An Empirical Study on the Effect of Face Occupancy on the Generalization Performance of CNN Models

Jialin Tian

Beijing-Dublin International College, Beijing University of Technology, Beijing, 100124, China
email: tianjialin@emails.bjut.edu.cn

Abstract. This empirical study investigated the impact of face occupancy on the generalization performance of Convolutional Neural Networks (CNNs), specifically focusing on three widely-used architectures: ResNet50, VGG16, and MobileNetV2. The face occupancy ratio, defined as the proportion of the image occupied by the face, is hypothesized to affect the model's ability to generalize across varying conditions. The study employs two benchmark datasets, EFE and FER-2013, to conduct this study's experiments. The datasets are preprocessed, and face detection is performed using Multi-task Cascaded Convolutional Networks (MTCNN). The face occupancy ratio is calculated for each image and categorized into bins for detailed analysis. This study trains and evaluates the CNN models on these datasets and calculates key performance metrics, including accuracy, loss, and mean absolute error (MAE). This study's analysis includes the Pearson correlation coefficient to measure the linear relationship between face occupancy and model accuracy. Additionally, the study visualizes model performance using confusion matrices and scatter plots with regression lines, highlighting the trends in model accuracy relative to face occupancy. Results indicate a strong positive correlation between face occupancy and model accuracy across all three models, with ResNet50 showing the highest Pearson correlation coefficient, followed by VGG16 and MobileNetV2. These findings suggest that higher face occupancy ratios contribute to better generalization performance in CNNs, offering valuable insights for improving facial recognition systems.

Keywords: Face Occupancy, Facial Expression Recognition, CNN Generalization.

1 Introduction

Facial expressions are a fundamental component of human non-verbal communication, serving as external manifestations of an individual's emotional state and intent. As such, they carry rich information about emotions and intentions. Typically, the basic categories of facial expressions include happiness, sadness, anger, surprise, fear, disgust, and neutrality (Fig. 1). Facial expression recognition represents a significant area of research within the fields of computer vision and artificial intelligence. It aims to automate the detection and classification of different facial expressions by analyzing image

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_61

or video data of human faces. This capability is crucial for enhancing human-computer interaction, social security monitoring, and psychological health assessment. For instance, in the educational sector, analyzing students' facial reactions enables teachers or intelligent tutoring systems to refine teaching strategies and methods to align with the students' emotional states and comprehension levels [1].

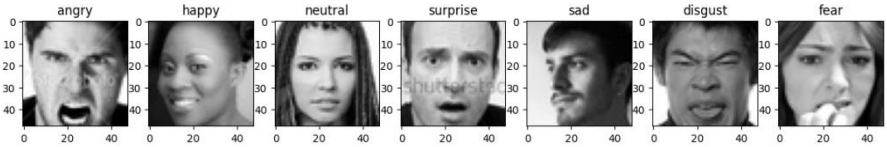

Fig. 1. The sample images of the FER-2013 dataset.

In facial expression recognition research, understanding dataset diversity—including race, gender, age, image clarity, grayscale, resizing, and obstructions—is crucial. Models trained on limited datasets like FER-2013, EFE, and CK+ in controlled labs often underperform in the diverse conditions of real-world environments due to privacy and security concerns [2].

Facial proportion, the ratio of the face to the entire image, varies with size, angle, and coverage in user-generated images, impacting model performance. Models trained on high facial proportion images may perform poorly on real-world images with lower proportions or unconventional layouts [3]. Thus, it's crucial to test models across diverse facial proportions to enhance their practicality and generalization. Training CNNs on varied datasets can improve their adaptability and accuracy, fostering technological progress and supporting equitable Artificial Intelligence (AI) applications.

Facial expression recognition has historically been widely studied as an important branch in the field of computer vision and artificial intelligence. Early research relied heavily on feature engineering, such as Local Binary Pattern (LBP) and Gabor filters, which focused on extracting manually formulated features from facial images [2]. With the development of deep learning techniques, CNNs began to be widely used in this field, significantly improving the accuracy of recognition. For example, classical CNN architectures such as AlexNet and VGGNet have been shown to achieve superior performance on publicly available datasets such as CK and FER-2013 [3].

Despite significant progress, facial expression recognition research still faces challenges with real-world data due to models often being trained and tested under controlled conditions that lack variety and complexity. Additionally, many studies fail to account for variations in facial occupancy ratios, common in images from surveillance cameras or mobile devices. Improving the model's ability to recognize images with different facial occupancy ratios can enhance usability and accuracy in applications such as sentiment analysis, human-computer interaction, and security monitoring. This research aims to provide valuable insights for developing robust and intelligent systems that perform consistently across varied environments. Although deep learning, espe-

cially CNNs, has significantly improved recognition accuracy, challenges remain regarding the complexity of expressions, different poses, racial differences, and adaptability to real-world conditions.

Despite significant advances in current facial expression recognition techniques, most models are still trained and tested on specific standardized datasets that usually contain only frontal and clear expression images. This approach ignores the diversity of facial images that is common in real-world applications, especially the different ways in which facial occupancy is represented in the images. The aim of this study is to fill this gap by systematically evaluating the performance of CNN models under different facial occupancy ratios, thereby improving the model's adaptability and robustness to real-world data. This study focuses on the performance of different convolutional neural network models (e.g., ResNet [4], VGG [5], MobileNet [6]) in the task of face expression recognition, specifically considering how the proportion of faces in photos from different datasets (e.g., FER-2013 [7] and EFE) affects the model's generalization ability. In this way, the study hopes to gain a better understanding of the effectiveness and limitations of different models when working with real-world data.

2 Methodology

2.1 Dataset Preparation and Analysis

The research utilizes two significant datasets known for their application in facial expression recognition: FER-2013 and the Extended Facial Expression (EFE). The FER-2013 dataset, created for the Facial Expression Recognition Challenge at the International Conference on Machine Learning (ICML) 2013, comprises 35,887 grayscale images of human faces. Each image, sized at 48×48 pixels, is labeled with one of seven emotion categories: happiness, sadness, surprise, fear, disgust, anger, and neutral, providing a standardized input size for consistent processing across various computational models [8]. The distributions of these two datasets are provided in Fig. 2.

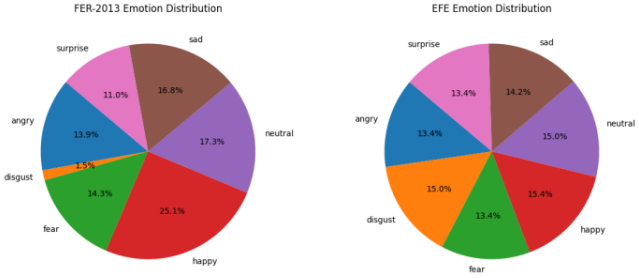

Fig. 2. The facial expression distribution of datasets.

In contrast, the EFE dataset shown in Fig. 2 is designed to augment the diversity of facial expression datasets by including a wider range of sources and more nuanced expressions. It contains approximately 5,000 images, each labeled with similar emotion

categories but featuring more subtle expressions and a broader range of emotional intensities. Unlike FER-2013, the images in the EFE dataset vary in size, commonly around 640×480 pixels, and are in color, offering a more challenging dataset for expression recognition due to the additional color information. These datasets were selected to offer a broad spectrum of facial expressions captured under varying conditions, enabling a comprehensive analysis of the models' performance across different environments.

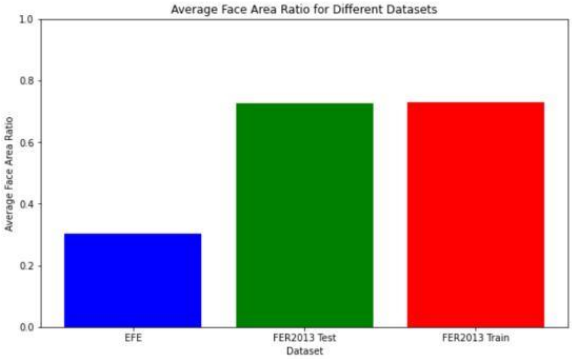

Fig. 3. Average Face Area Ratio for Different Datasets.

Additionally, the bar chart shown in Fig. 3 illustrates the average face area ratio for three datasets: EFE, FER2013 Test, and FER2013 Train. The face area ratio is defined as the proportion of the image occupied by the face. The EFE dataset has a significantly lower average face area ratio (approximately 0.3) compared to the FER2013 Test and Train datasets, which both have average face area ratios around 0.8.

2.2 Preprocessing

To prepare the images for effective training and evaluation, the study incorporated several crucial preprocessing steps. Firstly, normalization was applied where the pixel values of each image were scaled to a range of [0,1]. This was achieved by dividing each pixel value by 255, the maximum possible value, which helps in faster convergence during neural network training and standardizes the input features for enhanced performance. Additionally, images from the EFE dataset, which were originally in RGB format, were converted to grayscale. This conversion is vital to match the input format of the FER-2013 dataset and eliminate color-related variations that are irrelevant to expression recognition tasks [9].

Furthermore, all images from the EFE dataset were resized to 48×48 pixels to ensure consistency with the FER-2013 dataset. This uniformity is essential for the CNN architectures used in this study, which require fixed-size inputs. To increase the robustness of the models, image data augmentation techniques such as rotation, zoom, and horizontal flipping were employed. These techniques simulate different angles and variations in facial expressions, thus enhancing the generalizability of the trained models.

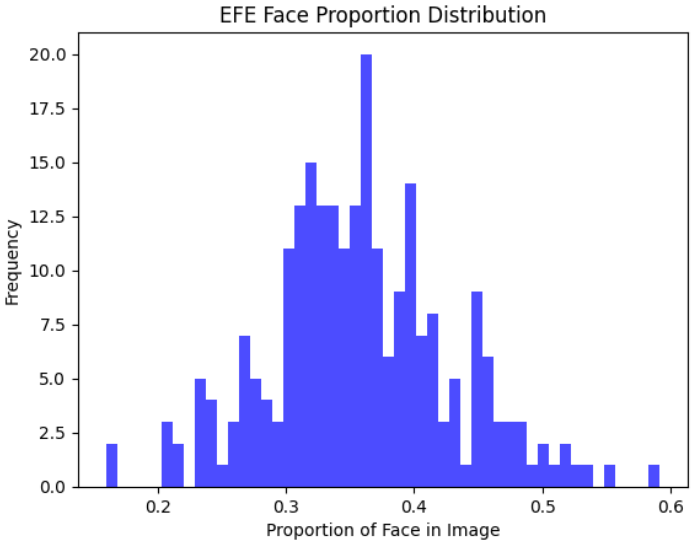

Fig. 4. EFE face proportion distribution.

An important aspect of this research involves the analysis of facial proportions, defined as the ratio of the face area to the total image area shown in Fig. 4. This was calculated using the MTCNN algorithm [10], which detects faces and their bounding boxes in images. Analyzing this metric helps in understanding how well the models perform when the face occupies different portions of the image, a common occurrence in real-world scenarios [11]. To quantify this relationship, this study used the Pearson correlation coefficient to measure the correlation between facial proportions and model performance (loss and MAE) in both datasets. This comprehensive approach to data preprocessing aims to refine the accuracy and applicability of expression recognition across varied conditions.

2.3 CNN Models

In this study, CNNs are employed due to their efficacy in visual imagery analysis. CNNs consist of multiple layers that adaptively learn spatial feature hierarchies through backpropagation, including convolutional layers that apply learnable filters, pooling layers that reduce spatial size to decrease computational complexity, and fully connected layers that interpret these features. A key feature of CNNs is their ability to develop an internal representation of a two-dimensional image, learning location-invariant features ideal for image recognition tasks. This capability is enhanced by shared weights in convolutional layers, which reduce the model's memory footprint and improve generalization across visual scenes.

The study incorporates several specific CNN models: Residual Network (ResNet), which uses residual learning and shortcut connections to address vanishing gradients,

making it suitable for recognizing fine details across deep network layers; Visual Geometry Group Network (VGG), which achieves good performance through deep layers of convolutional networks with small receptive fields followed by max-pooling; and MobileNet, designed for mobile and resource-constrained environments using depthwise separable convolutions for efficiency, making it ideal for real-time applications.

For training and evaluation, the FER-2013 dataset is utilized, with 70% allocated for training the models. This subset provides a diverse range of facial expressions under controlled conditions, allowing comprehensive feature learning. The remaining 30% serves as the primary testing set, evaluating model performance under similar conditions. Additionally, the EFE dataset acts as a secondary testing set to assess performance in varied, real-world conditions, helping to determine the robustness and generalization capabilities of the models. This methodological setup ensures a thorough evaluation of the models' abilities and their applicability in real-world scenarios, demonstrating the effectiveness and adaptability of CNNs in facial expression recognition across diverse environments.

3 Results and Discussion

The graphs in Fig. 5 show the training and validation loss and accuracy over epochs for ResNet50, VGG16, and MobileNetV2 models. In each plot, the blue and orange lines represent training and validation loss, respectively, while the green and red lines represent training and validation accuracy. All three models exhibit a decrease in loss and an increase in accuracy over time, indicating that they are effectively learning and improving.

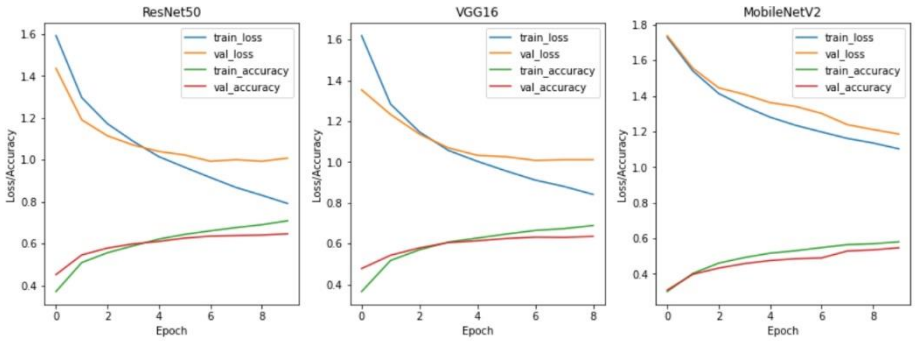

Fig. 5. Model performance over epochs for ResNet50, VGG16, and MobileNetV2 models

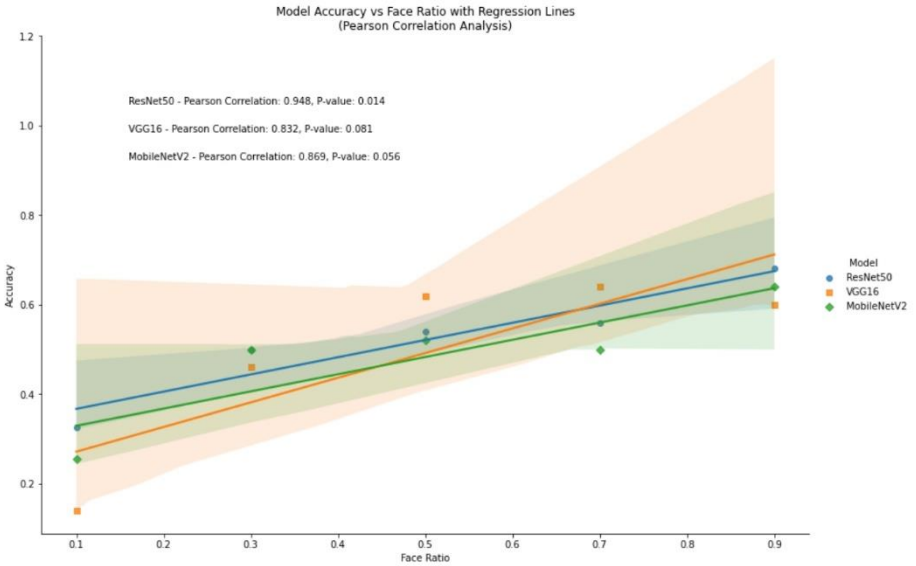

Fig. 6. The scatter plot shows the correlation between model accuracy and face ratio for three different models

The results from the Pearson correlation analysis and visual representations (Fig. 6) provide clear insights into how face occupancy affects the generalization performance of CNN models. The strong positive correlation exhibited by ResNet50 (Pearson correlation coefficient: 0.948, p-value: 0.014) indicates that its accuracy significantly improves with higher face occupancy ratios. This suggests that ResNet50, while highly effective with images that have a high proportion of face area, may struggle with images where the face occupies a smaller portion of the frame. The bar plot in Fig. 7 below shows that ResNet50's accuracy steadily increases with higher face occupancy bins, reinforcing this observation.

VGG16 demonstrated a moderate positive correlation (Pearson correlation coefficient: 0.832, p-value: 0.081), indicating that it benefits from higher face occupancy but to a lesser extent than ResNet50. The bar plot in Fig. 1 reveals that VGG16 performs consistently well across varying face occupancy bins, suggesting better adaptability and robustness. The bar plot in Fig. 7 shows that MobileNetV2 maintains a balanced performance across different face occupancy bins, and the confusion matrix in Fig. 8 highlights its efficiency in correctly classifying diverse images. The confusion matrix in Fig. 8 for VGG16 shows fewer misclassifications compared to ResNet50, further supporting its superior generalization capability. MobileNetV2 showed a moderate positive correlation (Pearson correlation coefficient: 0.869, p-value: 0.056), indicating its suitability for real-time applications where computational efficiency and generalization capability are both crucial.

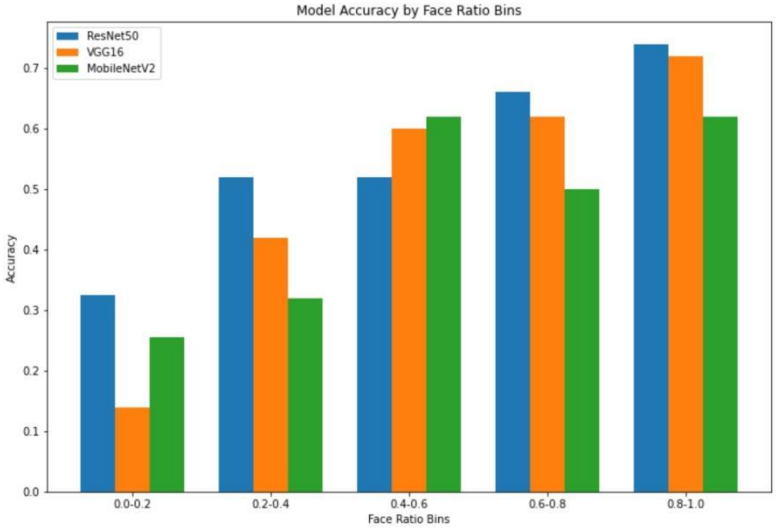

Fig. 7. Model Accuracy by Face Ratio Bins.

The scatter plot with regression lines in Fig. 7 provides a visual representation of the correlation between face occupancy and model accuracy. The regression lines show that all three models' accuracy increases as face occupancy increases, with ResNet50 showing the steepest slope, followed by MobileNetV2 and VGG16. This visual analysis corroborates the numerical results and underscores the importance of face occupancy in training datasets. The confusion matrices in Fig. 8 offer a detailed view of each model's performance across different categories. ResNet50's confusion matrix shows more significant misclassifications in categories with lower face occupancy ratios. In contrast, VGG16's confusion matrix demonstrates better performance with fewer misclassifications, highlighting its robust generalization. MobileNetV2's confusion matrix indicates balanced performance, with moderate accuracy and fewer extreme errors.

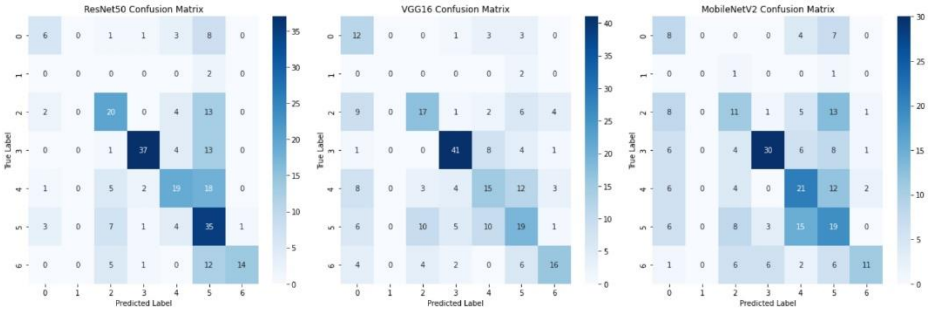

Fig. 8. Confusion matrixs of the ResNet50, VGG16, MobileNetV2.

The strong positive correlations between face occupancy and model accuracy across all three models emphasize the need for considering face occupancy in training datasets.

Training models on datasets with higher face occupancy ratios can significantly enhance their generalization performance. This study highlights that while models like ResNet50 can achieve high accuracy with high face occupancy images, they may require more diverse training data to handle real-world scenarios effectively. In conclusion, VGG16 demonstrated the best overall performance across both datasets, indicating robustness and better generalization capabilities. ResNet50, while highly accurate on images with high face occupancy, struggled with images where the face occupancy was lower. MobileNetV2 balanced efficiency and accuracy, performing moderately well across varying face occupancy scenarios. The positive correlation between face occupancy and model performance underscores the critical role of considering face occupancy in training datasets to improve the generalization of CNN models in diverse real-world conditions.

4 Conclusion

This study examined the effect of face occupancy on the generalization performance of CNN models, specifically ResNet50, VGG16, and MobileNetV2, using the FER-2013 and EFE datasets. As highlighted in the introduction, accurate facial expression recognition is crucial for enhancing human-computer interaction, security monitoring, and psychological health assessments. The methodology included comprehensive preprocessing steps and the use of Pearson correlation coefficients to understand the relationship between face occupancy ratios and model performance. The results revealed significant performance differences among the models. VGG16 demonstrated the best overall performance and robustness across both datasets, indicating superior generalization capabilities. ResNet50, while performing well on images with high face occupancy, struggled significantly with lower face occupancy images. MobileNetV2 showed balanced efficiency and accuracy, performing moderately well in diverse conditions but still lagging behind VGG16 in adaptability. The correlation analysis confirmed a strong positive relationship between face occupancy ratios and model performance for all three models, with ResNet50 exhibiting the highest Pearson correlation coefficient (0.948, p-value: 0.014), followed by MobileNetV2 (0.869, p-value: 0.056) and VGG16 (0.832, p-value: 0.081). This emphasizes the importance of training models on datasets with higher face occupancy ratios to improve their generalization performance in real-world applications. The findings align with the hypothesis that models trained on high face occupancy images perform better, highlighting the need for diverse training data that includes a wide range of face occupancy scenarios.

References

1. Mou, Y., Lei, Z., Tian, X., & Liu, X.: Application Research on Classroom Facial Expression Recognition Technology Based on Deep Learning. Journal of Hunan Radio and Television University, (01), 17–24 (2023).

2. Adjabi, I., Ouahabi, A., Benzaoui, A., & Taleb-Ahmed, A.: Past, Present, and Future of Face Recognition: A Review. Electronics (2020).
3. Peng, X., & Qiao, Y.: Progress and Challenges in Facial Expression Analysis. Journal of Image and Graphics, 25(11), 2337–2348 (2020).
4. He, K., Zhang, X., Ren, S., & Sun, J.: Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778 (2016).
5. Simonyan, K., & Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
6. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510-4520 (2018).
7. Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y.: Challenges in representation learning: A report on three machine learning contests. In Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20, pp. 117-124. Springer, Berlin, Heidelberg (2013).
8. Wu, Y., & Chen, X.: Facial Expression Recognition System Based on Improved ResNet. Information and Communication, 07, 37–39 (2020).
9. Wang, W., Zhou, X., He, X., Qing, L., & Wang, Z.: Facial Expression Recognition Based on Improved MobileNet Network. Computer Applications and Software, 37(04), 137–144 (2020).
10. Zhang, K., Zhang, Z., Li, Z., & Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, 23(10), 1499-1503 (2016).
11. Fan, W., & Tian, X.: Research on Facial Expression Recognition System Based on Deep Learning. Modern Information Technology, 6(20), 90-93+97 (2022).