataset are set to 10, the initial learning rate is 0.0008. And the epochs of fine-tuning are set to 5, the initial learning rate is 0.0001.

The hardware and software experimental platform is PC, the operating system is Windows10, the CPU is Core i7-9750H with 8GB memory, the GPU is NVIDIA GeForce GTX 1650 with a 4G memory. The code is written in Python and is based on the Pytorch deep learning framework.

B. RESULTS AND ANALYSIS

The overall accuracy (OA), which is the percentage of all the correctly classified samples and can represent the overall performance of a model, is used to evaluate the performance of different methods, the OA is defined as:

$$OA = \frac{K}{N} \times 100\% \tag{2}$$

where N is the number of testing samples, K is the number of samples that are correctly classified.

1) SOURCE DOMAIN EXPERIMENTS AND ANALYSIS

First, the experiments based on AlexNet and VGG16 are designed to verify the superiority of transfer learning on the semisynthetic dataset.

We set up the following controlled methods: the networks are trained from scratch on the semisynthetic dataset without transfer learning. AlexNet-TSS (AlexNet, training from scratch on the semisynthetic dataset) and VGG16-TSS (VGG16, training from scratch on the semisynthetic dataset); the networks pre-trained on ImageNet do the transfer learning on the real dataset. AlexNet-TITR (AlexNet, training on ImageNet and transfer learning on the real dataset), VGG16-TITR (VGG16, training on ImageNet and transfer learning on the real dataset); And the networks pre-trained on the semisynthetic dataset do the transfer learning on the real dataset (what we propose). AlexNet-TSTR (AlexNet, training

TABLE 1. Overall accuracy of different methods.

Methods	OA(%)
AlexNet-TSS	41.43
AlexNet-TITR	68.45
AlexNet-TSTR(Ours)	79.29
VGG16-TSS	67.86
VGG16-TITR	80.84
VGG16-TSTR(Ours)	90.16

on the semisynthetic data and transfer learning on the real dataset), VGG16-TSTR (VGG16, training on the semisynthetic dataset and transfer learning on the real dataset), all methods are tested on the real dataset in the same environment. The experimental results are shown in Table 1.

From Table 1, we can see that:

1) compared with the transfer learning method (-TITR and -TSTR) and the none transfer learning method (-TSS), the OA of the methods with transfer learning is higher. The reason why the -TSS method has a lower OA is that there are existing differences of data distribution between the real data and the semisynthetic data, especially the masked part of semisynthetic images lacks the same rich and changeable style as the real data. Using transfer learning, whether the source domain is ImageNet or the semisynthetic dataset, the OA can be significantly improved. So the model generalization is enhanced through transfer learning in the real scenarios.
2) The effect of transfer learning from semisynthetic dataset (-TSTR) is better than that from ImageNet (-TITR) because the semisynthetic masked face pose dataset as the source domain has more pertinence to the target domain. Even if there is a little difference between the real data and the semisynthetic data, both datasets are similar and have the same task. Besides, the model pre-trained on ImageNet lacks adaptability when transferring to a small dataset, especially in fine-grained image classification tasks such as face pose image classification.
3) Both efficient networks AlexNet and VGG16 have excellent performance in the -TSTR method. But the OA of VGG16 is higher than AlexNet because VGG16 has a more complex network structure which can extract more diverse and complex features. Therefore, the effect of VGG16 is better than AlexNet.

2) TRANSFER LEARNING METHOD EXPERIMENTS AND ANALYSIS

We discuss the proposed transfer learning method on the basis of using the semisynthetic dataset as the source domain. The model that needs to be transferred is divided into two parts: the convolutional base and the MLP. The MLP can be directly transferred to the target domain for training and fine-tuning without any modification because the source task has the same number of classifications as the target task. The convolutional base is used as a feature extractor and we focus on studying its performance.

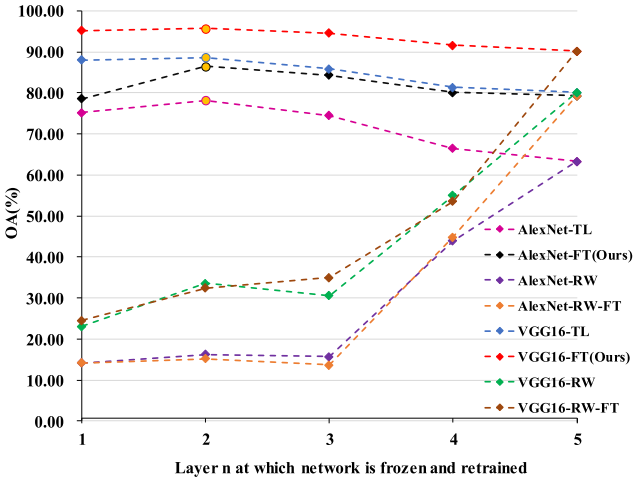

FIGURE 10. The results display of different methods with the hierarchically frozen training.

We set up the following controlled methods by hierarchically freezing convolutional layers/blocks (Conv1, Conv2, Conv3, Conv4, Conv5): transfer learning method (-TL); transfer learning with fine-tuning method (-FT); transfer learning with random weight method (-RW); and transfer learning with fine-tuning and random weight method (-RW- FT). The -TL method is that the parameters of the convolutional layers/blocks that need to be frozen are fixed, the remaining unfrozen part and the MLP are retrained on the target domain; the -FT method (ours) is based on -TL method, which is to unfreeze and fine-tune the entire networks after the -TL method; the -RW method is also based on the -TL method, but the difference is that the frozen part can keep the pre-trained parameters while the unfrozen part are set to random parameters; the -RW-FT method is to unfreeze and fine-tune the entire networks after the -RW method.

Fig. 10. shows the experimental results, we can see that:

1) From the four curves in the upper half of Fig. 10., it can be seen that all the methods in which Conv1 and Conv2 are frozen while Conv3, Conv4 and Conv5 are retrained can achieve the best effect (the yellow dot). And as the number of frozen layers/blocks increases, the effect gradually decreases. Conv1 and Conv2 are almost well transferred to the new domain, indicating that the features of the first two layers/blocks are general and have a positive effect on new tasks; the performance decreases when the layer/block is frozen to Conv3, and the decrease is more significant when the layer/block is frozen to Conv4 and Conv5. This is because the features in Conv3, Conv4 and Conv5 are little general but more specific, the features that are specific to the semisynthetic dataset lack adaptation in the real dataset. The method, which is to freeze Conv1 and Conv2 while retraining Conv3, Conv4 and

Conv5, can optimize the specific features to adapt to the real dataset.

2) From the comparison of the red curve with the blue curve and the black curve with the rose red curve, we know that the -FT method is superior to the -TL method. The OA of the -FT method is 3.43∼16.83% higher than the -TL method. The networks are divided into the frozen part and the unfrozen part, and the freezing operation breaks the co-adaptability existing between the convolutional layers. So the entire networks need to be unfrozen and fine-tuned after training the deep convolutional parts. Recovering the co-adaptability can further improve the performance of the networks.
3) From the comparison of the four curves in the upper half and the four curves in the lower half, it can be seen that the network performance is damaged severely when the unfrozen part is reset with random parameters. In particular, the OA is dramatically decreased when the first three layers are frozen. Setting the networks with random parameters is equivalent to training these parts from scratch on the new dataset. When the training dataset is small, the ability of feature extraction is not established and the convolution base cannot get rich and robust features. So the network performance is greatly decreased; Comparing the brown curve with the green curve and the purple curve with the yellow curve, it can be seen that the fine-tuning can hardly improve the performance. This is because the co-adaptability between the frozen part and the unfrozen part does not exist when the unfrozen part is not fully trained; the OA has a certain degree of rebound when only Conv5 is reset to random parameters and can achieve the highest when the entire networks are frozen. This phenomenon also shows that retaining pre-trained weights has a positive effect on performance during transferring and training from scratch on a small dataset cannot get a good result.

3) SKIP-CONNECTED STRUCTURE EXPERIMENTS AND ANALYSIS

We can see from Table 2, the OA of the -FT method based on AlexNet (AlexNet-FT) can reach 86.42%, and that of VGG16 (VGG16-FT) can reach 95.71%. But we are curious about how the accuracy of each pose class, so we give the accuracy of each class (CA) to evaluate the accuracy of each masked face pose class in different methods. The CA is defined as:

$$CA = \frac{K_i}{N_i} \times 100\% \tag{3}$$

where N_i is the number of testing samples of class i, K_i is the number of correctly classified samples of class i, $i = 1, 2, ...7$.

In our further research, we find that the CA of classes with subtle differences is lower. Fig. 11. shows that whether it is based on AlexNet or VGG16, the CA in attitudes ±45° and ±22° is lower than that of ±67° and 0°. AlexNet has a simpler structure than VGG16, its CA decreases more significantly. The face images in attitudes ±67° and 0° are obviously different from other poses, whereas the face images in attitudes ±45° and ±22° have subtle differences to be hardly distinguished, so the CA in ±45° and ±22° is lower. It can be seen from Fig. 12. that the ±45° and ±22° attitudes are easily misclassified as similar attitudes. For example, (a) in Fig. 12. shows that there are 13 correctly classified as −45° attitude, 5 samples are incorrectly classified as the −22° attitude and 3 samples are incorrectly classified as the −67° attitude, the total number of test samples of −45° attitude was 20. The further improvement of the OA is subject to the CA of classes with subtle differences, so solving this problem is the key to further improving the OA.

TABLE 2. Overall accuracy of different methods.

Methods	OA(%)
AlexNet-FT	86.42
AlexNet-SF	59.17
AlexNet-FT-FF(Ours)	96.43
VGG16-FT	95.71
VGG16-SF	77.72
VGG16-FT-FF(Ours)	99.29

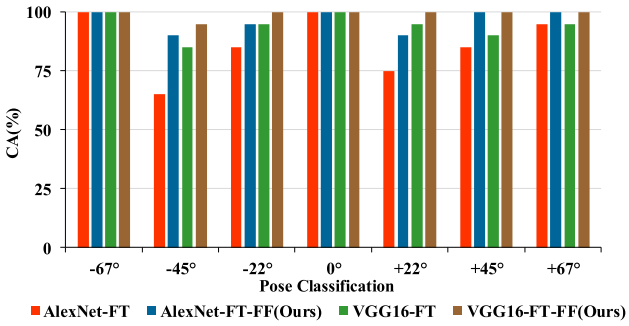

FIGURE 11. Each classification accuracy of different methods.

As mentioned in section D of the methodology, detailed information is more important than semantic information, and the local detailed information can effectively improve the accuracy of classification. We supplement the lost detailed information into the MLP by the skip-connected structure. The designed networks are AlexNet-FT-FF (AlexNet-FT with Feature Fusion) and VGG16-FT-FF (VGG16-FT with Feature Fusion).

Fig. 11. shows that the CA in ±45° and ±22° attitudes are improved by the skip-connected structure (AlexNet-FT and VGG16-FT). The accuracy of the classes with subtle differences is increased by 5%-25%, the improvement of AlexNet is more obvious. By comparing (a) with (b) and (c) with (d) in Fig. 12, we know that the misclassification in attitudes ±45° and ±22° is obviously decreased. The skip-connected structure effectively solves the problem of low accuracy in the classes with subtle differences. From

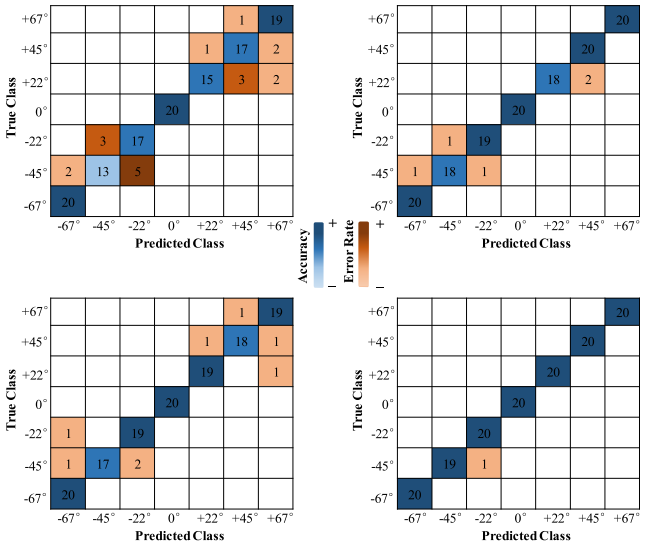

FIGURE 12. Confusion matrices for the different methods. (a) is in the first column of the first row (AlexNet-FT method). (b) is in the second column of the first row (AlexNet-FT-FF method). (c) is in the first column of second first row (VGG16-FT method). (d) is in the second column of second first row (VGG16-FT-FF method).

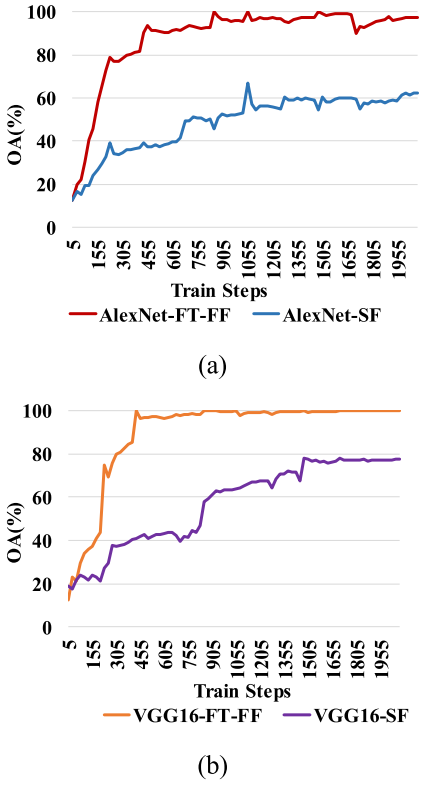

FIGURE 13. Convergence of networks training process. (a) is based on AlexNet and (b) is based on VGG16.

the OA of AlexNet-FT-FF and VGG16-FT-FF in Table 2, it can be seen that the overall performance of the networks with skip-connected structure is improved. The OA of VGG16 increases from 95.71% to 99.29%, an increase of 3.58%; the OA of AlexNet increases from 86.42% to 96.43%, an increase of 10.01%. The enhancement of AlexNet is more obvious, although AlexNet does not have the same excellent structure as VGG16. Therefore, the results prove the effectiveness of the skip-connected structure and the performance can be further improved by the skip-connected structure.

From the above theories and experiments, we know that face pose classification is sensitive to detailed information such as textures features, Gabor filters, color blobs, etc. Fusing the detailed features into the MLP can effectively improve the performance of the network. This makes us wonder if we can get the same good result by just sending the Conv2 features into the MLP? We set up two networks, AlexNet-SF (AlexNet with shallow features) and VGG16-SF (VGG16 with shallow features) to verify what we guess. Specifically, only the Conv2 features of AlexNet or VGG16 are sent to the MLP, and the features of Conv5 are discarded. Fig. 13. shows the training process. Fig. 13 shows that the performance of the model without Conv5 features is severely damaged whether in AlexNet or VGG16. The training curve of AlexNet-FT-FF and VGG16-FT-FF can converge well, and the models achieve high accuracy at about 500 iterations. However, the convergence curves of AlexNet-SF and VGG16-SF are more twists and turns. The models seem difficult to converge and get a low accuracy at the final. Table 2 shows that the OA of AlexNet-SF is 59.17%, and the OA of VGG16-SF is 77.27%, which is severely decreased compared to AlexNet-FT-FF and VGG16-FT-FF. So we can summarize that although the general detailed information is beneficial to the classes with subtle differences, performance will be severely damaged if high-level semantic information is discarded. The results show that the general detailed features are sensitive to the classes with subtle differences, but the high-level semantic features are also important to the classification.

V. CONCLUSION

The focus of this paper is that we propose an efficient transfer learning strategy combined skip-connected structure to fulfill the new requirements for masked face pose classification in the absence of data. In our work, we analyze how features in the shallow and deep networks influence the accuracy of masked face pose classification in the transfer learning process firstly. Then we make improvements in the source domain and the transfer learning approach for solving the specificity problem. In the comparative experiment, it is proved that the semisynthetic dataset as the source domain can improve the pertinence of transfer learning. Besides, the proposed transfer learning approach can optimize specific features by retraining them on the target domain. Finally, we propose a skip-connected structure to send detailed features in the shallow networks into the MLP, which further improves the overall accuracy by effectively improving the accuracy of the classes with subtle differences. The experimental results illustrate the importance of feature fusion and the effectiveness of the skip-connected structure.

The method proposed in this paper effectively achieves masked face pose estimation, but it is specific to the masked face objects. In the future, it may be a common phenomenon that faces with and without masks coexist in public areas because of the improvement of people's awareness of health. Therefore, it is practical to propose a general and efficient method of face pose estimation. In future work, we aim to study the pose estimation method that is suitable for the objects of faces with and without masks together.

REFERENCES

[1] V. Chamola, V. Hassija, V. Gupta, and M. Guizani, "A comprehensive review of the COVID-19 pandemic and the role of IoT, drones, AI, blockchain, and 5G in managing its impact," *IEEE Access*, vol. 8, pp. 90225–90265, 2020.

[2] A. A. Hussain, O. Bouachir, F. Al-Turjman, and M. Aloqaily, "AI techniques for COVID-19," *IEEE Access*, vol. 8, pp. 128776–128795, 2020.

[3] Q.-V. Pham, D. C. Nguyen, T. Huynh-The, W.-J. Hwang, and P. N. Pathirana, "Artificial intelligence (AI) and big data for coronavirus (COVID-19) pandemic: A survey on the state-of-the-arts," *IEEE Access*, vol. 8, pp. 130820–130839, 2020.

[4] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic," *Meas., J. Int. Meas. Confederation*, vol. 167, May 2020, Art. no. 108288.

[5] M. Jamshidi, A. Lalbakhsh, J. Talla, Z. Peroutka, F. Hadjilooei, P. Lalbakhsh, M. Jamshidi, L. La Spada, M. Mirmozafari, M. Dehghani, A. Sabet, S. Roshani, S. Roshani, N. Bayat-Makou, B. Mohamadzade, Z. Malek, A. Jamshidi, S. Kiani, H. Hashemi-Dezaki, and W. Mohyuddin, "Artificial intelligence and COVID-19: Deep learning approaches for diagnosis and treatment," *IEEE Access*, vol. 8, pp. 109581–109595, Dec. 2020.

[6] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "CovidGAN: Data augmentation using auxiliary classifier GAN for improved COVID-19 detection," *IEEE Access*, vol. 8, pp. 91916–91923, 2020.

[7] S. Hu, Y. Gao, Z. Niu, Y. Jiang, L. Li, X. Xiao, M. Wang, E. F. Fang, W. Menpes-Smith, J. Xia, H. Ye, and G. Yang, "Weakly supervised deep learning for COVID-19 infection detection and classification from CT images," *IEEE Access*, vol. 8, pp. 118869–118883, 2020.

[8] X. Ouyang, J. Huo, L. Xia, F. Shan, J. Liu, Z. Mo, F. Yan, Z. Ding, Q. Yang, B. Song, F. Shi, H. Yuan, Y. Wei, X. Cao, Y. Gao, D. Wu, Q. Wang, and D. Shen, "Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2595–2605, Aug. 2020.

[9] Z. Hu, Q. Ge, S. Li, E. Boerwincle, L. Jin, and M. Xiong, "Forecasting and evaluating intervention of covid-19 in the world," 2020, *arXiv:2003.09800*. [Online]. Available: http://arxiv.org/abs/2003.09800

[10] V. Chenthamarakshan, P. Das, S. C. Hoffman, H. Strobelt, I. Padhi, K. Wai Lim, B. Hoover, M. Manica, J. Born, T. Laino, and A. Mojsilovic, "CogMol: Target-specific and selective drug design for COVID-19 using deep generative models," 2020, *arXiv:2004.01215*. [Online]. Available: http://arxiv.org/abs/2004.01215

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, vol. 3, 2015, pp. 41.1–41.12.

[12] N. Otberdout, A. Kacem, M. Daoudi, L. Ballihi, and S. Berretti, "Automatic analysis of facial expressions based on deep covariance trajectories," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 3892–3905, Oct. 2020.

[13] H. Jung, S. Lee, S. Park, I. Lee, C. Ahn, and J. Kim, "Deep temporal appearance-geometry network for facial expression recognition," 2015, *arXiv:1503.01532*. [Online]. Available: http://arxiv.org/abs/1503.01532

[14] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.

[15] J. Wang, F. Ullah, Y. Cai, and J. Li, "Non-stationary representation for continuity aware head pose estimation via deep neural decision trees," *IEEE Access*, vol. 7, pp. 181947–181958, 2019.

[16] J. Li, J. Wang, and F. Ullah, "An end-to-end task-simplified and anchor-guided deep learning framework for image-based head pose estimation," *IEEE Access*, vol. 8, pp. 42458–42468, 2020.

[17] D. Lee, M.-H. Yang, and S. Oh, "Head and body orientation estimation using convolutional random projection forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 107–120, Jan. 2019.

[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[19] B. Ahn, J. Park, and I. S. Kweon, "Real-time head orientation from a monocular camera using deep neural network," in *Proc. Asian Conf. Comput. Vis.*, in Lecture Notes in Computer Science, vol. 9005, 2015, pp. 82–96.

[20] M. Patacchiola and A. Cangelosi, "Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods," *Pattern Recognit.*, vol. 71, pp. 132–143, Nov. 2017.

[21] M. Raza, Z. Chen, S.-U. Rehman, P. Wang, and P. Bao, "Appearance based pedestrians' head pose and body orientation estimation using deep learning," *Neurocomputing*, vol. 272, pp. 647–659, Jan. 2018.

[22] Y. Lu, S. Yi, N. Hou, J. Zhu, and T. Ma, "Deep neural networks for head pose classification," in *Proc. World Congr. Intell. Control Autom.*, Sep. 2016, pp. 2787–2790.

[23] J. Bao and M. Ye, "Head pose estimation based on robust convolutional neural network," *Cybern. Inf. Technol.*, vol. 16, no. 6, pp. 133–145, Dec. 2016.

[24] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2155–2164.

[25] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[26] P. Li, Y. Li, and L. Tan, "Transfer useful knowledge for headpose estimation from low resolution images," *Multimedia Tools Appl.*, vol. 75, no. 15, pp. 9395–9408, Aug. 2016.

[27] Y. Liu, P. Lasang, S. Pranata, S. Shen, and W. Zhang, "Driver pose estimation using recurrent lightweight network and virtual data augmented transfer learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3818–3831, Oct. 2019.

[28] Y. Yan, R. Subramanian, O. Lanz, and N. Sebe, "Active transfer learning for multi-view head-pose classification," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2012, pp. 1168–1171.

[29] P. Sreekanth, U. Kulkarni, S. Shetty, and M. S. M., "Head pose estimation using transfer learning," in *Proc. Int. Conf. Recent Trends Advance Comput. (ICRTAC)*, Sep. 2018, pp. 73–79.

[30] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 4, Jan. 2014, pp. 3320–3328.

[31] R. K. Samala, H.-P. Chan, L. Hadjiiski, M. A. Helvie, C. D. Richter, and K. H. Cha, "Breast cancer diagnosis in digital breast tomosynthesis: Effects of training sample size on multi-stage transfer learning using deep neural nets," *IEEE Trans. Med. Imag.*, vol. 38, no. 3, pp. 686–696, Mar. 2019.

[32] H. Pan, Z. Pang, Y. Wang, Y. Wang, and L. Chen, "A new image recognition and classification method combining transfer learning algorithm and MobileNet model for welding defects," *IEEE Access*, vol. 8, pp. 119951–119960, 2020.

[33] Z. Wang, L. Du, J. Mao, B. Liu, and D. Yang, "SAR target detection based on SSD with data augmentation and transfer learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 150–154, Jan. 2019.

[34] G. Huo, Z. Wu, and J. Li, "Underwater object classification in sidescan sonar images using deep transfer learning and semisynthetic training data," *IEEE Access*, vol. 8, pp. 47407–47418, 2020.

[35] D. Malmgren-Hansen, A. Kusk, J. Dall, A. A. Nielsen, R. Engholm, and H. Skriver, "Improving SAR automatic target recognition models with transfer learning from simulated data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1484–1488, Sep. 2017.

[36] Y.-X. Wang, D. Ramanan, and M. Hebert, "Growing a brain: Fine-tuning by increasing model capacity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3029–3038.

[37] T. Shermin, S. Wei Teng, M. Murshed, G. Lu, F. Sohel, and M. Paul, "Enhanced transfer learning with ImageNet trained classification layer," 2019, *arXiv:1903.10150*. [Online]. Available: http://arxiv.org/abs/1903.10150

[38] B. Zhao, B. Huang, and Y. Zhong, "Transfer learning with fully pretrained deep convolution networks for land-use classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1436–1440, Sep. 2017.

[39] T. Shermin, M. Murshed, G. Lu, and S. Wei Teng, "Transfer learning using classification layer features of CNN," 2018, *arXiv:1811.07459*. [Online]. Available: http://arxiv.org/abs/1811.07459

[40] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian, "Coarse-to-fine description for fine-grained visual categorization," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4858–4872, Oct. 2016.
[41] L. Zheng, Y. Zhao, S. Wang, J. Wang, and Q. Tian, "Good practice in CNN feature transfer," 2016, *arXiv:1604.00133*. [Online]. Available: http://arxiv.org/abs/1604.00133
[42] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, Nov. 1998.
[43] 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.
[44] S. Jialin Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[46] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2656–2666.
[47] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The CAS-PEAL large-scale chinese face database and baseline evaluations," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 38, no. 1, pp. 149–161, Jan. 2008.

SENQIU CHEN received the bachelor's degree in measurement and control technology and instrumentation from the Nanjing University of Aeronautics and Astronautics, China, in 2019, where he is currently pursuing the M.S. degree. His research interests include image processing and deep learning.

WENBO LIU received the B.S. and M.S. degrees in electronic engineering from the Harbin Shipbuilding Engineering Institute, Harbin, China, in 1990 and 1993, respectively, and the Ph.D. degree from the Department of Measurement and Testing, Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2001.

Since 1993, she has been working with the College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, where she was promoted to a Full Professor, in 2006. Her research interests include control and measurement technology, signal processing and information processing methods, and nonlinear system analysis and design.

GONG ZHANG (Member, IEEE) was born in 1964. He received the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, in 2002. He is currently a Professor with the Nanjing University of Aeronautics and Astronautics. His research interests include SAR image processing, target detection, and target recognition.