Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048304

ARTICLE

# Expression Recognition Method Based on Convolutional Neural Network and Capsule Neural Network

**Zhanfeng Wang**[1] **and Lisha Yao**[2,*]

[1]School of Computer Science and Artificial Intelligence, Chaohu University, Hefei, 238000, China

[2]School of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei, 230088, China

*Corresponding Author: Lisha Yao. Email: jsjyaolisha@163.com

Received: 04 December 2023 Accepted: 12 March 2024 Published: 25 April 2024

## ABSTRACT

Convolutional neural networks struggle to accurately handle changes in angles and twists in the direction of images, which affects their ability to recognize patterns based on internal feature levels. In contrast, CapsNet overcomes these limitations by vectorizing information through increased directionality and magnitude, ensuring that spatial information is not overlooked. Therefore, this study proposes a novel expression recognition technique called CAPSULE-VGG, which combines the strengths of CapsNet and convolutional neural networks. By refining and integrating features extracted by a convolutional neural network before introducing them into CapsNet, our model enhances facial recognition capabilities. Compared to traditional neural network models, our approach offers faster training pace, improved convergence speed, and higher accuracy rates approaching stability. Experimental results demonstrate that our method achieves recognition rates of 74.14% for the FER2013 expression dataset and 99.85% for the CK+ expression dataset. By contrasting these findings with those obtained using conventional expression recognition techniques and incorporating CapsNet's advantages, we effectively address issues associated with convolutional neural networks while increasing expression identification accuracy.

## KEYWORDS

Expression recognition; capsule neural network; convolutional neural network

## 1 Introduction

Given the rapid pace of scientific and technological advancements, coupled with ongoing progress in computer science, artificial intelligence, and related fields, there is a constant need for improving human-computer interaction. In face-to-face communication, nonverbal cues such as facial expressions and body movements play a crucial role in conveying messages and helping the audience understand the speaker's intentions [1]. Facial expressions serve as manifestations of human thoughts, emotions, and conditions. As science and technology continue to advance with dedicated research efforts, the potential ability of computers to accurately and efficiently recognize facial expressions holds great promise for enhancing the naturalness and harmony of human interactions. The theoretical and practical applications of facial expression recognition technologies are numerous.

According to Mehrabian's investigation [2], verbal communication accounts for only 7% of expressing emotions, while other non-verbal means such as rhythm, voice, speed of speech, and particularly facial expressions contribute to 38%. Among these non-verbal cues, facial expressions alone represent 55% [3]. Therefore, valuable insights into human thoughts and emotions can be derived from analyzing facial expressions. The primary objective of facial expression recognition technology is to develop an effective and efficient system capable of accurately identifying various human emotions conveyed through expressions including neutrality, surprise, disgust, anger, sadness, and happiness [4].

The rapid advancement of deep learning technology [5–9], artificial intelligence technology [10–13], and computer hardware has significantly impacted facial expression recognition technology. Facial recognition [14] is a representative interdisciplinary, involving psychology, sociology, psychology and psychology. The continuous development of facial expression recognition technology will surely attract more scholars' attention, and research in various fields will promote the development of many fields.

Facial expression recognition algorithms can be categorized into machine learning algorithms and deep learning-based methods. Traditional recognition algorithms typically employ specific feature extraction techniques based on the research subject, and these traditional methods have long been the focus of human facial expression recognition research. However, traditional facial expression recognition methods rely on manually designed feature extraction, which is susceptible to interference from irrelevant factors, resulting in relatively low-level semantic features being extracted for most facial expressions. With advancements in computer hardware performance, researchers have increasingly turned to deep learning-based approaches as the mainstream method for facial expression recognition. Deep learning-based algorithms address the limitations of traditional methods by eliminating the need for manual design of facial expression feature extraction [15,16]. By increasing network depth and width, higher-level semantic features can be extracted. Currently, convolutional neural networks (CNN) are widely used in facial expression recognition tasks; however, CNNs require a large amount of training data and may struggle with identifying small sample sizes while also being prone to overfitting or underfitting issues. Additionally, CNNs exhibit poor performance when recognizing complex scenes and fail to accurately handle changes in image angles or reflect relationships between internal feature levels [17–19].

The research motivation of this paper is to address the limited generalization ability of convolutional neural network expression recognition methods when processing small sample data, such as expression data, and to overcome the challenges posed by complex scenes, objects at different scales, and image transformations. Our objective is to optimize the model in order to enhance its recognition capability for small samples, improve its performance in complex scenes, enable better identification of facial images with varying scales, effectively handle rotation, scaling and translation transformations of facial images, and enhance the model's robustness for recognizing different types of transformations.

The present paper introduces the capsule neural network (CapsNet) and proposes an expression recognition method based on both convolutional neural network and capsule neural network. Firstly, by leveraging the concept of capsules, the data dimension is reduced to enhance the model's processing capability for small sample data and facilitate better identification of objects with varying scales. Secondly, the capsule neural network effectively handles the relationship between local features and global features, thereby improving recognition performance in complex scenes. Lastly, spatial orientation information is stored and propagated using the capsule structure within the capsule neural network.

## 2 Related Work

Since the 19th century, foreign scholars have conducted systematic analyses and research on facial expressions. In 1872, biologist Darwin confirmed in his book "The Emotional Expression of Man and Animals" that facial expressions are shared characteristics between humans and animals. In 1971, psychologists Ekman and Friesen developed a comprehensive classification system for facial expressions, encompassing emotions such as happiness, fear, sadness, surprise, anger, and disgust [20].

Deep learning-based techniques and machine learning algorithms are two categories under which facial expression recognition systems fall. The deep learning approach, surpassing machine learning algorithms and yielding more abstract information, has gained popularity in recent years. In 2016, Li et al. proposed a facial expression recognition method based on Gabor and Conditional Random Fields (Gabor+PCA+CRF [21]). This method utilized the Gabor characteristics extracted by five scales to reduce dimensionality and employed State Random Fields (CRF) for facial expression identification and classification. Similarly, in the same year, AT Lopes presented the Lopes algorithm [22] in 2017, which jointly processed images using convolutional neural networks and specific preprocessing methods. Yan et al.'s cooperative discriminant multi-metric learning algorithm (CDMML) [23] was suggested for video-based facial expression identification by integrating voice and video modalities to enhance recognition performance. In 2021, Pourmirzaei et al. [24], employing self-supervised learning techniques, improved recognition performance and reduced error rates through fine-grained tasks of facial feature expression analysis. Aouayeb et al. [25], on the other hand, proposed a Transformer model for facial expression recognition based on attention mechanisms. The present study introduces a novel joint extrusion and excitation (SE) block for learning visual Transformer in order to address the limited training data issue of the transformer model in FER task. Experimental results demonstrate the competitive performance of this approach. Minaee et al. proposed a deep learning technique based on attention convolution networks (Deep Emotion [26]), which effectively focuses on facial key regions and significantly enhances the performance compared to previous models across multiple datasets. The current status of image-based facial expression recognition was assessed based on the suggestion made by Christopher Pramerdorfer (Inception [27]). To highlight algorithm differences and their performance effects, CNNS were employed. Fard et al. proposed an algorithm model called adaptive correlation (AD-CORRE [28]) loss to guide the correlation between samples in the classroom and network level samples, aiming to embed feature vectors with reduced correlation. Khaireddin and his colleagues proposed the adoption of VGGNet [29] architecture, which was fine-tuned with precise adjustments to its approximate parameters and implemented using various optimized experimental methodologies, leading to favorable outcomes.

The aforementioned research demonstrates that although deep learning has made some progress in face recognition, there are still challenges with convolutional neural networks (CNNs), such as their limited ability to accurately handle changes in angles and directions of images, as well as the relationship between internal feature levels. In 2017, Sabour et al. [30] proposed a novel deep learning network structure called Capsule Neural Network (CapsNet), which initially outperformed CNNs and RNNs in image recognition. Despite being at an early stage of research, several scholars have applied CapsNet in various fields including natural language processing and computer vision, yielding competitive results. For instance, Kim et al. [31] introduced an abnormal driving centerline intersection detection method based on CapsNet. One advantage of using CapsNet is that it recognizes objects as vectors containing all state information within capsules. Suri et al. [32], utilizing signals collected by a specially designed wearable IMU system, proposed a novel one-dimensional deep CapsNet architecture for continuous Indian sign language recognition. Compared to CNNs, CapsNet exhibits higher performance demonstrating its applicability. Li et al. [33] employed CapsNet to determine whether natural text contains secret information and achieved robustness and accuracy. By enhancing the direction and size of information, CapsNet achieves vectorization without losing spatial information from images thereby overcoming limitations of traditional CNNs. To address existing issues with CNNs and enhance expression recognition performance, this paper proposes a model combining CNNs with CapsNet for expression recognition purposes due to the latter's superior ability in recognizing spatial orientation.

## 3 Method

The VGG16 model serves as the fundamental framework in this study, while the incorporation of a capsule layer significantly enhances facial expression recognition performance. Referred to as CAPSULE-VGG, our proposed convolutional neural network and capsule neural network model represents a fusion CapsNet classification approach. Specifically, we introduce a capsule neural network before the fully connected layer within the existing VGG16 architecture.

### *3.1 Convolutional Neural Network*

The convolutional layer, pooling layer, activation layer, fully connected layer, and output layer collectively constitute a convolutional network [34–36], which shares similarities with various fundamental neural networks while maintaining a relatively straightforward architectural design. The convolutional operations in a general convolutional neural network consistently consider the construction of local mean or maximum feature pooling levels [37–40]. Utilizing a multi-feature extraction network can significantly reduce the resolution of multi-feature images. The output results from the fully connected layer are fed into the softmax layer, enabling node-based multi-feature extraction and comparison of feature graphs [41,42].

The convolution layer performs comprehensive feature extraction by employing multiple convolution kernels, as demonstrated in Eq. (1) for precise calculation.

$$y_j = \alpha \left( \sum_{i=1}^{N_j} \omega_i * x_i + b_j \right) \tag{1}$$

The $i$-th feature map of the upper layer, denoted as $x_i$, serves as the input at this stage. Correspondingly, $\omega_i$ represents the convolution kernel associated with it. $y_j$ refers to the $j$-th feature map, $b_j$ denotes the bias term, and $N_j$ signifies the number of features in each feature map. The activation function $f$ can take various forms such as Tanh or ReLu.

The primary objective of the pooling layer is to reduce the dimensions (width, length, and number of channels) of the preceding layer while minimizing computation, memory, and parameter usage. This facilitates achieving specific scale and space invariance objectives without relying on overly aggressive fitting techniques. For instance, after obtaining the output from the last convolutional layer, eigenvalues representing characteristic quantities of $\mathrm{n} * \mathrm{n} * \mathrm{M}$ can be utilized for $2 * 2$ pooling operations within a window. Consequently, the pooled output would have dimensions of $\mathrm{n}/2 * \mathrm{n}/2 * \mathrm{M}$. Various common calculation methods such as random pooling, maximum pooling, and average pooling are commonly employed.

### *3.2 CapsNet*

The CapsNet model was developed based on the principle of human visual object recognition [43–46]. Hinton posits that during object recognition, humans tend to disregard irrelevant details and transmit information about key parts of the observed object from the visual central nervous system to the higher central nervous system for higher-level decision-making. Similarly, in our representation of image features as capsules, upper-level capsules do not fully incorporate all information from each lower-level capsule but selectively integrate salient information.

The CapsNet architecture, illustrated in Fig. 1, represents a high-performance network structure for deep learning. In this configuration, each feature is represented by a vector known as a capsule [47–49]. By training this architecture, effective image feature extraction can be achieved. The characteristics of each spatial entity are encapsulated within vectors and gradually combined using clustering methods. To preserve the spatial information of the data, CapsNet replaces the pooling layer commonly found in CNNs, resulting in an impressively low test error rate of only 0.25% on the MNIST dataset.

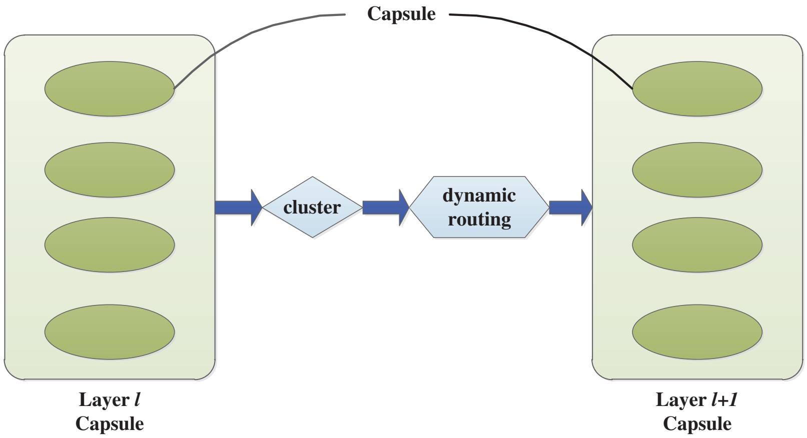


**Figure 1:** Capsule neural network structure diagram

The feature vector of a capsule possesses the property that its direction represents a specific feature, while its modulus length (L2 norm) indicates the likelihood of this feature's existence. Furthermore, it is important to note that the modulus length of all capsules always remains less than 1. In contrast to conventional neural networks, the utilization of a dynamic routing algorithm facilitates effective information transfer between upper and lower layers within CapsNet.

The instantiation parameters of a specific entity type are represented by the input and output vectors of the capsule. The direction of these vectors indicates certain attributes of the entity, while their length signifies the likelihood of its existence. A transformation matrix predicts the instantiation parameters for capsules within the same layer. The dynamic routing algorithm plays a crucial role in prediction, ensuring consistency among multiple predictions within this layer and activating capsules in the subsequent layer. The output vector length of each "capsule" reflects its occurrence probability in the current input, necessitating that it falls within a range from 0 to 1. Nonlinear compression through "squashing" guarantees that short vectors approach zero length, while long vectors are reduced to unit length. At the capsule level, activation function is achieved using discriminant learning approach as expressed by Eq. (2).

$$v_j = \frac{\left\|s_j^2\right\|}{1 + \left\|s_j^2\right\|} \frac{s_j}{\left\|s_j\right\|} \tag{2}$$

The input vector of capsule $j$, denoted as $s_j$, and its corresponding output vector $v_j$ are defined in this case. Moreover, the sum of the vector weights from all capsules $i$ in the preceding layer to capsule

$j$ is equal to $s_j$. This non-linear compression not only preserves the original direction of the vector, thereby maintaining the attributes of instantiated parameters unchanged but also restricts the length of the vector within a range between 0 and 1. The calculation process for obtaining the input vector involves two steps: Linear combination and Routing, as expressed by Eq. (3).

$$s_j = \sum_i c_{ij}\hat{u}_{j|i}, \hat{u}_{j|i} = w_{ij}u_i \tag{3}$$

Among them, $\hat{u}_{j|i}$ is a linear combination of $u_i$, which, although not identical, exhibits comparability to the fully connected network. It demonstrates the strength of the connection between the $i$-th capsule in the preceding layer and the $j$-th capsule in the subsequent layer or represents the prediction vector obtained by multiplying the output vector of the $i$-th capsule in the preceding layer with its corresponding weight matrix. The process of calculating $s_j$ is designed to iteratively update $c_{ij}$ using a dynamic routing algorithm. Through routing, this procedure can acquire the input vector $s_j$ of the subsequent layer of capsules. Dynamic routing is a pivotal component in capsule neural networks, as its dynamic nature enhances the network's explanatory power and improves performance on noisy samples. By iteratively adjusting the coupling coefficient between parent and child capsules, dynamic routing facilitates dynamic connections between them. This mechanism identifies which sub-capsules should be updated during backward propagation and highlights entities in the image that require attention, thereby encouraging sub-capsules to transmit information to their parent capsules with higher coupling coefficients.

Fig. 2 illustrates the process of implementing dynamic routing, which involves obtaining the prediction vector $\hat{u}_{j|i}^{(l+1)} \in R^{d^{(l+1)}}$ by multiplying sub-capsule $u_i^{(l)}$ with learnable viewpoint matrix $W_{i,j}^{(l+1)} \in R^{d^{(l+1)} \times d^{(l)}}$. The coupling coefficient between capsules is determined using Softmax function as the cumulative prior and set to 0 in the first iteration. Then, squashing function is used for weighted sum of $\hat{u}_{j|i}^{(l+1)}$, coupling coefficient, parent capsule $u_j^{(l+1)}$, and corresponding prediction vector $\hat{u}_{j|i}^{(l+1)}$. Finally, scalar product is accumulated with $t_{i,j}^{(l+1)}$ to update coupling coefficient. This mechanism enables dynamic routing to quickly establish connection between child capsule and parent capsule.

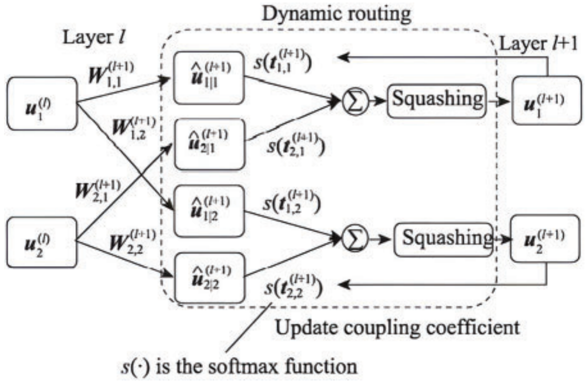


**Figure 2:** Execution process of dynamic routing

Table 1 displays the steps of the dynamic routing algorithm.

**Table 1:** Dynamic routing algorithm

| 1 | *Rounting* $\left(\hat{u}_{j\|i}, r, l\right)$ |
|---|---|
| 2 | Regarding every $l$+1-layer capsule $j$ and every $l$-layer capsule $i$: $b_{ij} \rightarrow 0$ |
| 3 | Iterate $r$ times |
| 4 | Regarding every $l$-layer capsule, $i$: $c_{ij} \leftarrow softmax(b_{ij})$ |
| 5 | For all capsules $j$ with $l$+1 layers: $s_i = \sum_i c_{ij}\hat{u}_{j\|i}$ |
| 6 | For all capsules $j$ with $l$+1 layers: $v_j \leftarrow squash(s_j)$ |
| 7 | Regarding every $l$+1-layer capsule $j$ and every $l$-layer capsule $i$: |
| | $b_{ij} = b_{ij} + \hat{u}_{j\|i} \cdot v_j$ |
| | Return $v_j$ |

### 3.3 Loss Function and Optimization Algorithm

The internal weight matrix of parameters should be adjusted in accordance with the loss function, while the coupling factor needs to be dynamically modified through routing.

The probability of a capsule's representation content is denoted by its vector magnitude, and the sum of the output probabilities does not necessarily equal 1. Therefore, this paper employs interval loss to formulate the network's loss function, as opposed to the commonly used cross entropy loss in traditional classification tasks. The interval loss function can be mathematically represented as Eq. (4).

$$L_c = T_c \max\left(0, m^+ - \|v_c\|^2\right) + \lambda(1 - T_c)\max(0, \|v_c\| - m^-)^2 \tag{4}$$

In Eq. (4), $c$ represents the category. $T_c$ indicates the presence or absence of a class $c$ intrusion. Within the output layer, $v_c$ denotes the length of the capsule, specifically representing the likelihood that a sample belongs to class $c$. The maximum penalty for false positives is denoted as $m^+$, whereas $m^-$ represents the minimum penalty for false negatives. The coefficient of proportionality $\lambda$ is utilized to adjust their specific gravity accordingly. In this study, we set $\lambda$, $m^+$ and $m^-$ to 0.25, 0.9 and 0.1, respectively. Although dynamic routing algorithm resolves weight update issues between capsules exclusively, it remains crucial to implement backpropagation in order to enhance network convergence capabilities effectively. To achieve smooth convergence and minimize loss values iteratively while updating neuron weights, we employ Adam method as an optimization algorithm for our loss function.

### 3.4 CAPSULE-VGG

The VGG16 network is employed in this study for facial expression recognition. Fig. 3 illustrates the model structure constructed using the VGG16 network, which consists of 13 convolutional layers accompanied by 5 max-pooling and 3 fully connected levels. In terms of the network architecture, a combination approach integrating CapsNet with the VGG16 network is adopted.

Due to the parameter sharing in convolution and the local effect of the pooling layer, CNN exhibits translation invariance, implying that when the position and direction of an entity change during prediction, neurons that were active for the original entity will not be activated. Conversely, the pooling layer results in significant degradation of spatial information. To better preserve the spatial properties of data, we optimize the basic VGG16 network and propose a CAPSULE-VGG model. The features extracted from VGG16 are further analyzed using CapsNet, thereby enhancing the representation of spatial information within these features. The structural diagram is illustrated in Fig. 4.

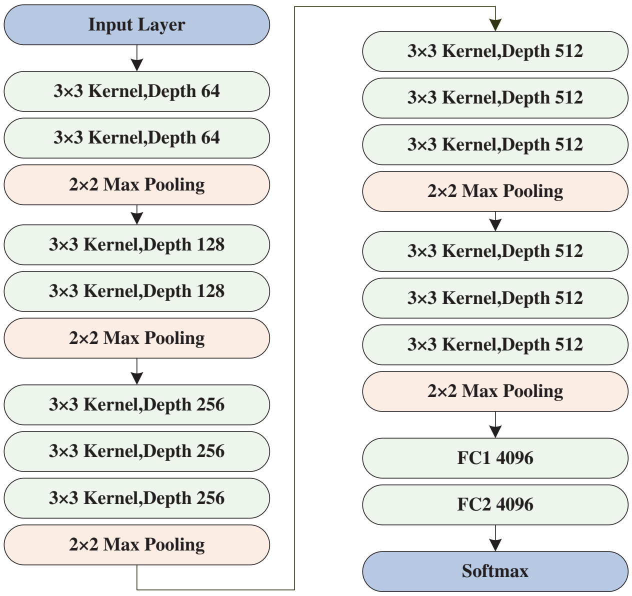


**Figure 3:** VGG-16 structure diagram

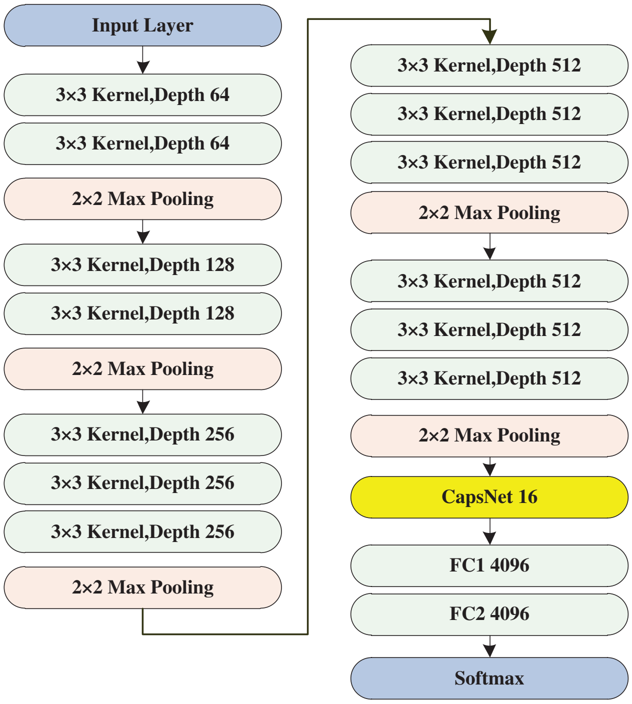


**Figure 4:** CAPSULE-VGG structure diagram

The CAPSULE-VGG model fully mines the preprocessed expression data set and extracts its features using the VGG module. These features are then aggregated by the capsule network layer, with weights between capsules updated through dynamic routing algorithm. The number of capsules is set to 16, and the final classification result is outputted by a Softmax classifier. Table 2 provides detailed configuration information for the CAPSULE-VGG model.

**Table 2:** Detailed configuration of CAPSULE-VGG model

| | |
|---|---|
| Conv1-1 | Convolution kernel: 3 × 3, 64 |
| Conv1-1 | Convolution kernel: 3 × 3, 64 |
| Maxpool | 2 × 2 |
| Conv2-1 | Convolution kernel: 3 × 3, 128 |
| Conv2-2 | Convolution kernel: 3 × 3, 128 |
| Maxpool | 2 × 2 |
| Conv3-1 | Convolution kernel: 3 × 3, 256 |
| Conv3-2 | Convolution kernel: 3 × 3, 256 |
| Conv3-3 | Convolution kernel: 3 × 3, 256 |
| Maxpool | 2 × 2 |
| Conv4-1 | Convolution kernel: 3 × 3, 512 |
| Conv4-2 | Convolution kernel: 3 × 3, 512 |
| Conv4-3 | Convolution kernel: 3 × 3, 512 |
| Maxpool | 2 × 2 |
| Conv5-3 | Convolution kernel: 3 × 3, 512 |
| Conv5-3 | Convolution kernel: 3 × 3, 512 |
| Conv5-3 | Convolution kernel: 3 × 3, 512 |
| Maxpool | 2 × 2 |
| Activation function | ReLu |
| CapsNet | Layer number = 1, Capsule numbers = 16 |
| FC1 | 4096 |
| FC2 | 4096 |
| Activation function | Softmax |

## 4 Experimental Results and Analysis

### *4.1 Data Set*

The performance of the model is evaluated in this study using two facial expression datasets, namely FER2013 [50] and CK+ [51,52]. The CK+ dataset is an extension and supplementation of the original Cohn-Kanade data from 2010. Data collection for CK+ was conducted indoors. To construct a comprehensive network model, we extracted frames from videos and gathered photographs captured by 123 staff members. Each video yielded 593 photos encompassing seven distinct expressions. A representative image is depicted in Fig. 5.

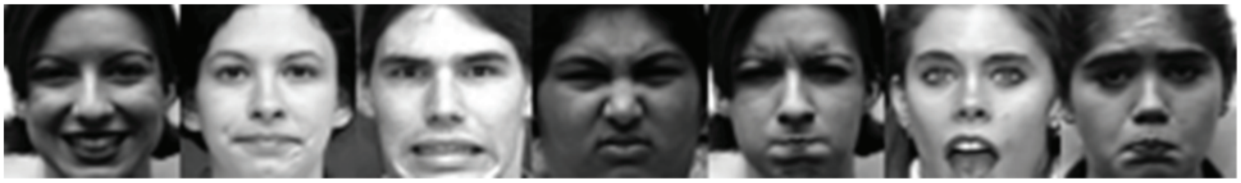


**Figure 5:** CK+ expression data set example

The specific configuration of the original dataset is presented in Table 3.

**Table 3:** CK+ raw data set data configuration

| | Label | Training set | Test set | Total |
|---|---|---|---|---|
| Anger | 0 | 1230 | 120 | 1350 |
| Disgust | 1 | 1590 | 180 | 1770 |
| Happy | 2 | 1860 | 210 | 2070 |
| Fear | 3 | 660 | 90 | 750 |
| Sad | 4 | 750 | 90 | 840 |
| Surprise | 5 | 480 | 60 | 540 |
| Neutral | 6 | 480 | 60 | 540 |
| Total | – | 8820 | 990 | 9810 |

The FER2013 dataset is a comprehensive collection of freely available data, obtained using the Google image retrieval API. After removing any defective frames and adjusting the cropping area, all images were standardized to a resolution of $48 * 48$ pixels. The training set consists of 28,709 images, while the verification and test sets each contain 3589 images. These datasets encompass seven distinct emotional tags: Neutral, happy, sad, surprised, disgusted, and fearful. An illustrative sample image is presented in Fig. 6.

The distribution of different types of data exhibits non-uniformity, and the facial expression dataset is limited in size. This study enhances the dataset to augment the model's capacity for generalization by incorporating techniques such as random horizontal flip, random cropping, and random rotation.

### *4.2 Experimental Parameters*

The parameters utilized in the experimental model are presented in Table 4, encompassing the learning rate, test iterations, learning decay rate, batch size, quantity of capsules within the capsule layer, dynamic routing iterations, as well as training optimizer and loss function parameters.

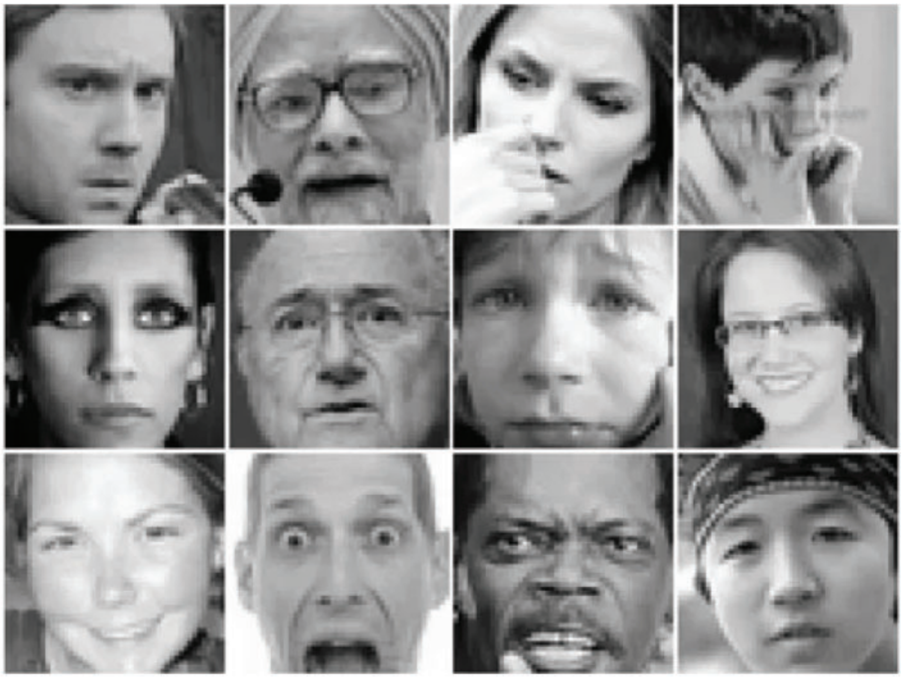


**Figure 6:** FER2013 expression data set example

**Table 4:** Model parameter

| Parameter | CK+ | FER2013 |
|---|---|---|
| Learning rate | 0.00001 | 0.000001 |
| Test iterations | 200 | 80 |
| Weight decay | 0.0005 | 0.0005 |
| Batch_size | 20 | 10 |
| Capsule numbers | 16 | 16 |
| Route_iterations | 3 | 3 |
| Optimizer | Adam | Adam |
| $\lambda$ | 0.25 | 0.25 |
| $m^+$ | 0.9 | 0.9 |
| $m^-$ | 0.1 | 0.1 |

## 4.3 Experimental Analysis

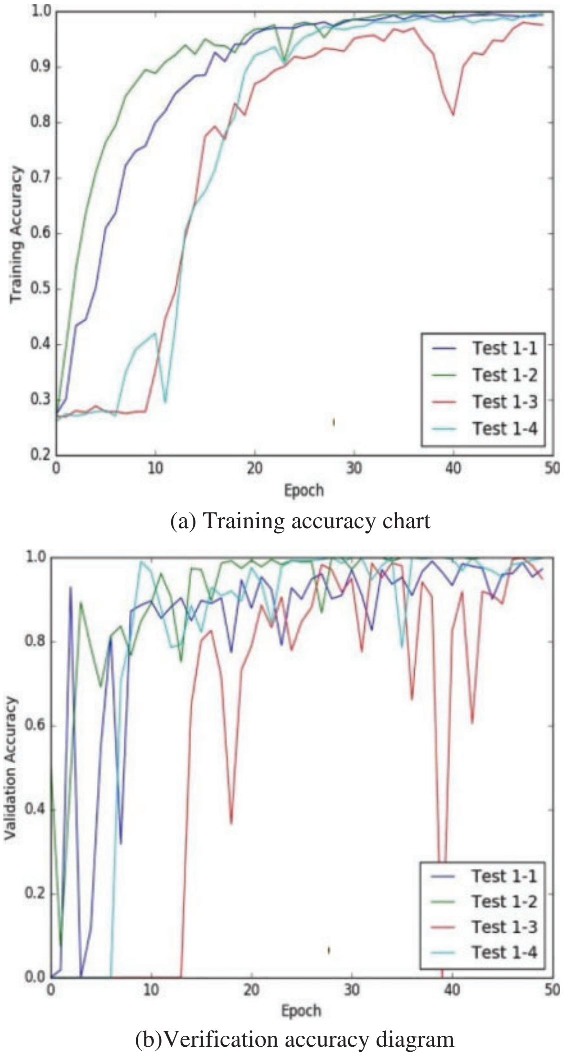
### 4.3.1 Preparatory Experiment

A preliminary experiment was conducted on the CK+ dataset to investigate the impact of CapsNet on the novel network architecture and determine the optimal number of capsule layers for achieving superior performance. Notably, the number of test iterations was fixed at 50, while other parameters were set as specified in Table 4. The experimental parameters are presented in Table 5.

**Table 5:** Preliminary experimental parameters

| Experimental name | Number of capsule layers |
|---|---|
| Test 1–1 | 0 |
| Test 1–2 | 1 |
| Test 1–3 | 2 |
| Test 1–4 | 3 |

Fig. 7 presents a comparative diagram of the training and validation accuracies for various capsule layers in the preliminary experiment.



(a) Training accuracy chart

(b)Verification accuracy diagram

**Figure 7:** FER2013 expression data set example

The experiment compares the training accuracy, validation accuracy, test accuracy, and duration required for each epoch. The findings are presented in Table 6.

**Table 6:** Prepare experimental results

| | Train Acc (%) | Val ACC (%) | Test Acc (%) | Epoch Time (s) |
|---|---|---|---|---|
| Test 1–1 | 99.33 | 97.29 | 85.48 | 27 |
| Test 1–2 | 99.87 | 100.00 | 88.22 | 28 |
| Test 1–3 | 97.57 | 94.88 | 87.13 | 29 |
| Test 1–4 | 99.44 | 99.80 | 87.05 | 30 |

The mesh structure with a single capsule layer and without a capsule layer exhibits the highest motion accuracy at 30∼40 epochs, along with enhanced stability, surpassing the performance of mesh structures with two or three capsule layers (Fig. 7a). Additionally, Fig. 7b demonstrates that after 30 epochs, the network featuring a single capsule level achieves superior verification accuracy initially and tends to stabilize over time. Although Table 6 indicates some influence of the quantity of capsule layers on experimental accuracy, an increase in the number of capsule layers does not correspondingly improve classification accuracy during practical usage. Consequently, this model is designed to have only one specified capsule layer.

### *4.3.2 Ablation Experiment*

Firstly, this paper evaluates the rationality of the CAPSULE-VGG model. The experimental results are presented in Table 7. Among them, VGG represents the baseline model without incorporating the capsule neural network, while CAPSULE-VGG is a composite model that combines convolutional neural networks with capsule neural networks. The findings demonstrate that by introducing the capsule network, training convergence is accelerated, model accuracy becomes more stable, and recognition accuracy improves by 2.24% and 0.86% on CK+ and FER2013 datasets, respectively. These results indicate that the capsule network compensates for limitations inherent to convolutional neural networks, thereby facilitating enhanced learning of facial expression features and improved overall network performance.

**Table 7:** Experimental comparison of the model on CK+ and FER2013

| Model | Accuracy/% | |
|---|---|---|
| | CK+ | FER2013 |
| VGG | 97.61 | 73.28 |
| CAPSULE-VGG | 99.85 | 74.14 |

Through the analysis of CK+ expression and FER2013 emotion, we have obtained prediction results for different types of facial expressions using the CAPSULE-VGG network model. The emotions of each category were predicted and corresponding results were provided. Tables 8 and 9 present the different types of mixing matrices for CK+ and FER2013 datasets in the CAPSULE-VGG model.

**Table 8:** CK+ confusion matrix %

| Expression | Anger | Disgust | Fear | Happy | Neutral | Sad | Surprise |
|---|---|---|---|---|---|---|---|
| Anger | 99.88 | 0.07 | 0 | 0 | 0 | 0.05 | 0 |
| Disgust | 0 | 100 | 0 | 0 | 0 | 0 | 0 |
| Fear | 0.03 | 0 | 99.89 | 0 | 0.03 | 0.05 | 0 |
| Happy | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| Neutral | 0.05 | 0 | 0 | 0 | 99.67 | 0.17 | 0.11 |
| Sad | 0.04 | 0.04 | 0.16 | 0 | 0.25 | 99.51 | 0 |
| Surprise | 0 | 0 | 0 | 0 | 0 | 0 | 100 |

**Table 9:** FER2013 confusion matrix %

| Expression | Anger | Disgust | Fear | Happy | Neutral | Sad | Surprise |
|---|---|---|---|---|---|---|---|
| Anger | 73.11 | 0.99 | 6.11 | 3.26 | 6.25 | 8.65 | 1.63 |
| Disgust | 9.68 | 71.56 | 4.13 | 1.56 | 3.69 | 7.56 | 1.82 |
| Fear | 7.56 | 0.09 | 66.72 | 0.16 | 8.23 | 13.68 | 3.56 |
| Happy | 3.23 | 0.89 | 3.12 | 83.02 | 5.89 | 2.98 | 0.87 |
| Neutral | 5.97 | 0.86 | 4.68 | 4.77 | 73.69 | 9.14 | 0.89 |
| Sad | 7.56 | 1.77 | 7.77 | 2.13 | 11.23 | 68.69 | 0.85 |
| Surprise | 1.89 | 0.08 | 7.98 | 2.75 | 2.98 | 2.15 | 82.17 |

Comparing the confusion matrix of the CK+ dataset with that of FER2013, it is evident that the image quality in CK+ surpasses that of FER2013, leading to a significantly higher recognition accuracy for facial expressions. In comparison to the CK+ dataset, FER2013 presents more challenging characteristics. The images in FER2013 exhibit a wider range of pose angles and cover a broader age spectrum, thereby resembling real-world scenarios more closely. The lower recognition rate achieved by CAPSULE-VGG on the FER2013 dataset (74.14%) can be attributed largely to the substantial variations in facial postures within this dataset.

### *4.3.3 Contrast Experiment*

To further validate the proposed model, we compared the experimental results of the CAPSULE-VGG model on the CK+ expression dataset with the following experimental approaches.

Model 1: Collaborative Discriminant Multi-Metric Learning (CDMML). Yan et al. [23] introduced CDMML, an algorithm that combines speech and video modalities to improve facial expression recognition.

Model 2: Nonlinear eval on SL + SSL puzzling. Pourmirzaei et al. [24] utilized self-supervised learning for fine-grained tasks in facial feature expression recognition, resulting in improved performance and reduced error rate.

Model 3: ViT + SE. Aouayeb et al. [25] proposed a Transformer model based on the attention mechanism for recognizing facial expressions. To address the lack of training data issue in Transformer models, they incorporated extrusion and excitation (SE) blocks into vision Transformers. Experimental results demonstrate the competitiveness of this approach.

This paper presents a comparative analysis of four models applied to the CK+ expression dataset: CDMML, Nonlinear evaluation on SL + SSL perplexity, and ViT + SE models.

The comparison experiment of the CK+ dataset's accuracy is illustrated in Fig. 8, which demonstrates that this paper achieves the best recognition performance with an accuracy of 99.85%. Subsequently, we conducted experiments to compare the CAPSULE-VGG model's results on the FER2013 expression dataset.

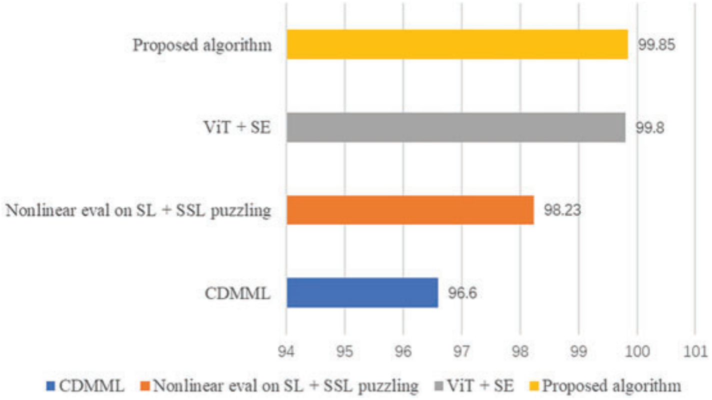


**Figure 8:** The accuracy of CK+ data set comparison experiment (unit: %)

Model 1: DeepEmotion, proposed by Shervin Minaee, presents a deep learning approach based on attention convolutional networks (DeepEmotion) [26]. This model effectively focuses attention on the primary facial regions and significantly enhances the performance compared to previous models across multiple datasets.

Model 2: Inception, introduced by Christopher Pramerdorfer in a comprehensive review of the state-of-the-art techniques, demonstrates image-based facial expression recognition capabilities (Inception [27]). The utilization of CNNs highlights algorithmic distinctions and their impact on performance outcomes.

Model 3: Ad-Corre is an algorithmic framework developed by Fard et al., which incorporates adaptive correlation loss (AD-CORRE [28]) to guide feature vector embedding with reduced correlation between level samples and classroom samples.

Model 4: VGGNet, proposed by Khaireddin et al., adopts the VGGNet architecture [29] and achieves commendable results through precise parameter tuning and various optimization tests.

During experimentation, CAPSULE-VGG was compared against DeepEmotion model, Inception model, Ad-Corre model, and VGGNet model using FER2013 expression dataset.

The accuracy of the FER2013 dataset comparison experiment is illustrated in Fig. 9. Upon comparison, it has been observed that the suggested algorithm exhibits lower accuracy compared to alternative approaches. Although there has been an improvement in the model's recognition performance on the FER2013 dataset, further enhancements are still required due to its relatively weaker generalization ability when compared to other datasets.

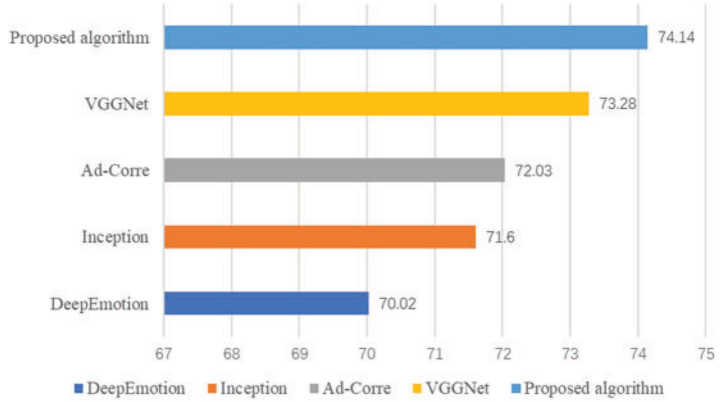


**Figure 9:** The accuracy of FER2013 data set comparison experiment (unit: %)

## 5 Conclusion

The paper introduces the CAPSULE-VGG neural network model of the capsule neural network to address the limitation of convolutional neural networks in considering information such as relative position and angle between image features, resulting in insufficient generalization ability when dealing with complex small sample data like expression data. A novel expression recognition method based on a combination of convolutional neural networks and capsule neural networks is proposed. VGG16 is utilized as the feature extraction component during training, while an additional capsule neural network layer is incorporated to enhance attention towards direction and location features of facial expressions, thereby improving interpretability and network stability. Experimental results on both CK+ expression dataset and FER2013 expression dataset demonstrate that the Capsule-VGG model achieves accuracy rates of 99.85% and 74.14%, respectively, outperforming basic VGG16 by enhancing recognition accuracy by 2.24% and 0.89%. Furthermore, it exhibits faster training convergence, improved model speed, and enhanced stability in terms of accuracy.

**Acknowledgement:** None.

**Funding Statement:** This paper was supported by the following funds: The Key Scientific Research Project of Anhui Provincial Research Preparation Plan in 2023 (Nos. 2023AH051806, 2023AH052097, 2023AH052103); Anhui Province Quality Engineering Project (Nos. 2022sx099, 2022cxtd097); University-Level Teaching and Research Key Projects (Nos. ch21jxyj01, XLZ-202208, XLZ-202106); Special Support Plan for Innovation and Entrepreneurship Leaders in Anhui Province.

**Author Contributions:** Study conception and design: Z. Wang, L. Yao; data collection: Z. Wang; analysis and interpretation of results: Z. Wang, L. Yao; draft manuscript preparation: Z. Wang, L. Yao. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** The accompanying author can provide some of the models or code created or utilized during the study upon request. The datasets generated and/or analysed during the current study are available in the GitHub repository https://GitHub.com/kaiwang960112/Challenge-condition-FER-dataset and http://www.jeffcohn.net/Resources/ with corresponding permission.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] H. B. Liao and B. Xu, "Robust face expression recognition based on gender and age factor analysis," *J. Comput. Res. Dev.* , vol. 58, no. 3, pp. 528–538, 2021.

[2] T. T. Li, Y. L. Hu, and F. L. Wei, "Improved facial expression recognition algorithm based on GAN and application," *J. Jilin Univ. (Sci. Ed.)*, vol. 58, no. 3, pp. 605–610, 2020.

[3] L. Yan, "Application of face expression recognition technology in skilled unsupervised course based on ultra-wide regression network," *J. Intell. Fuzzy Syst.*, vol. 38, no. 11, pp. 1–11, 2020. doi: 10.3233/JIFS-179794.

[4] J. Y. Chen, C. Guo, R. Y. Xu, K. Zhang, Z. K. Yang and H. H. Liu, "Toward Children's empathy ability analysis: Joint facial expression recognition and intensity estimation using label distribution learning," *IEEE Trans. Ind. Inf.*, vol. 18, no. 1, pp. 16–25, 2022. doi: 10.1109/TII.2021.3075989.

[5] J. B. Liu, Y. Bao, W. T. Zheng, and S. Hayat, "Network coherence analysis on a family of nested weighted n-polygon networks," *Fractals*, vol. 29, no. 8, pp. 2150260, 2021. doi: 10.1142/S0218348X21502601.

[6] F. Golpelichi and H. Parastar, "Quantitative mass spectrometry imaging using multivariate curve resolution and deep learning: A case study," *J. Am. Soc. Mass Spectrom.*, vol. 34, no. 2, pp. 236–244, 2023. doi: 10.1021/jasms.2c00268.

[7] Q. Chai, Y. Deng, H. Li, Y. Yu, and S. Ming, "Survey on human action recognition based on deep learning," *Comput. Sci.*, vol. 47, no. 4, pp. 85–93, 2020.

[8] M. Kalfaoglu, S. Kalkan, and A. Alantan, "Late temporal modeling in 3D CNN architectures with BERT for action recognition," in *Proc. of ECCV*, Glasgow, UK, 2020.

[9] S. Yan *et al.*, "Multiview transformers for video recognition," in *Proc. of CVPR*, New Orleans, USA, 2022, pp. 271–283.

[10] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," arXiv preprint arXiv:2203.12602, 2022.

[11] D. Zhao, J. Zhang, C. Guo, D. Zhao, and M. A. S. Hakimi, "Review of video action recognition method based on the depth of learning," *Telecommun. Sci.*, vol. 5, no. 12, pp. 99–111, 2019.

[12] Q. Liang, Y. Li, B. Chen, and K. Yang, "Violence behavior recognition of two-cascade temporal shift module with attention mechanism," *J. Electron. Imag.*, vol. 30, no. 4, pp. 43009, 2021. doi: 10.1117/1.JEI.30.4.043009.

[13] J. B. Liu, S. Nadeem, K. Muhammad, S. Ashraf, and R. Khan, "Single-valued neutrosophic eutrosophic set with quaternion information: A promising approch to assess image quality," *Fractals*, vol. 31, no. 6, pp. 2340074, 2023. doi: 10.1142/S0218348X23400741.

[14] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: Past, present and future," *Comput. Vis., Image. Underst.*, vol. 31, no. 4, pp. 1–24, 2015. doi: 10.1016/j.cviu.2015.03.015.

[15] Y. I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 2, pp. 97–115, 2001. doi: 10.1109/34.908962.

[16] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affective Comput.*, vol. 13, no. 3, pp. 1195–1215, 2020. doi: 10.1109/TAFFC.2020.2981446.

[17] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *Int. J. Comput. Vis.*, vol. 126, no. 5, pp. 550–569, 2018. doi: 10.1007/s11263-017-1055-1.

[18] M. I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019. doi: 10.1109/ACCESS.2019.2917266.

[19] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li and A. M. Dobaie, "Facial expression recognition via learning deep sparse autoencoders," *Neurocomputing*, vol. 273, no. 1, pp. 643–649, 2018. doi: 10.1016/j.neucom.2017.08.043.

[20] Q. M. Liu and Y. Y. Xin, "Face expression recognition based on end-to-end low-quality face images," *J. Chin. Comput. Syst.*, vol. 41, no. 3, pp. 668–672, 2020.

[21] J. Li and B. Zhang, "Facial expression recognition based on Gabor and conditional random fields," in *Proc. 2016 IEEE 13th Int. Conf. Signal Process. (ICSP)*, Chengdu, China, 2016, pp. 752–756.

[22] A. T. Lopes, E. D. Aguiar, A. F. D. Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, no. 12, pp. 610–628, 2017. doi: 10.1016/j.patcog.2016.07.026.

[23] H. Yan, "Collaborative discriminative multi-metric learning for facial expression recognitionin video," *Pattern Recognit.*, vol. 75, pp. 33–40, 2018.

[24] M. Pourmirzaei, G. A. Montazer, and F. Esmaili, "Using self-supervised auxiliary tasks to improve fine-grained facial representation," arXiv preprint arXiv:2105.06421, 2021.

[25] M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, and R. Seguier, "Learning vision transformer with squeeze and excitation for facial expression recognition," arXiv preprint arXiv:2107.03107, 2021.

[26] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-Emotion: Facial expression recognition using attentional convolutional network," *Sens.*, vol. 21, no. 9, pp. 3046, 2021. doi: 10.3390/s21093046.

[27] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: State of the art," arXiv preprint arXiv:1612.02903, 2016.

[28] A. P. Fard and M. H. Mahoor, "Ad-Corre: Adaptive correlation-based loss for facial expression recognition in the wild," *IEEE Access*, vol. 10, pp. 26756–26768, 2022. doi: 10.1109/ACCESS.2022.3156598.

[29] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on FER2013," 2021. [Online]. Available: https://arxiv.org/abs/2105.03588

[30] S. Sabour, N. Frosst, and E. Hinton, "Dynamic routing between capsules," arXiv preprint arXiv:1710.09829, 2017.

[31] M. Kim and S. Chi, "Detection of centerline crossing in abnormal driving using CapsNet," *J. Supercomput.*, vol. 75, no. 1, pp. 189–196, 2019. doi: 10.1007/s11227-018-2459-6.

[32] K. Suri and R. Gupta, "Continuous sign language recognition from wearable IMUs using deep capsule networks and game theory," *Comput. Electr. Eng.*, vol. 78, pp. 493–503, 2019. doi: 10.1016/j.compeleceng.2019.08.006.

[33] H. Li and S. Jin, "Text steganalysis based on capsule network with dynamic routing," *IETE Tech. Rev.*, vol. 38, no. 1, pp. 72–81, 2021. doi: 10.1080/02564602.2020.1780959.

[34] I. Mecheter, M. Abbod, H. Zaidi, and A. Amira, "Brain MR images segmentation using 3D CNN with features recalibration mechanism for segmented CT generation," *Neurocomputing*, vol. 49, no. 2, pp. 232–243, 2022. doi: 10.1016/j.neucom.2022.03.039.

[35] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (CNN) in vegetation remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, no. 2, pp. 24–49, 2021. doi: 10.1016/j.isprsjprs.2020.12.010.

[36] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[37] O. Kpüklü, X. Wei, and G. Rigoll, "You only watch once: A unified CNN architecture for real-time spatiotemporal action localization," arXiv preprint arXiv:1911.06644, 2019.

[38] L. S. Yao, S. X. He, K. Su, and Q. T. Shao, "Deep learning method of facial expression recognition based on gabor filter bank combined with PCNN," *Wirel. Pers. Commun.*, vol. 125, no. 2, pp. 1483–1500, 2022.

[39] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet?," in *Proc. CVPR*, New York, USA, 2018, pp. 6546–6555.

[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 36, no. 6, pp. 91–99, 2015.

[41] Z. F. Wang, L. S. Yao, S. X. Yu, and H. H. Wang, "A combination of TEXTCNN model and Bayesian classifier for microblog sentiment analysis," *J. Comb. Optim.*, vol. 45, no. 4, pp. 109, 2023. doi: 10.1007/s10878-023-01038-1.

[42] L. S. Yao and H. F. Zhao, "Deep learning method of facial expression recognition based on gabor filter bank combined with PCNN," *Wireless Pers. Commun.*, vol. 131, no. 4, pp. 955–971, 2023. doi: 10.1007/s11277-023-10463-8.

[43] X. Jiang, Y. Wang, W. Liu, S. Li, and J. Liu, "CapsNet, CNN, FCN: Comparative performance evaluation for image classification," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 6, pp. 840–848, 2019. doi: 10.18178/ijmlc.2019.9.6.881.

[44] N. Garau and N. Conci, "CapsulePose: A variational CapsNet for real-time end-to-end 3D human pose estimation," *Neurocomputing*, vol. 523, no. 7, pp. 81–91, 2023. doi: 10.1016/j.neucom.2022.11.097.

[45] C. Zhang, Y. Li, Z. Yu, X. Huang, J. Xu and C. Deng, "An end-to-end lower limb activity recognition framework based on sEMG data augmentation and enhanced CapsNet," *Expert. Syst. Appl.*, vol. 227, no. 24, pp. 120257, 2023. doi: 10.1016/j.eswa.2023.120257.

[46] Y. Li, C. Yu, and Y. Cui, "TPCaps: A framework for code clone detection and localization based on improved CapsNet," *Appl. Intell.*, vol. 53, no. 13, pp. 16594–16605, 2023. doi: 10.1007/s10489-022-03158-3.

[47] Z. Wang, C. Chen, J. Li, F. Wan, Y. Sun and H. Wang, "Linking spatial and temporal attention with capsule network for P300 detection improvement," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, no. 4, pp. 991–1000, 2023. doi: 10.1109/TNSRE.2023.3237319.

[48] Y. Lei, Z. Wu, Z. Li, Y. Yang, and Z. Liang, "BP-CapsNet: An image-based deep learning method for medical diagnosis," *Appl. Soft Comput.*, vol. 146, no. 4–5, pp. 110683, 2023. doi: 10.1016/j.asoc.2023.110683.

[49] B. Sivaiah, N. P. Gopalan, C. Mala, and S. Lavanya, "FL-CapsNet: Facial localization augmented capsule network for human emotion recognition," *Signal Image Video Process.*, vol. 17, no. 4, pp. 1705–1713, 2023. doi: 10.1007/s11760-022-02381-2.

[50] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on FER2013," arXiv preprint arXiv:2105.03588, 2021.

[51] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. Fourth IEEE Int. Conf. Automatic Face Gesture Recognit. (FG'00)*, Grenoble, France, Mar. 2000, pp. 46–53.

[52] P. Lucey *et al.*, "The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression," in *Proc. Third Int. Workshop CVPR Human Commun. Behav. Anal. (CVPR4HB 2010)*, San Francisco, USA, Jun. 2010, pp. 94–101.