Enhanced Facial Expression Recognition Based on ResNet50 with a Convolutional Block Attention Module

Liu Luan Xiang Wei, Nor Samsiah Sani
Center for Artificial Intelligence Technology-Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, Selangor, 43600, Malaysia

***Abstract*—Deep learning techniques are becoming increasingly important in the field of facial expression recognition, especially for automatically extracting complex features and capturing spatial layers in images. However, previous studies have encountered challenges such as complex data sets, limited model generalization, and lack of comprehensive comparative analysis of feature extraction methods, especially those involving attention mechanisms and hyperparameter optimization. This study leverages data science methodologies to handle and analyze large, intricate datasets, while employing advanced computer vision algorithms to accurately detect and classify facial expressions, addressing these challenges by comprehensively evaluating FER tasks using three deep learning models (VGG19, ResNet50, and InceptionV3). The convolutional block attention module is introduced to enhance feature extraction, and the performance of the model is further improved by hyperparameter tuning. The experimental results show that the accuracy of VGG19 model is the highest 71.7\% before the module is integrated, and the accuracy of ResNet50 is the highest 72.4\% after the module is integrated. The performance of all models was significantly improved through the introduction of attention mechanisms and hyperparameter tuning, highlighting the synergistic potential of data science and computer vision in developing robust and efficient in facial expression recognition systems.**

***Keywords*—Data science; computer vision; deep learning; facial expression recognition**

I. INTRODUCTION

Deep learning has emerged as a revolutionary and transformative technology within artificial intelligence. Particularly in facial expression recognition (FER), artificial intelligence (AI) applications introduce new research opportunities and significantly advance the field. Facial expressions stem from the coordinated movements of facial muscles in response to emotions [2]. Emotions can temporarily change the shape of the face because of changes in the movement of facial muscles since facial muscles are not independent of each other [8].

Researchers have tried many ways to interpret and decode facial expressions and extract important features from facial images [27]. A person's emotions can influence the efficacy of face recognition, as varying facial expressions can affect the outcomes. Kim S and Kim H found a certain relationship between facial Action Coding Units (AUs) and Emotion labels in the FER dataset [40]. Being a primary means of expressing human emotions, facial expressions are crucial for social interaction. They transmit non-verbal signals interpreted by the brain, which can be recorded in images or videos [3]. The human brain can automatically recognize emotions without delay [6]. Cha et al. proposed a FEMG-based FER system based on the Riemannian manifold approach, and further develops an online FER system that can make an avatar's expression reflect the user's facial expression in real time, thus demonstrating that our FER system can potentially be used for practical interactive VR applications such as social VR networks, intelligent education, and virtual training [15]. However, this is a challenging task for computers [50]. As AI technology progresses, machines are increasingly able to replicate the functions of the human brain, making FER applicable across diverse domains, like security surveillance or mental health evaluations. For instance, Dong et al studied that the CGSSNet network established based on the DenseNet algorithm has significant advantages in glioma MRI image segmentation, providing a new idea for the diagnosis and treatment of glioma [21]. Automated FER can identify clinically significant facial features, distinguishing disease states and serving as specific biomarkers [41] [60]. Li et al. (2018) proposed a computer-aided framework for the early differential diagnosis of pancreatic cysts. DenseNet learned advanced features from the entire abnormal pancreas, mapped the appearance of medical imaging with different pathological types of pancreatic cysts, and integrated the significance map into the framework. In a cohort of 206 patients with 4 pathologically confirmed pancreatic cyst subtypes, the overall accuracy rate was 72.8%, significantly higher than the baseline accuracy rate of 48.1% [43]. Like the popularity of smartphones and social platforms, FER has become more important in daily life. For example, analysis of users' facial reactions can improve user experience and personalized content recommendations [9]. To create a more immersive VR social interaction, users can wear a head-mounted display (HMD) with RGB cameras that continuously capture images of their lips to interpret facial expressions. Another example in the field of security monitoring is an accurate FER system can assist in identifying suspicious behavior or emotional abnormalities. Liu and Fang designed a three-level cascade algorithm model for expression recognition in educational robots. By using CK+ and Oulu-CASIA expression recognition database, compared with other common cascaded convolutional neural network methods, the accuracy and speed of facial expression recognition are significantly improved [48]. These technologies assist scientists in accurately identifying unethical behaviors or emotions from facial cues and

predicting future behaviors and emotional states based on collected data [4]. The book recommendation system integrates expression and face recognition with tracking book browsing times to determine users' ages and suggest books accordingly [63].

Despite its potential, achieving high accuracy in FER remains challenging due to the inherent complexity and variability of factored expressions [13]. Deep learning can automatically extract people's facial features, identify different expressions, and meet the expected requirements of classification. Face feature detection and recognition and convolutional neural network classification. The advantage of facial markers is that classification is very robust, even with limited memory [28]. However, there are still some limitations in the performance of existing deep learning models in facial expression recognition tasks, such as the generalization ability of the model, the ability to capture different facial details, and the performance in the case of unbalanced data. Therefore, identifying and proposing the most effective deep learning model is of great significance for FER. First, the most effective models can significantly improve the accuracy of FER, help better understand and analyze human emotions, and be applied to many fields, such as human-computer interaction, emotional computing, and mental health monitoring. Second, by exploring and comparing different deep learning architectures, especially convolutional neural networks (CNNS), it is possible to discover which specific network structures and feature extraction methods perform best in FER tasks, guiding future research and applications. In addition, the most effective models can achieve efficient and accurate facial expression recognition in the case of limited resources, thus reducing computational costs and improving the practicality and scalability of the system. Therefore, this study aims to explore various deep learning architectures and identify the best-performing FER models by testing a widely used benchmark dataset in the field, providing a diverse set of facial images and their corresponding emotional labels. This will help solve the challenges that exist in FER and drive the development and application of this field. Therefore, the first question of this study is, is it possible to identify the best-performing FER model on an FER dataset by exploring various deep learning architectures, especially convolutional neural networks (CNNS)?

Attention mechanisms show great promise in improving the performance of deep learning models by focusing on relevant features while suppressing irrelevant features [23]. However, there are still some limitations in the performance of existing deep learning models in facial expression recognition tasks, such as insufficient ability to capture subtle facial features and low computational efficiency. Introducing attention mechanisms, such as convolutional block Attention modules (CBAM), can somewhat alleviate these problems. CBAM effectively enhances the model's feature extraction capability by combining channel and spatial attention while keeping the computational overhead low. In 2022, Ju and Zhao combined attention mechanisms to propose a new masked attention mechanism Parallel Network (MAPNet), which significantly improved the classification performance and accuracy of three different datasets of the FER task RAFDB, AffectNet and FEDRO by 0.001, 0.0118 and 0.0325, respectively [33]. In 2023, Putro et al. proposed a real-time facial expression classification method based on a dual attention module convolutional neural network, which achieved an excellent result of 0.9865 and 0.9688 in CK+ and JAFFE datasets, respectively [57]. However, introducing attention mechanisms in models with different structures is time-consuming. To solve this problem, Sanghyun et al. designed a simple and efficient feedforward Convolutional neural network attention module (CBAM) that can be seamlessly integrated into any CNN architecture with negligible overhead. Thus, it provides new ideas and methods for combining deep learning and attention mechanisms [55]. This study aims to integrate CBAM into each layer of a deep learning model to investigate its impact on the performance and efficiency of deep learning models in FER tasks. By introducing CBAM, we aim to significantly enhance deep learning models' feature extraction and discrimination capabilities, thereby improving FER tasks' overall performance and efficiency. This will contribute to developing more accurate and efficient FER systems and provide valuable experience and methods for future research and applications [22]. Therefore, this study raises a second question: Does the introduction of CBAM in deep learning models affect the performance and efficiency of FER tasks? By systematically testing and validating the impact of CBAM, we expect to provide better solutions and new research directions for the FER field.

Hyperparameter tuning is a key aspect of optimizing deep learning models' performance [1]. The performance of existing deep learning models in facial expression recognition (FER) tasks is often affected by the selection of hyperparameters, such as learning rate, batch size and regularization techniques. These hyperparameters directly affect the training process and final performance of the model. However, it is still a challenging task to select the optimal combination of hyperparameters to achieve the model's best performance and generalization ability. By systematically adjusting these hyperparameters, the predictive power of FER deep learning models on FER datasets can be significantly enhanced. Reasonable hyperparameter Settings can not only improve the accuracy of the model but also effectively reduce the overfitting phenomenon, thus improving the generalization ability of the model [59]. Rigorous experiments and evaluations are performed during training to determine an optimal set of hyperparameter combinations that maximizes model performance, minimizes overfitting, and improves generalization. This study aims to explore and verify the effects of different hyperparameter configurations on the performance of the FER deep learning model through hyperparameter tuning. Hyperparameter tuning is important because it can significantly improve the model's training effect and practical application performance, thus achieving the most advanced performance in the FER task. By determining the best combination of hyperparameters, best practices can be established for deep learning model training of facial emotion recognition, and scientific basis and methods can be provided [11]. Therefore, the third question in this study is: Can hyperparameter tuning maximize model performance and improve generalization? Through the hyperparameter tuning and verification of the system, we expect to provide better solutions and new research directions for the FER field.

This study aims to achieve the following three objectives:

- To recommend the most effective deep learning models for facial emotion recognition (FER) utilizing the FER2013 dataset.
- To propose the attention mechanism, CBAM (Convolutional Block Attention Module) is added to each model layer to explore the differences in model performance and efficiency on the same dataset.
- Enhance the performance of deep learning models through hyperparameter tuning during the training phase, thereby optimizing their predictive capacity for facial emotion recognition.

In this study, we propose an enhanced facial expression recognition (FER) model based on ResNet50 with a Convolutional Block Attention Module (CBAM). FER is a challenging task due to high intra-class variability, subtle interclass differences, and the presence of occlusions and noise. ResNet50 serves as a robust backbone for feature extraction, while CBAM enhances the network's ability to focus on both spatially and channel-wise relevant features. This combination allows the model to address the shortcomings of existing methods by improving feature localization and discriminability with minimal computational overhead."

II. Literature Review

A. Facial Emotion Recognition

Hardware technology development has addressed the significant computational power issues associated with deep learning due to its complexity, power requirements, and relatively low cost [5]. CNN has emerged as one of the most revolutionary technologies for FER. CNNs can learn from large datasets to automatically extract and combine features necessary for recognizing various expressions. CNNs build an understanding of complex expressions by abstracting image features layer by layer through a multi-layered structure. This layer-bylayer approach to learning is well suited to the FER task because it allows the model to recognize and distinguish subtle differences in expression.

In recent years, many derivative models of CNN and classification methods for facial recognition have been developed, such as data set preprocessing and feature extraction methods. Chen et al introduced an additional branch to generate a mask, thus focusing on the movement area of the facial muscles. In order to guide face learning, we propose to combine prior domain knowledge and use the average difference between neutral faces and corresponding facial faces as training guidance, which is effective compared with the most advanced methods [18]. Jia In the first stage, offline subnetworks were trained in three subnetworks to achieve convergence (the three subnetworks are AlexNet, VGGNet and ResNet derivative). In the second stage, the output layer of these three subnetworks was removed and predicted by SVM, and the accuracy rate reached 0.7127[26]. Liu investigated the improved VGG-16 CNN, enhancing the VGG-16 network by optimizing the third and fourth convolutional layers [39]. Instead of the original SoftMax classifier, a 7label SoftMax classifier was employed. It replaced the original ReLU activation function with LeakyReLU. Experiments on the FER2013 showed an accuracy of 0.7242, higher than the previous rates of 0.6631 and 0.7138 without improvements to VGG and ResNet, respectively. Part of FER datasets are shown in Fig. 1:

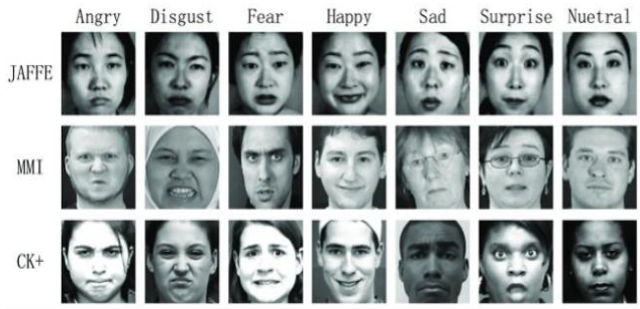

Fig. 1. Presentation of different FER datasets (Part).

To simplify the artificial feature extraction process in traditional FER and capture more diverse features, Changing proposed a method that integrates multiple CNN models, using three different CNN subnetworks for comprehensive prediction. This approach achieved an accuracy of 0.701 on the FER2013 [12]. Dwijayanti et al. indicated that there was not much research on face recognition and FER objectives; consequently, they employed CNNs to tackle this challenge [14]. Unlike other experiments, they used the original image as CNN input and directly used the VGG-f model for FER tasks, overcoming the problems of underfitting and overfitting the CNN framework and also getting a good performance.

Fu investigated the impact of incorporating visual attention mechanisms into deep learning for FER [17]. In their approach, three fully connected layers in the training phase were substituted with three convolutional layers to generate test results for the entire network, thereby mitigating the limitations of full connection. Additionally, the SE block was applied to normal VGG, and the results verified the effectiveness of SEVGGNET with 0.668 accuracy. Das and Neelima proposed an improved pretreatment stage. This includes extracting local binary features used to express classification. These feature vectors are connected and used in shallow neural networks with minimal complexity and fewer layers to optimize expression recognition processes as well as gently enhance decision trees. Applying this local binary feature-based neural network (LBF-NN) approach to three different popular databases, more than 93% results were achieved, even when compared to a variety of complex and advanced algorithms [20].

Mohamed et al. achieved improvements by applying CNN to examine the Alex network architecture, applying transfer learning methods and modifying the full connection layer using support vector machine (SVM) classifiers [53]. The improved model has a classification recognition rate of about 0.6429 for the selected expressions. The system has achieved satisfactory results on the ice-MEFED dataset. The improved model has a classification recognition rate of about 0.6429 for the selected expressions. Lee et al. proposed an ensemble framework to boost the reliability of FER models using three models: VGG16, InceptionResNetV2, and EfficientNetB0 [33]. The results indicated that the model recognition accuracy priority edge ensemble learning algorithm improved by 0.0281.

B. Deep Learning

Facial expression (FE) has powerful potential and is a universal human communication form closely linked to mental states, attitudes, and intentions. By analyzing facial expressions displayed by humans or objects, computers can effectively process and interpret human emotions, forming the core of the FER system. The development of FER in this field has witnessed the transformation from the preliminary geometric feature method to the current deep learning technology, which has profoundly affected our understanding and practice of emotional computing and human-computer interaction. Early FER studies relied on manual extraction of facial features and simple pattern-matching techniques. Researchers try to identify expressions by pinpointing key features. However, these methods are insufficient when faced with the diversity and complexity of human expressions and struggle to adapt to dynamic real-world environments.

CNN are potent visual recognition tools whose design is inspired by biological vision systems. It is mainly used for image processing and classification [6]. Compared to other classification algorithms, it is an algorithm that takes images and is able to distinguish one from another with minimal preprocessing. Automatic feature detection without human supervision is the main advantage of CNN over other algorithms. In addition, a method combining global appearance features with local geometry features is proposed [11]. Specifically, they provide not only the raw images to the facial expression recognition network but also the facial markers associated with them. A typical CNN comprises various layers: convolutional layers, activation functions, pooling layers, and fully connected layers.

Transfer Learning is an important method in the field of deep learning. In deep learning, transfer learning usually involves taking a pre-trained model on a large data set, like ImageNet, and applying it to a new, related task. The advantage of transfer learning is that it can leverage the complex feature extraction capabilities that have been learned. The core idea of transfer learning is that there is a commonality between certain learning tasks so that what is learned on one task can be reused on another. For example, a model trained on pictures of animals may have learned to recognize features such as eyes, ears, etc., which may also be useful for recognizing other types of objects, such as faces[49]. Lee et al. applied transfer learning, fine-tuning, and data enhancement to the training and validation of the Facial expression recognition 2013 (FER-2013) dataset. Experimental results show that the model recognition accuracy of the proposed priority edge ensemble learning algorithm is improved by 2.81% [42].

Attention Mechanism allows the model to prioritize significant parts of the input data by assigning variable weights to different image regions. This reflects the importance of these regions for the final task. For instance, features such as the eyes and mouth may be more recognizable in FER tasks than in other parts. Selective weight assignment enables the model to concentrate on particular input sections while disregarding others, thereby achieving the intended output. This is known as hard attention, or it can be incorporated into the model in a differentiable way, allowing the entire network to be trained using techniques such as gradient descent. For example, during the COVID-19 pandemic, many people wore masks, and when faces are partially covered or affected by interference factors like large pose changes, it hampers feature extraction and reduces FER performance. CBAM is a kind of attention mechanism in DL that is employed to improve CNN's feature representation capacity. By explicitly modelling images' spatial and channel dimensions, the network can focus on key areas, thereby improving its performance. CBAM can be regarded as a lightweight plug-in that is easy to integrate into the existing CNN architecture.

C. Attention Mechanism

Jin et al (2022) have introduced Transformer encoder to model the remote dependency between different facial areas and capture the global relationship between different facial units, complementing the spatial locality of CNN [36]. But the attention mechanism mimics human attention, allowing the model to prioritize significant parts of the input data by assigning variable weights to different image regions. This reflects the importance of these regions for the final task. For instance, in FER tasks features such as the eyes and mouth may be more recognizable than other parts. Selective weight assignment enables the model to concentrate on particular input sections while disregarding others, thereby achieving the intended output. Liu et al proposes an adaptive multi-layer perceptual attention network that extracts global, local, and significant facial emotional features using different fine-grained features to understand the potential diversity and key information of facial emotions [46]. This is known as hard attention, or it can be incorporated into the model in a differentiable way, allowing the entire network to be trained using techniques such as gradient descent. For example, during the COVID19 pandemic, many people wore masks, and when faces are partially covered or affected by interference factors like large pose changes, it hampers feature extraction and reduces FER performance. The flow diagram of the attention mechanism on FER is shown in Fig. 2.

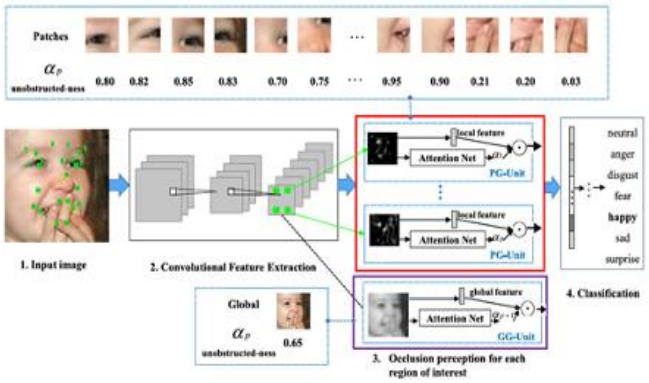

Fig. 2. Diagram of attention mechanism example in FER.

Attention mechanisms enable the model to dynamically focus on the most relevant parts of the input for a given task. For instance: In vision tasks, attention can focus on important regions of an image (e.g., objects or edges). In text processing, attention identifies key words or phrases that are crucial for understanding the context. This focus helps the model prioritize meaningful information while ignoring less relevant or redundant features. Attention mechanisms enhance feature

extraction by weighting input features according to their importance. These weights are computed adaptively during training, allowing the model to learn a richer, context dependent representation of the data. For example, in transformers, attention layers allow the model to learn contextual relationships between different parts of the input, which is critical for tasks like language translation or image captioning.

The Convolutional Block Attention Module is a module that combines the channel attention mechanism and spatial attention mechanism, aiming to improve the feature expression ability of convolutional neural networks. It is also an attention mechanism in deep learning. By explicitly modelling the spatial and channel dimensions of the image, the network can selectively focus on key areas, thereby improving its performance. Its authors Woo et al. (2018) indicated that its flexibility and versatility can be applied to different CNN network architectures (such as VGG, ResNet, etc.), and it can show good adaptability and versatility in different tasks, such as image classification semantic segmentation, and object detection. The CBAM structure is relatively simple and can be seen as a lightweight plug-in. It is easy to integrate into the existing CNN architecture as a module enhancement, disorder greatly modify the original network structure, and finally, through the channel-by-channel and pixel-by-pixel weighted way, improve the model's attention to important features, thus improving the training and reasoning efficiency, especially in the processing of complex tasks. The specific architecture of CBAM is shown in Fig. 3:

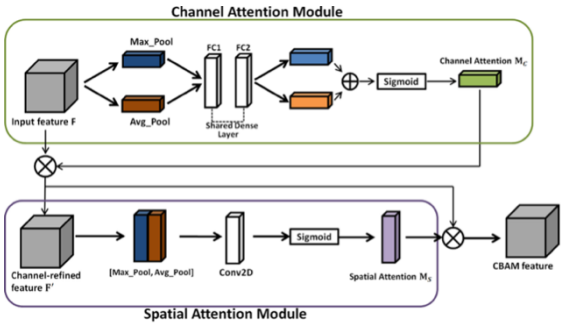

Fig. 3. Diagram of attention mechanism example in CBAM.

D. Deep Learning Models

Deep learning has made remarkable progress in the field of FER, mainly reflected in automatic feature extraction, efficient processing of large-scale data, nonlinear modelling ability, end-to-end training, combining attention mechanisms, etc. Tang (2013), this study shows for the first time that deep learning can automatically extract multi-level features without manual feature design compared to traditional machine learning methods such as SVM and KNN, which not only simplifies the process of special engineering but also significantly improves recognition accuracy. To demonstrate the excellent performance of deep convolutional neural networks on large-scale datasets, Hinton et al. (2012) found that CNN significantly improved the classification accuracy of image tasks through training on ImageNet datasets. Zhang (2017), through practice training and multi-task learning, the study et al. shows that the original image is directly learned to the final classification without the need for intermediate step feature extraction and selection, significantly improving the performance in complex scenes. After Woo et al. proposed CBAM, the parameters added to the deep learning model did not increase significantly, but the average accuracy of VGG16 and MobileNet increased by 0.0015 and 0.0024, respectively. Jin et al. (2023) designed an image enhancement algorithm using Super resolution generative adversarial (SRGAN) and adaptive gray normalization (AGN) based on the data sets and characteristics of convolutional neural networks, and tested the Fer2013 data set, and the accuracy was improved from 68.03% to 70.04% [37].

TABLE I. COMPARISON OF DIFFERENT DEEP LEARNING ALGORITHM STRUCTURES BASED ON FER-2013

Author	Model	Accuracy (%)	Year
Rajesh Kumar [38]	CNN EmotionNet	67 66.71	2023
Xu & Zhao [64]	AlexNet OneNet	64.29 54.29	2020
Putro et al. [57]	VGG13 ResNet	73.03 72.4	2020
Lu et al. [47]	VGG Inception ResNet	72.7 71.6 72.4	2023
	Ensemble CNN	75.2	
Pramerdorfer & Kampel et al. [56]	Inception ResNet	71.6 72.4	2016
Sahoo et al. [58]	6-layer CNN 10-layer CNN	66.67 68.34	2023
	VGG-16	63.68	
Lee et al. [39]	Fine-Tune VGG16 InceptionResNetV2 Fine-Tune EfficientNetB0	66.65 67.71 67.46	2022
	Priority Ensemble CNN Algorithm	70.52	
Chen et al. [16]	FERW	71	2018
Jia et al. [32]	Ensemble CNN	71.27	2020
Joseph et al. [35]	CNN	67.18	2022
Zhang et al. [36]	LeNet-5/VGG-16	70.1	2023
Liu[45]	VGGNet	71.42	2023
Muhamad et al. [51]	CNN	54	2021
Meena et al. [49]	InceptionV3	73.09	2023
Alexeevskaya et al. [10]	CNN	60.54	2022
Fu [24]	VGG16 VGG16+SENet	65 66.8	2022
	MobileNet	68.03	
Jin et al. [34]	SRGAN-MobileNet AGN-MobileNet	69.07 68.92	2023
	SRGAN+AGN-MobileNet	70.04	

Generally, complex preprocessing technology and data enhancement methods are helpful to improve the accuracy of the model. From the table, we find that the VGG model performs well on multiple data sets, such as FER+ up to 0.806, FER2013 up to 0.7303 and CK+ up to 0.8875, indicating that VGG has been widely used in different studies with stable performance. Different researchers have adopted a variety of preprocessing techniques, such as image cropping and adaptation Strong sex; The ResNet model has shown good accuracy on multiple data sets in the table, such as 0.724 accuracy on FER-2013 data set and 0.8726 accuracy on RAF-DB data set. ResNet solves the

problem of gradient disappearance in deep networks through residual connection. Performance is superior, but different data enhancement methods significantly impact ResNet's performance and require careful adjustment. The Inception model has an accuracy of 0.7309 on the FER-2013 dataset and 0.727 on the CK+ dataset. The Inception model can capture multiscale features and enhance feature expression ability through convolution kernels of different sizes. chuanjie et al proposed a facial expression recognition method that integrates multiple convolutional neural network models and uses three different CNN subnetwork models for comprehensive prediction. Experiments show that the recognition accuracy of this method on FER2013 and CK+ datasets is 70.1% and 94.9%, respectively [19].

Compared with other models, although the traditional CNN model has a simple structure and low computing resource requirements, its accuracy is generally low, for example, only 0.67 on the FER-2013 dataset. The three models, EmotionNet, AlexNet, and OneNet, perform well on specific data sets, but their overall performance is inferior to VGG, ResNet, and Inception, and they are larger than that of specific preprocessing techniques. Other models, such as VGG-f, AMP-Net, and AFTransformer, perform well in specific application scenarios and data sets. For example, the accuracy rate of AMP-Net on the RAF-DB dataset is 0.8925, but the model is relatively special and has low universality. Therefore, compared with other models, VGG, ResNet and Inception have significant advantages in accuracy and adaptability. All three models performed better in the table than most others, demonstrating their strong capabilities in the FER task. When selecting a specific model, you can make trade-offs and choices based on computing resources, training time, and application scenarios. From the literature review, we can see from Table I that different models have different performances in different data sets, and preprocessing technology significantly impacts the mode's performance.

Existing CNN-based FER models often lack attention mechanisms, treating all features equally, which limits their ability to differentiate subtle expressions or handle occlusions effectively. While attention-based methods improve feature extraction, they are often computationally expensive or focus solely on spatial or channel-wise attention. Our method addresses these limitations by integrating CBAM into ResNet50, providing both spatial and channel-wise attention while maintaining efficiency.

III. METHODOLOGY

A. Research Framework

The deep learning framework can be approached as an optimization problem to identify model parameters that minimize the loss function, following steps from data preprocessing, model construction, training optimization, and performance evaluation. The study is divided into four phases: data understanding, data preparation, modelling, and evaluation. Fig. 4 illustrates the summary of tasks in each phase. The overview of each stage of this study is as follows:

1) Business understanding: This phase includes evaluating the current state of the application of deep learning-based models to FER tasks by reviewing existing publications. The aim is to identify research gaps and set research goals.

2) Data understanding: This stage starts with an initial exploration of the data set, involving data collection and distribution checks to grasp the basic features and structure of the data. The goal is to familiarize yourself with the data and spot any potential data quality issues.

3) Data preparation: In this stage, raw data needs to be converted into clean data suitable for deep learning development. The third chapter also expounds on this stage. This includes data transformation, data enhancement, data segmentation, and coding.

4) Modelling: In this phase, the selection and implementation of modelling techniques will be explained, and the techniques needed to build predictive models will be discussed. Additionally, it involves fine-tuning the model parameters to optimize performance.

5) Evaluation: This stage encompasses reviewing the entire development process for the model, assessing the performance of the developed model, and evaluating its stability and validity through various evaluation parameters and statistical tests. It also includes verifying whether the research objectives have been met.

6) Deployment: In this phase, the insights gained from developing the FER deep learning model are communicated to the stakeholders. In this study, this stage is limited to presenting the results of developed deep learning models.

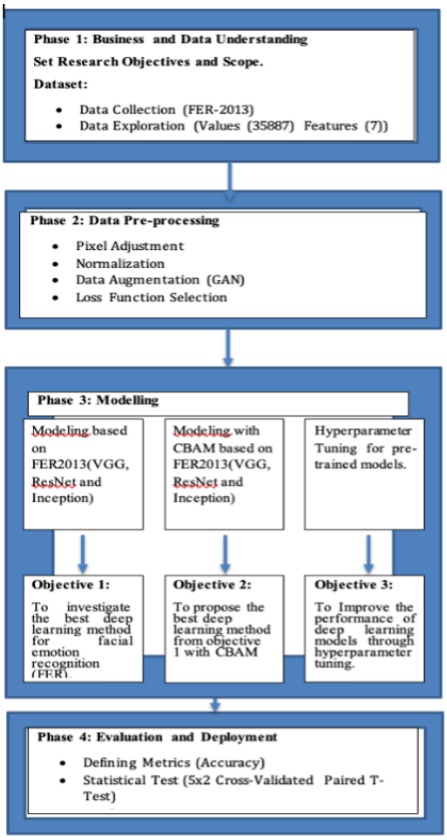

Fig. 4. Research framework.

B. Phase 1: Business and Data Understanding

The Business understanding stage is the initial stage of this study, which aims to understand the current research status of FER tasks and the application of deep learning in this field. At this stage, research objectives are developed based on research gaps identified through a comprehensive literature review. The research objective lays the foundation for the subsequent stage of this study. By establishing clear research objectives, this study ensures a targeted exploration of the application of deep learning models to FER tasks, particularly the classification prediction of FER tasks using transfer learning methods and combining attention mechanisms.

Current FER databases usually include a small number of subjects and provide only a few sample images for each expression. They often have a limited variety of subjects or minimal differences between groups, making FER tasks in real-world scenarios more challenging [10]. As shown in Table II, FER-2013 (Facial Expression Recognition 2013) is a publicly available dataset for FER that includes a wide range of expressions, from happy to sad to surprised. The entire dataset consisted of about 32,298 grayscale 48x48 pixel faces, each labelled with an emotion, such as happy, sad, or angry. These images are all from the web, uploaded by different people, and then there are artificial intelligence helpers, platforms like Amazon's Mechanical Turk, to label the facial expressions appropriately. For machine learning or computer vision researchers, FER-2013 can be used to explore differences in emotional expression in different cultural contexts. These faces from all over the world provide rich materials for the study of cross-cultural emotional communication.

They serve as a benchmark assessment for the performance of FER algorithms. Its advantage is that all images are preprocessed and are uniformly 48x48 pixels and each image is labelled, which is why it has become the standard for comparing the FER algorithm. However, limitations also exist because all images are grey and lack binary colour information, which may limit the features the model learns, and the images that are collected from the Internet may not fully reflect the natural state of people's expressions in real-life scenarios. Moreover, the dataset may lack diversity regarding race, age, and background. Another significant issue is the imbalance in the dataset; some categories have substantially more samples than others. For instance, the happy category contains 8989 samples, far exceeding the 547 samples in the Disgust category.

C. Phase 2: Data Preprocessing

This study mainly tests three deep learning models: VGG, ResNet and Inception. They are all CNN-derived models, and the models are optimized by adjusting parameters and hyperparameters. Data preprocessing is a key step in ML and data analysis, aiming to convert original data into a suitable form for further analysis and modelling. The usual steps of the FER system are to preprocess the image, extract the features from the preprocessed image, and classify the extracted features [25]. Fig. 5 and Fig. 6 show the imbalance between the test set and the training set of the FER2013. Table III shows the number of data sets after the planned data enhancement.

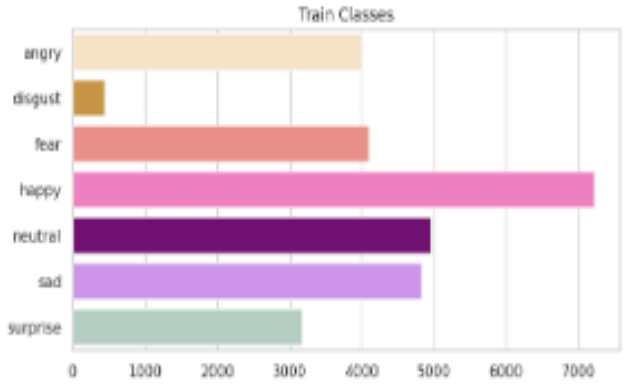

Fig. 5. Bar chart of FER-2013 trainset.

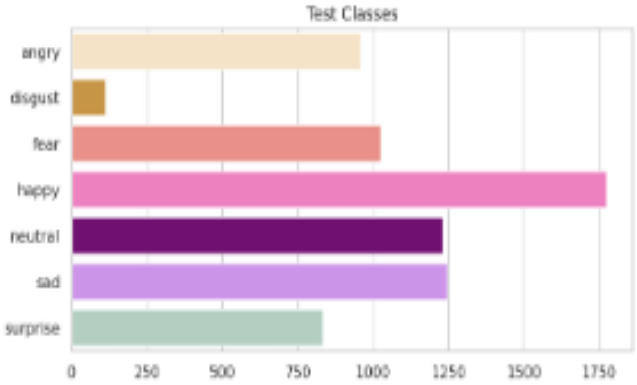

Fig. 6. Bar chart of FER-2013 testset.

TABLE II. DESCRIPTION OF FER-2013

Attribute ID	Attribute Name	Attribute Testset	Attribute Trainset	In Total
0	Angry	958	3995	4953
1	Disgust	111	436	547
2	Fear	1024	4097	5121
3	Happy	1774	7215	8989
4	Sad	1247	4830	6077
5	Surprise	831	3171	5002
6	Neutral	1233	4965	6198
In Total		3589	28709	32298

TABLE III. AUGMENTATION COUNTS PER CLASS

Name of Class	Augmentation Counts
Angry	1675
Disgust	5234
Fear	1573
Happy	0
Neutral	705
Sad	840
Surprise	2499
Total Amount after Augmentation	41235

Image Processing: Different deep learning models have different requirements for the input of images, and adjusting the input images to a uniform size ensures that the model can handle them indiscriminately [7]. If the image is saved at a larger size, it means more pixels and higher computational complexity. By adjusting the image size, enough information can be retained while reducing the need for computational resources, and in some cases, adjustment Size helps models better focus on key features of facial expressions rather than other irrelevant parts of the image. I set a standard for image input that we call the standard form of input, and we require the image to always be in the standard form [44]. While my test models usually require larger input sizes, for example, VGG and ResNet require 224 x 224, Inception requires 299 x 299, so the image needs to be adjusted to the desired size before entering the model without destroying the aspect ratio of the image, so as not to affect the model representation. The resized in FER2013 dataset is shown in Fig. 7:

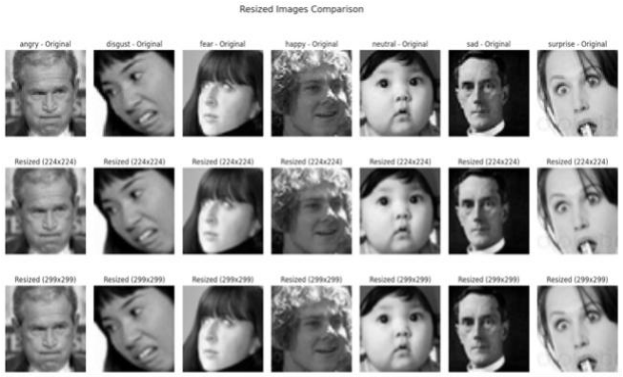

Fig. 7. Part resized images comparison of the dataset FER-2013.

Normalization: Normalization is a process in data preprocessing which is used to change the range of numerical data so that it is located in a specific cell, such as [0,1] or [1,1]. In image processing, normalization is a common practice. The essence of the method is some layer input data of the neural network that is preprocessed with zero mathematical expectations and unit variance with the intention of improving the stability and efficiency of the training process [4]. For FER tasks, normalization can give different features similar to ranges. Unnormalized data may lead to unstable gradient problems during model training. In deep learning models, normalized data may lead to a gradient that is too large or too small, thus affecting the learning effect of the model. The normalization in FER2013 dataset is shown in Fig. 8:

Fig. 8. Part normalized images comparison of the dataset FER-2013.

Feature Extraction: Feature extraction is an important factor in determining the recognition result. Some of the environmental and pose issues that need to be addressed in an image containing a complete face [61]. If the features are not good, even the best classifiers will not get the best results. In most cases, feature extraction produces a large number of features [35]. In this paper, we compare the performance of the model before and after the introduction of the attention mechanism, and for the FER task, features such as the eyes and mouth may be easier to identify than other parts. When assigning weights, we can selectively focus on certain parts of the input and directly ignore others to get the output we want most, which is called hard attention, or it can be integrated into the model in a differentiable way that allows end-to-end training of the entire network using standard techniques like gradient descent. For example, during the COVID-19 epidemic, many people are wearing masks. When the face is partially covered or interfered with [30], such as large pose changes, it can hinder useful feature extraction and greatly reduce the performance of FER predictions. The feature extraction in FER2013 dataset is shown in Fig. 9:

Fig. 9. Part feature extraction images comparison of the dataset FER-2013.

Data Augmentation: Data Augmentation is a technique to generate new, modified data points by transforming some original data columns. In the small-scale deep model data set, the deep model is redundant, complex, and easy to overfit. To solve the redundancy problem, data enhancement techniques were used to extend the original dataset [34]. Data enhancement will enable the model to introduce more variables during training, helping the model learn more generalized features and thus perform better on previously unseen data. In this paper, there is a data imbalance in the FER2013 dataset. The generation of data is enhanced using generative adversarial networks (GANs), the core idea of which is based on an adversarial process in which two networks - generator and discriminator - compete against each other [51]. For example, Hu et al (2019) used GAN to generate reference expressions and compared them with original expressions to generate differential features, avoiding interference of irrelevant information on expression recognition [31]. The generator takes random noise as input and outputs as real data as possible, such as high resolution images. In contrast, the discriminator takes real data or the data generated by the generator as input and outputs the probability of the data being true or false to distinguish the real data

generated by the generator from the false data. A generated discriminant representation can be obtained by separating and interpolating different expressions in a face image [62]. The learned representations not only generate more training samples of unpaired input images but also contribute to better FER performance. So, we involve generators and discriminators in GAN training, where generators try to generate more and more real data, and discriminators better distinguish between real and fake data. The most common form is the minimax game given by the following formula:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p\ data\ (x)}[\log D(x)] + \mathbb{E}_{z\sim pz(z)}[\log(1-D(G(z)))] \quad (1)$$

Where: $\mathbb{E}_{x\sim p_{data}(x)}[logD(x)]$represents the effect of real data x on discriminator D.

$\mathbb{E}_{z\sim pz(z)}[\log(1-D(G(z)))$ represents the effect of the data generated by the generator G on the discriminator D.

Cross-entropy loss effectively measures the disparity between the model's output probability distribution and the actual label distribution. The cross-entropy loss function usually has the following form:

$$L = \sum_{i=0}^{c} y_i \log(\hat{y}_l) \quad (2)$$

Where: the C is the number of classes (for FER-2013, it is 7, representing 7 basic emotions), y_i is the one-hot encoding of the real tag,$\hat{y}_l$ is the predicted probability for category i. Hence, a regularization term is added to encourage model to adopt smaller weights, thus reducing the model's complexity. L1-regularization: L1-regularization sums the absolute values of weights and tends to produce sparse weight matrices, which is conducive to feature selection.

$$L1 = \lambda\Sigma|\omega| \quad (3)$$

L2 Regularization: L2 regularization sums the squares of the weights, tends to uniformly assign errors, and is often used to prevent neural networks from overfitting.

$$L2 = \lambda\Sigma\omega^2 \quad (4)$$

Where ω indicates the weight of the model, λ is the regularization coefficient.

D. Phase 3: The Development of Algorithms and Models

Transfer learning is commonly training to assign specific weights to a pre-trained model and then train it with the dataset. Rajesh Kumar, C.G. Patil et al. and Sahoo et al. which usually perform better than statistical and traditional machine learning algorithms[32], [56], [54]. In addition, Lu et al., Martin Kampel et al. and Yichen Liu, compared to deep learning models, show superior performance [39], [45], [52]. Therefore, VGG19, ResNet50 and Inception V3 were used in this study to develop facial expression recognition models. Using the TensorFlow software library developed by the Google Brain team to develop the training model, the following sections outline the proposed architectures for the VGG19, ResNet50, and Inception models, including the model architecture with the addition of CBAM.

VGG19: The VGG model adopted in this study is VGG19, where 19 indicates that there are 19 learnable layers in the network, among which the convolutional layer uses multiple small dimensions (3×3). Use ReLU as the activation function between convolutional layers to increase nonlinearity. After each convolutional layer, use the maximum pooling layer (2×2). The network's top consists of three FC layers, with two layers comprising 4096 units each and the final FC layer matching the number of target categories. In this study, the target for the FER2013 dataset classification is seven; thus, the final fully connected layer is configured for seven categories. The output layer employs Softmax activation functions to convert the outputs into probability distributions as shown in Fig. 10:

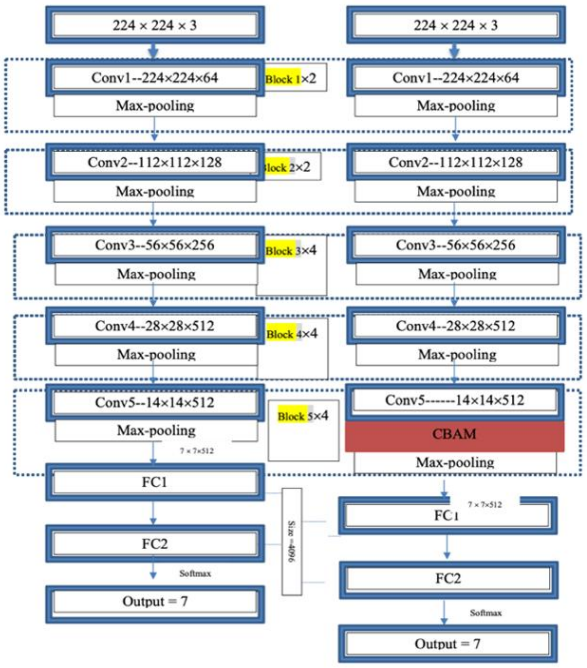

Fig. 10. Structure diagram of VGG19 and add CBAM.

ResNet50: ResNet model used in this study is ResNet50, which is a variant of the residual network, where 50 in ResNet50 refers to a weight layer in which the network contains 50 layers deep. Residual connections allow the inputs of the network to skip directly over one or more layers by adding outputs to the layer, which helps solve the problem of disappearing gradients in deeper networks. During training, this structure allows the gradient to flow directly over the jump connection, increasing the speed and effectiveness of training. The architecture starts with a 7x7 convolutional layer with a stride of 2, which is followed by a 3×3 max pooling layer with a stride of 2. The main component consists of several residual blocks, each containing multiple convolutional layers. In ResNet50, these residual blocks typically have three different configurations (different number of convolution layers and convolution kernel sizes) and are repeated multiple times. When the feature map's dimension needs modification, a convolution with a specific stride length is used to downsample the residual block input, and the number of channels is adjusted accordingly to match the output. Towards the end of the network, a global average pooling layer is employed instead of the conventional fully connected layer, thereby reducing the model's parameter count. As shown in Fig. 11:

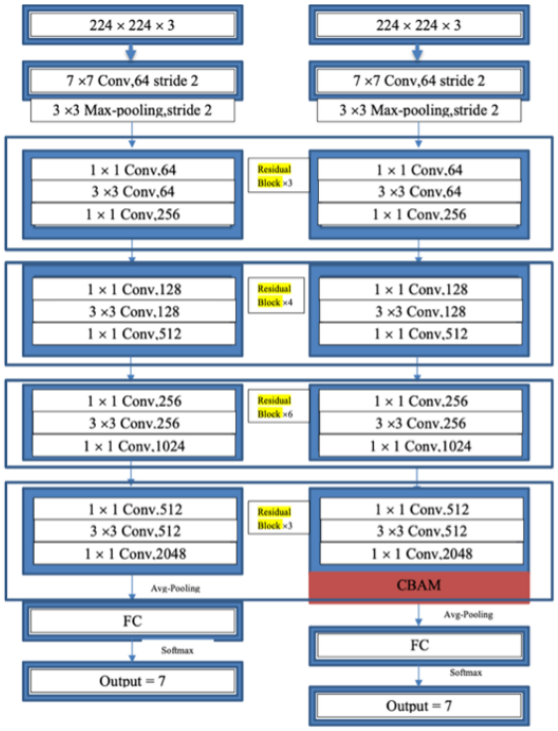

Fig. 11. Structure diagram of ResNet50 and add CBAM.

Inception V3: The Inception model used in this study is Inception V3. The core feature of Inception V3 is its "modular" design, which constructs the entire network through different modular building blocks. Inception V3 optimizes the original Inception module, for example, by introducing the concept of "factorization into smaller convolutions" to decompose large convolution kernels (e.g. 5×5) into smaller continuous convolution operations (e.g., two 3×3 convolution operations). An auxiliary classifier is added to the network as an output of the middle layer, which aids in gradient flow, provides additional regularization, and prevents overfitting during deep network training. At the end of the network, an overall average pooling layer takes the place of the conventional fully connected layer. The specific structure of Inception V3 used in this paper is depicted in Fig. 12:

Hyperparameter Tuning: Unlike the model training process, hyperparameter tuning involves finding the optimal set of hyperparameter values to maximize a performance metric on unseen data. This study uses the hyperparameter optimization library method and the grid search for hyperparameter tuning. In the hyperparameter optimization library, the scope and objective function of hyperparameter search is defined so that it accepts hyperparameters as input and returns the performance index of the model on the verification set (such as loss rate or accuracy rate) as the optimization target. By systematically searching the hyperparameter space, the hyperparameter combination that optimizes the model performance is found. For these models of deep learning, the hyperparameters tuned include learning _rate, batch size, regularization coefficient, activation function and epoch. Table IV are the details of each hyperparameter tuned in this study:

1) Learning rate: This determines how quickly the model moves in the gradient's direction or the number of steps taken. If the learning rate is excessively high, the optimizer may overshoot the minimum, preventing convergence. Conversely, a low learning rate makes the optimization process slow and may remain stuck at a suboptimal local minimum.

2) Batch size: This denotes the number of samples handled in a single forward and backward pass through the neural network. The batch size affects the optimization's efficiency and speed. A larger batch size improves memory utilization and makes the gradient descent direction more stable.

3) Regularization coefficient: This refers to the realization of a minimization strategy in which penalty terms are added to the empirical risk. Typically, it is a function of monotonically increasing model complexity. The more complex the model, the higher the penalty value.

4) Optimizer: An algorithm or method used in deep learning to update model parameters to minimize the loss function. The primary goal is to reduce the loss function as much as possible by adjusting the model parameters, ensuring the model fits the training set well and performs effectively on the test set.

5) Epoch: Defined as one complete pass of the dataset through the neural network. A single pass is insufficient; the dataset needs to be iterated multiple times to achieve convergence. An iterative method called gradient descent is employed to optimize learning. As the epoch count increases, the weights in the neural network are updated more frequently, transitioning from underfitting to overfitting. The optimal number of epochs varies and should be determined using validation sets or cross-validation.

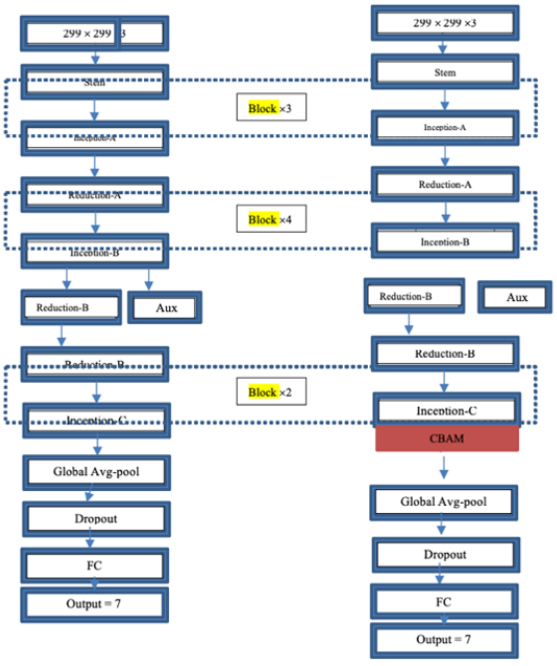

Fig. 12. Structure diagram of InceptionV3 and adds CBAM.

TABLE IV. DEFINITION OF VARIOUS HYPERPARAMETERS FOR DEEP LEARNING MODELS

Hyperparameter	Default Value	Tuning Value
Learning rate	0.01	0.001, 0.005
Batch size	30	60, 120
Regularization coefficient	0.001	0.01, 0.1
Optimizer	SGD	Adam RMSProp
epoch	50	100, 200

E. Phase 4: Evaluation

Key Metrics for Model Performance: A confusion matrix is a tool used to measure performance and is commonly utilized in supervised learning and classification problems. It helps visualize an algorithm's performance, particularly when dealing with two or more classes. The confusion matrix itself does not provide performance metrics like accuracy or precision, but they can be calculated from it. By examining the confusion matrix, we can further analyze the model's performance across different categories and identify its strengths and weaknesses. For example, some expression categories in the FER-2013 dataset may have more samples than others. Metrics such as confusion matrices and precision rates can help identify and address this imbalance, ensuring that the model has good recognition across all categories. The confusion matrix and accuracy rate provide a detailed perspective for diagnosing such issues. The confusion matrix can also reveal the model's tendency to misclassify one category into another, which can be very helpful for further tweaking and optimizing the model.

Statistical Test on Model Performance: After model evaluation, statistical tests are performed to determine which model shows better performance than other models that have been developed. Statistical testing is essential to determine whether a particular model is statistically significantly better than others, a process that can help us understand whether the differences in performance are significant enough to support or guard against assumptions about substantial differences between models. In this paper, the average performance of the best-performing VGG, ResNet and InceptionV3 models and the CBAM models were compared using the T-test, and the holdout method and cross-validation were employed. The statistical test in this study was performed using a 5×2 cross-validated paired t-test. One reason for conducting a 5x2 cross-validated paired T-test is its acceptable likelihood of type I errors. Because the comparison is carried out on the same set of data, the error caused by random variation of the data set is reduced. The 5x2 cross-validation paired T-test enables effective detection of performance differences even with small sample sizes. The 5x2 cross-validation provides ten independent performance evaluations that can be used to compare two models using statistical tests such as paired Ttests. Because the data is re-split each time, there is less correlation between the test results, reducing the estimates' variance. The paired T-test can also better estimate variation for 2-fold cross-validation since the training sets do not overlap, compared to 10-fold cross-validation.

IV. RESULTS AND DISCUSSION

In this section, compare the performance metrics of the developed VGG19, ResNet50, and InceptionV3, with and without the CBAM module. Model efficacy was gauged using traintest segmentation (TTS) and cross-validation (CV) techniques. The dataset was partitioned with a training-to-test ratio of 80:20, and a 10-fold cross-validation was employed. Hyperparameter tuning was executed on six models, each with five distinct hyperparameters, each tested at three varying levels. Consequently, this culminated in 243 potential hyperparameter configurations for each algorithm during the tuning process.

TABLE V. HYPERPARAMETER COMBINATIONS WITH THE HIGHEST TEST ACCURACY FOR VGG19 AND VGG19-CBAM WITH TRAIN-TEST SPLIT

Model	Hyperparameter	Default	Best Value
VGG19	Epoch	50	200
	Batch Size	512	128
	Learning Rate	0.01	0.001
	Regularization Coefficient	0.01	0.001
	Optimizer	Adam	Adam
VGG19-CBAM	Epoch	50	200
	Batch Size	512	128
	Learning Rate	0.01	0.001
	Regularization Coefficient	0.01	0.001
	Optimizer	Adam	Adam

TABLE VI. ACCURACY COMPARISON BETWEEN THE DEFAULT AND OPTIMAL HYPERPARAMETER CONFIGURATION OF VGG19 AND VGG19-CBAM

Model	Hyperparameter Configuration	Test Accuracy
VGG19	Default	0.7040
	Best	0.7170
VGG19-CBAM	Default	0.7104
	Best	0.7190

The hyperparameter adjustment results of the VGG19 and added CBAM models are visualized in Fig. 12 using the mesh search method based on the train-test split. Under the training-test segmentation, the test accuracy ranges from 0.7040 to 0.7104. Table V shows the best hyperparameter configuration of VGG19 based on the highest test accuracy in the case of training test segmentation. The highest test accuracy of training-test segmentation is 0.719, as shown in Table VI. This suggests that both models may have overfitted training datasets and cannot properly generalize to previously unknown data.

The hyperparameter adjustment results of the ResNet50 and added CBAM models are visualized in Fig. 8 using the mesh search method based on the train-test split. The test accuracy ranges from 0.7033 to 0.7124. Table VII shows the best hyperparameter configuration of the ResNet50 model based on the highest test accuracy in the case of training test segmentation. The highest test accuracy of training-test segmentation is 0.724, as shown in Table VIII.

TABLE VII. HYPERPARAMETER COMBINATIONS WITH THE HIGHEST TEST ACCURACY FOR RESNET50 AND RESNET50-CBAM WITH TRAIN-TEST SPLIT

Model	Hyperparameter	Default	Best Value
ResNet50	Epoch	50	200
	Batch Size	512	128
	Learning Rate	0.01	0.01
	Regularization Coefficient	0.01	0.001
	Optimizer	Adam	Sgd
ResNet50-CBAM	Epoch	50	200
	Batch Size	512	128
	Learning Rate	0.01	0.01
	Regularization Coefficient	0.01	0.001
	Optimizer	Adam	Sgd

TABLE VIII. ACCURACY COMPARISON BETWEEN THE DEFAULT AND OPTIMAL HYPERPARAMETER CONFIGURATION OF RESNET50 AND RESNET50-CBAM

Model	Hyperparameter Configuration	Test Accuracy
ResNet50	Default	0.7033
	Best	0.7150
ResNet50-CBAM	Default	0.7124
	Best	0.7240

TABLE IX. HYPERPARAMETER COMBINATIONS WITH THE HIGHEST TEST ACCURACY FOR INCEPTION V3 AND INCEPTION V3-CBAM WITH TRAIN-TEST SPLIT

Model	Hyperparameter	Default	Best Value
Inception V3	Epoch	50	200
	Batch Size	512	128
	Learning Rate	0.01	0.05
	Regularization Coefficient	0.01	0.001
	Optimizer	Adam	Sgd
Inception V3-CBAM	Epoch	50	200
	Batch Size	512	128
	Learning Rate	0.01	0.05
	Regularization Coefficient	0.01	0.001
	Optimizer	Adam	Sgd

The hyperparameter adjustment results of the Inception V3 and added CBAM models are visualized in Fig. 9 using the mesh search method based on the train-test split. The test accuracy ranges from 0.7002 to 0.7070. Table IX shows the best hyperparameter configuration of the ResNet50 model based on the highest test accuracy in the case of training test segmentation. The highest test accuracy of training-test segmentation is 0.711, as shown in Table X.

The analysis indicates that hyperparameter optimization significantly enhances the performance of deep learning models. The incorporation of CBAM further boosts the models' performance, with ResNet50-CBAM showing the most substantial improvement accuracy rate at 0.724. These findings highlight the importance of both hyperparameter tuning and advanced architectural modifications in achieving optimal model performance for facial expression recognition tasks.

Table XI shows the accuracy comparison of multiple deep learning models on a specific task, including the results of other researchers, providing rich information for analyzing the performance differences of the models. First, AlexNet [60] is an earlier deep-learning model with a relatively low performance on this task, with an accuracy of 0.643. Subsequently, the accuracy rate of VGG16 [24] was 0.65, which was slightly improved. The VGG16+SENet [24] combined with SENet module reaches 0.668, indicating that the addition of SENet module can improve the model performance to a certain extent. 10layer CNN [58] has an accuracy of 0.683, a significant improvement over than previous models. MobileNet [34], which combines the generation of an adversarial network and attention mechanism, reached 0.70, and its performance was further improved. The accuracy of the Priority Ensemble CNN [39] is 0.7052, which is further improved by the integrated approach. FERW [29], a model designed specifically for a specific task, achieved 0.71 and also performed very well. The Ensemble CNN [32] has an accuracy of 0.7127, which is further improved by integrating multiple CNN models. The VGG [45] achieved 0.714, an improvement over the traditional VGG16. VGG19 has an accuracy of 0.717, which is deeper and better than VGG16.

TABLE X. ACCURACY COMPARISON BETWEEN THE DEFAULT AND OPTIMAL HYPERPARAMETER CONFIGURATION OF INCEPTION V3 AND INCEPTION V3-CBAM

Model	Hyperparameter Configuration	Test Accuracy
Inception V3	Default	0.7002
	Best	0.7040
Inception V3-CBAM	Default	0.7072
	Best	0.7110

TABLE XI. THE SPECIFIC PERFORMANCE RESULTS OF THIS RESEARCH MODELS WITH THOSE OF PREVIOUS RESEARCH MODELS

Model	Accuracy
AlexNet[60]	0.643
VGG16[24]	0.650
VGG16+SENet[24]	0.668
10-layer CNN[58]	0.683
SRGAN + AGN - MobileNet[34]	0.700
Priority Ensemble CNN[39]	0.7052
FERW[29]	0.710
Ensemble CNN[32]	0.7127
VGG[45]	0.714
VGG19	0.717
ResNet50	0.715
Inception V3	0.704
VGG19-CBAM	0.719
ResNet50-CBAM	0.724
InceptionV3-CBAM	0.711

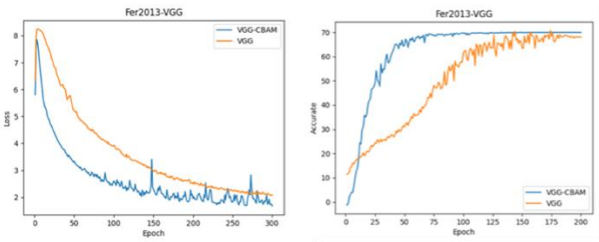

Fig. 13. Structure diagram of VGG19 and adds CBAM.

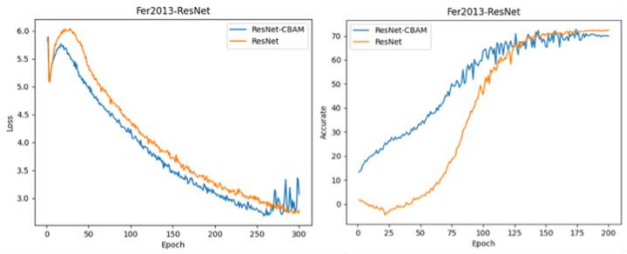

Fig. 14. Structure diagram of ResNet50 and adds CBAM.

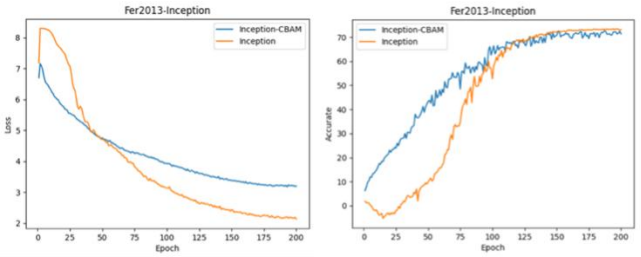

Fig. 15. Structure diagram of InceptionV3 and adds CBAM.

Fig. 13-15 shows the training verification line charts of VGG19, ResNet50 and InceptionV3, respectively, in which we can clearly observe their obvious trends. Detailed comparison: For Learning Speed, ResNet50 demonstrates the fastest learning speed among the three models, achieving significant reductions in loss and increases in accuracy in fewer epochs. VGG19 and InceptionV3 show slower but steady improvements. For Performance Stability, VGG19 shows the most stable performance with less fluctuation in its loss and accuracy curves. In contrast, ResNet50, while faster in learning, exhibits more fluctuations, indicating a more dynamic learning process with potential overfitting or regularization adjustments. InceptionV3 shows consistent improvement but at a slower rate.

The VGG series and the ResNet series exhibit relatively superior performance when handling the categories of "anger", "neutral", and "surprise", yet encounter substantial challenges when dealing with the complex categories of "fear" and "disgust". After the introduction of the CBAM, both demonstrate enhancements in the majority of categories, particularly in the manifestations of the "happy" and "surprised" categories. The Inception series model achieves the optimal classification effect for "happy" and "surprised", and the introduction of the CBAM attention mechanism further elevates the accuracy rates of these categories. Nevertheless, this model still presents considerable classification errors in the "disgust" and "fear" categories. Through the introduction of CBAM, the classification accuracy rates of all models have improved in most emotional categories, especially in the "happy" and "surprised" categories. However, in some difficult-to-distinguish categories (such as "disgust" and "fear"), the improvement effect of CBAM is limited. Among the six models, the ResNet-CBAM and Inception-CBAM models display the most outstanding overall performance, particularly in the classification performance of complex emotional categories. Nevertheless, all models still encounter notable classification difficulties in the emotions of "disgust" and "fear", indicating that the features of these emotional categories in the Fer2013 dataset might be challenging to differentiate. The confusion matrix generated by the six models trained in this study is shown in Fig. 16.

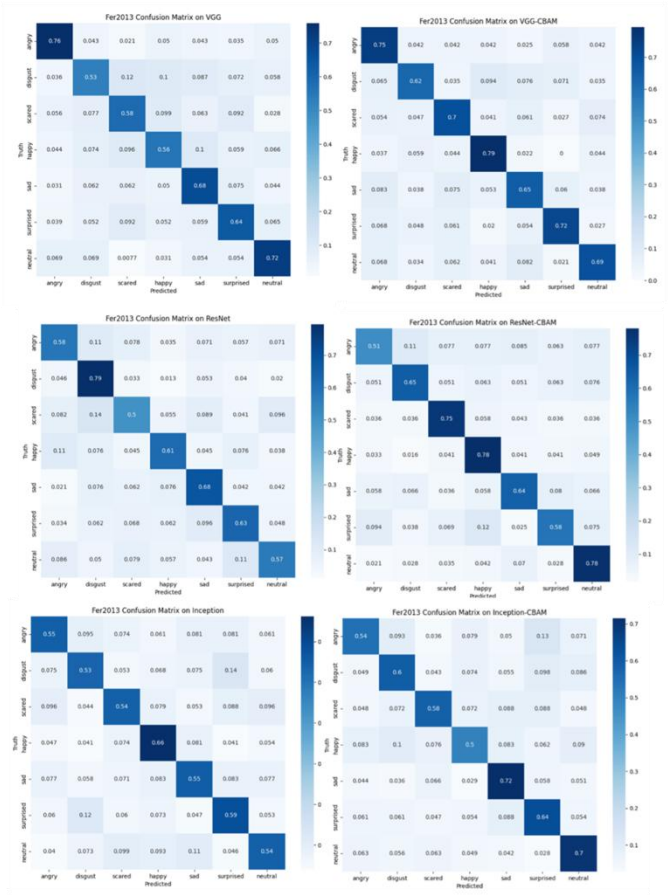

Fig. 16. The confusion matrix generated by the six models.

TABLE XII. HYPERPARAMETER COMBINATIONS WITH THE HIGHEST TEST ACCURACY FOR INCEPTION V3 AND INCEPTION V3-CBAM WITH TRAIN-TEST SPLIT

Model	t-value	p-value
VGG19 default vs. VGG19 optimize	-2.828	0.002
ResNet50 default vs. ResNet50 optimize	-3.238	0.032
Inception default vs. Inception optimize	-4.268	0.013
VGG19-CBAM default vs. VGG19-CBAM optimize	-5.266	0.006
ResNet50-CBAM default vs. ResNet50-CBAM optimize	-4.603	0.010
Inception V3-CBAM default vs. Inception V3-CBAM optimize	-5.411	0.006

TABLE XIII. RESULTS OF 5X2 CROSS-VALIDATED PAIRED T-TEST FOR MODEL PERFORMANCE COMPARISON OF VGG19, RESNET50, AND INCEPTION V3 AND ADDED CBAM MODELS BEFORE AND AFTER THE ADDITION OF CBAM COMPARED AFTER HYPERPARAMETERTUNING WITH EACH OTHER

Model	t-value	p-value
VGG19 optimize vs VGG19-CBAM optimize	-4.1	0.015
ResNet50 optimize vs ResNet50-CBAM optimize	-4.297	0.013
Inception V3 optimize vs Inception V3-CBAM optimize	-4.347	0.013
VGG19-CBAM optimize vs ResNet50-CBAM optimize	-3.033	0.039
VGG19-CBAM optimize vs Inception V3-CBAM optimize	8.421	0.001
Inception V3-CBAM optimize vs ResNet50-CBAM optimize	-12.858	0.002

To assess the statistical significance of changes in the accuracy performance of the developed models, a 5x2 cross-validated paired t-test was conducted. Among the models, ResNet50-CBAM demonstrated the highest performance during the test phase. To evaluate whether ResNet50-CBAM significantly outperformed the other models, the test compared the best performance of ResNet50-CBAM with VGG19-CBAM and Inception-CBAM under both default and optimized configurations. The algorithms were tested using the same training-test segmentation (TTS) and cross-validation (CV) hyperparameter configurations.

The t-test results summarized in Table XII test the null hypothesis that there is no significant difference in the accuracy of the three models of CBAM under default and tuned hyperparameters. Since all P-values are less than 0.05, the null hypothesis is rejected with 95\% confidence. A negative t value indicates that the default configured model performs worse on average compared to the optimized hyperparameters. The T-test results summarized in Table XIII verify the null hypothesis. After hyperparameter fine-tuning, there is no significant difference in the accuracy of each model, and all the P-values are less than 0.05, so the null hypothesis is rejected. The T-values show that there are differences between each other, and the rows of ResNet50-CBAM are significantly ahead of other models. Therefore, in summary, compared with VGG19-CBAM and Inception V3-CBAM models, the performance of ResNet50-CBAM model is statistically significant, which highlights the influence of CBAM and hyperparameter tuning on improving model accuracy.

V. CONCLUSION

This study set out to predict facial expressions and analyze the impact of the Convolutional Block Attention Module (CBAM) on the performance of deep learning models. The research successfully achieved three main objectives. The proposed method leverages the residual connections and hierarchical feature extraction capability of ResNet50, augmented with CBAM to enhance spatial and channel-wise focus. This combination improves the ability to distinguish subtle expressions, such as ‘neutral’ and ‘sad,’ while being robust to occlusions and lighting variations.”

The first objective was to identify the optimal deep learning model for classifying facial expressions into seven categories within the FER2013 dataset. The study trained three key models: VGG19, ResNet50, and Inception V3. VGG19 emerged as the top performer with a test accuracy of 0.717, slightly surpassing ResNet50 and Inception V3, which achieved 0.715 and 0.704, respectively. Despite these close results, VGG19 demonstrated a marginal but consistent advantage over the other models.

The second objective focused on assessing the impact of CBAM on these models. The results showed that integrating CBAM led to notable improvements across all three models, particularly ResNet50. The ResNet50-CBAM model achieved the highest accuracy of 0.7124, outperforming both VGG19-CBAM and Inception-CBAM, which reached 0.7104 and 0.7072, respectively. This demonstrates CBAM’s ability to enhance feature extraction by enabling models to dynamically adjust weights based on channel and spatial positions, thus improving their performance, particularly in deeper networks like ResNet50.

The third objective was to optimize model performance through hyperparameter tuning. The grid search method was employed to find the best hyperparameter combinations, leading to significant accuracy improvements. Post-tuning, the ResNet50-CBAM model achieved a test accuracy of 0.724, with VGG19-CBAM and Inception V3-CBAM following at 0.719 and 0.711, respectively. The 5x2 cross-validated paired t-test results confirmed that these enhancements were statistically significant, with p-values below 0.05 for all models. The findings underscore the critical role of hyperparameter tuning in maximizing model performance and demonstrate that ResNet50, when combined with CBAM and optimized hyperparameters, outperforms both VGG19 and Inception V3. In conclusion, ResNet50-CBAM emerged as the best-performing model in this study, demonstrating superior accuracy and effectiveness in facial expression recognition tasks. This study highlights the critical importance of integrating CBAM and optimizing hyperparameters to maximize model performance. The findings emphasize that advanced feature extraction techniques and careful model tuning can significantly enhance the accuracy and reliability of deep learning models, with ResNet50-CBAM setting the benchmark for excellence in this domain.

VI. FUTURE WORK AND LIMITATIONS

The limitations of this study relate to the hyperparameters of data sets, algorithms, and classification model development. Only the FER2013 dataset was trained in this study. Due to the early presentation time of FER2013, relatively low image quality, and certain noise and blurring, the faces in the data set are mainly positive faces, and lack of diverse images such as side faces and occluded images, which may lead to inadequate adaptation of the trained model in practical applications. In addition, the FER2013 only contains the basic expression types, and the number of samples for each expression type is not balanced, and the overall sample size is small, which leads to the problem of overfitting in the process of training and testing, affecting the generalization ability.

Since the data collected in the dataset is mainly facial expression image data from 2013, and higher quality datasets have also emerged due to improvements in image acquisition hardware and diversity, such as AffectNet, future improvements could focus only on the most reliable facial image datasets to

mitigate the effects of noise during model development. The attention mechanism is a feasible choice to optimize the model performance further, not just for facial expression recognition but for other tasks as well. For example, consider an image with a cat and a table in the image description generation task. The attention mechanism helps the model focus on two important areas of the image: the cat and the table. When the model generates a description, CBAM can automatically capture the importance of cats and tables and automatically assign weights, making the model more focused on this area, which allows the model to more accurately describe the content in the image, and the generated description is more interpretable.

In the FER task, the deployment of the FER platform faces great challenges due to the different channels of image acquisition and model training. Due to the early presentation of the FER2013 dataset, the image quality is low, and there are unbalanced noise, blurring, and other conditions. In the future, we can choose newer, higher quality facial expression recognition datasets such as AffectNet, RAF-DB, etc., which have higher resolution and diverse samples, or apply more advanced image enhancement and preprocessing technologies such as denoising, deblur, and contrast enhancement. In order to improve image quality, it is possible to recognize facial expressions more accurately, helping to improve the human interaction experience. Secondly, there may be bias in expression recognition of different ages and races. In future studies, we can pay special attention to and quantify the performance differences of different groups to ensure the fairness of the model for different groups. Design hierarchical or adaptive models that can adjust parameters or weights based on input characteristics, such as age, gender, and race, to improve the accuracy of identifying different populations.

In this study, only three model architectures were tested, and feature extraction only compared attention mechanisms. In future research, we can try more different types of deep learning models, such as recurrent neural networks (RNN), graph neural networks (GNN), and different model architectures, such as ResNet or Inception. In addition to attention mechanisms, other feature extraction methods can also be tried, such as multi-scale feature extraction, emotional feature extraction, Vision Transformer (ViT), etc. By comparing the deep learning model of facial expression recognition, this study can identify facial expressions more accurately, which helps to improve the human-computer interaction experience, provide users with more intelligent and convenient services and experiences, and solve practical problems. In addition, this study can be used as a basic framework for developing face recognition based on deep learning models.

ACKNOWLEDGMENT

Funding: This research was funded by the Universiti Kebangsaan Malaysia (Grant code:FRGS/1/2024/ICT06/ UKM/02/3).

Authors' Contribution: The authors confirm contribution to the paper as follows: study conception and design: Liu Luan Xiang Wei, Nor Samsiah Sani; data collection: Liu Luan Xiang Wei; analysis and interpretation of results: Liu Luan Xiang Wei, Nor Samsiah Sani; draft manuscript preparation: Liu Luan Xiang Wei, Nor Samsiah Sani. All authors reviewed the results and approved the final version of the manuscript.

Conflict of Interest The corresponding author states that there is no conflict of interest on behalf of all authors.

Data Availability The data used in this study are available from the following resources in the public domain: https://www.kaggle.com/datasets/msambare/fer2013

REFERENCES

[1] Dobrojevic, M., Zivkovic, M., Chhabra, A., Sani, N. S., Bacanin, N., & Amin, M. M. (2023). Addressing internet of things security by enhanced sine cosine metaheuristics tuned hybrid machine learning model and results interpretation based on shap approach. PeerJ Computer Science, 9, e1405.https://doi.org/10.1109/CSASE48920.2020.9142065.

[2] Suwadi, N. A., Derbali, M., Sani, N. S., Lam, M. C., Arshad, H., Khan, I., & Kim, K. I. (2022). An optimized approach for predicting water quality features based on machine learning. Wireless Communications and Mobile Computing, 2022(1), 3397972.https://doi.org/10.1109/IPRIA59240.2023.10147196.

[3] Othman, Z. A., Bakar, A. A., Sani, N. S., & Sallim, J. (2020). Household overspending model amongst B40, M40 and T20 using classification algorithm. International Journal of Advanced Computer Science and Applications, 11(7).

[4] Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering analysis for classifying student academic performance in higher education. Applied Sciences, 12(19), 9467.

[5] Holliday, J., Sani, N., & Willett, P. (2018). Ligand-based virtual screening using a genetic algorithm with data fusion. Match: Communications in Mathematical and in Computer Chemistry, 80(3).https://doi.org/10.1109/ICISS50791.2020.9307567.

[6] Bassel, A., Abdulkareem, A. B., Alyasseri, Z. A. A., Sani, N. S., & Mohammed, H. J. (2022). Automatic malignant and benign skin cancer classification using a hybrid deep learning approach. Diagnostics, 12(10), 2472.

[7] Abdul-Hadi, M.H., Waleed, J.: Human Speech and Facial Emotion Recognition Technique Using SVM. In: 2020 International Conference on Computer Science and Software Engineering (CSASE). pp. 191–196 IEEE, Duhok, Iraq (2020). https://doi.org/10.1109/CSASE48920.2020.9142065.

[8] Afshar, E. et al.: Facial Expression Recognition using Spatial Feature Extraction and Ensemble Deep Networks. In: 2023 6th International Conference on Pattern Recognition and Image Analysis (IPRIA). pp. 1–6 IEEE, Qom, Iran, Islamic Republic of (2023). https://doi.org/10.1109/IPRIA59240.2023.10147196.

[9] Agrawal, I. et al.: Emotion Recognition from Facial Expression using CNN. In: 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC). pp. 01–06 IEEE, Bangalore, India (2021). https://doi.org/10.1109/R10-HTC53172.2021.9641578.

[10] Alexeevskaya, Y.A. et al.: Recognizing Human Emotions Using a Convolutional Neural Network. In: 2022 4th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE). pp. 1–6 IEEE, Moscow, Russian Federation (2022). https://doi.org/10.1109/REEPE53907.2022.9731391.

[11] Andrian, R., Supangkat, S.H.: Comparative Analysis of Deep Convolutional Neural Networks Architecture in Facial Expression Recognition: A Survey. In: 2020 International Conference on ICT for Smart Society (ICISS). pp. 1–6 IEEE, Bandung, Indonesia (2020). https://doi.org/10.1109/ICISS50791.2020.9307567.

[12] Avanija, J. et al.: Facial Expression Recognition using Convolutional Neural Network. In: 2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR). pp. 1–7 IEEE, Hyderabad, India (2022). https://doi.org/10.1109/ICAITPR51569.2022.9844221.

[13] Avula, H. et al.: CNN based Recognition of Emotion and Speech from Gestures and Facial Expressions. In: 2022 6th International Conference on Electronics, Communication and Aerospace Technology. pp. 1360–

1365 IEEE, Coimbatore, India (2022). https://doi.org/10.1109/ICECA55336.2022.10009316.

[14] Azimi, M.: Effects of Facial Mood Expressions on Face Biometric Recognition System's Reliability. In: 2018 1st International Conference on Advanced Research in Engineering Sciences (ARES). pp. 1–5 IEEE, Dubai, United Arab Emirates (2018). https://doi.org/10.1109/ARESX.2018.8723292.

[15] Cha, H.-S. et al.: Real-Time Recognition of Facial Expressions Using Facial Electromyograms Recorded Around the Eyes for Social Virtual Reality Applications. IEEE Access. 8, 62065–62075 (2020). https://doi.org/10.1109/ACCESS.2020.2983608.

[16] Chen, H. et al.: Facial Expression Recognition and Positive Emotion Incentive System for Human-Robot Interaction. In: 2018 13th World Congress on Intelligent Control and Automation (WCICA). pp. 407–412 IEEE, Changsha, China (2018). https://doi.org/10.1109/WCICA.2018.8630711.

[17] Chen, X. et al.: DD-CISENet: Dual-Domain Cross-Iteration Squeeze and Excitation Network for Accelerated MRI Reconstruction, http://arxiv.org/abs/2305.00088, (2023).

[18] Chen, Y. et al.: Facial Motion Prior Networks for Facial Expression Recognition. In: 2019 IEEE Visual Communications and Image Processing (VCIP). pp. 1–4 IEEE, Sydney, Australia (2019). https://doi.org/10.1109/VCIP47243.2019.8965826.

[19] Chuanjie, Z., Changming, Z.: Facial Expression Recognition Integrating Multiple CNN Models. In: 2020 IEEE 6th International Conference on Computer and Communications (ICCC). pp. 1410–1414 IEEE, Chengdu, China (2020). https://doi.org/10.1109/ICCC51575.2020.9345285.

[20] Das, A., N, N.: Facial Expression Recognition System with Local Binary Features of Neural Network. In: 2023 International Conference on Data Science and Network Security (ICDSNS). pp. 1–5 IEEE, Tiptur, India (2023). https://doi.org/10.1109/ICDSNS58469.2023.10244983.

[21] Dong, J. et al.: Segmentation Algorithm of Magnetic Resonance Imaging Glioma under Fully Convolutional Densely Connected Convolutional Networks. Stem Cells International. 2022, 1–9 (2022). https://doi.org/10.1155/2022/8619690.

[22] Dwijayanti, S. et al.: Facial Expression Recognition and Face Recognition Using a Convolutional Neural Network. In: 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). pp. 621–626 IEEE, Yogyakarta, Indonesia (2020). https://doi.org/10.1109/ISRITI51436.2020.9315513.

[23] Dy, M.L.I.C. et al.: Multimodal Emotion Recognition Using a Spontaneous Filipino Emotion Database. In: 2010 3rd International Conference on Human-Centric Computing. pp. 1–5 IEEE, Cebu, Philippines (2010). https://doi.org/10.1109/HUMANCOM.2010.5563314.

[24] Ekman, P., Friesen, W.V.: Facial Action Coding System, http://doi.apa.org/getdoi.cfm?doi=10.1037/t27734-000, (2019). https://doi.org/10.1037/t27734-000.

[25] Fu, S.: Research on Facial Expression Recognition Based on Deep Learning Method. In: 2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT). pp. 818–821 IEEE, Dali, China (2022). https://doi.org/10.1109/ICCASIT55263.2022.9987082.

[26] Gaman, Y. et al.: Adaptive Learning Method in Facial Expression Recognition Model Using Fuzzy-ART. In: 2019 IEEE 1st Global Conference on Life Sciences and Technologies (LifeTech). pp. 85–86 IEEE, Osaka, Japan (2019). https://doi.org/10.1109/LifeTech.2019.8884040.

[27] Ganatra, N. et al.: Classification of Facial Expression for Emotion Recognition using Convolutional Neural Network. In: 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT). pp. 1–5 IEEE, Trichy, India (2022). https://doi.org/10.1109/ICEEICT53079.2022.9768508.

[28] Gopalan, N.P. et al.: Facial Expression Recognition Using Geometric Landmark Points and Convolutional Neural Networks. In: 2018 International Conference on Inventive Research in Computing Applications (ICIRCA). pp. 1149–1153 IEEE, Coimbatore (2018). https://doi.org/10.1109/ICIRCA.2018.8597226.

[29] Grover, R., Bansal, S.: Facial Expression Recognition: Deep Survey, Progression and Future Perspective. In: 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT). pp. 111–117 IEEE, Gharuan, India (2023). https://doi.org/10.1109/InCACCT57535.2023.10141843.

[30] Han, J., Gopalakrishnan, A.K.: Real-time Evaluation of Food Acceptance From Facial Expressions Based on Exponential Decay. In: 2023 15th International Conference on Knowledge and Smart Technology (KST). pp. 1–5 IEEE, Phuket, Thailand (2023). https://doi.org/10.1109/KST57286.2023.10086796.

[31] Hu, S. et al.: Natural Scene Facial Expression Recognition based on Differential Features. In: 2019 Chinese Automation Congress (CAC). pp. 2840–2844 IEEE, Hangzhou, China (2019). https://doi.org/10.1109/CAC48633.2019.8997280.

[32] Imamura, N. et al.: Extraction of Useful Features from Neural Network for Facial Expression Recognition. In: 2019 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). pp. 221–226 IEEE, Toyama, Japan (2019). https://doi.org/10.1109/SNPD.2019.8935652.

[33] ˙Incetas¸, M.O. et al.: A novel image Denoising approach using super resolution densely connected convolutional networks. Multimed Tools Appl. 81, 23, 33291–33309 (2022). https://doi.org/10.1007/s11042-02213096-4.

[34] Islam, B. et al.: Human Facial Expression Recognition System Using Artificial Neural Network Classification of Gabor Feature Based Facial Expression Information. In: 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT). pp. 364–368 IEEE, Dhaka, Bangladesh (2018). https://doi.org/10.1109/CEEICT.2018.8628050.

[35] Jia, C. et al.: Facial expression recognition based on the ensemble learning of CNNs. In: 2020 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). pp. 1–5 IEEE, Macau, China (2020). https://doi.org/10.1109/ICSPCC50002.2020.9259543.

[36] Jin, R. et al.: AVT: Au-Assisted Visual Transformer for Facial Expression Recognition. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 2661–2665 IEEE, Bordeaux, France (2022). https://doi.org/10.1109/ICIP46576.2022.9897960.

[37] Jin, X. et al.: The Research and Improvement of Facial Expression Recognition Algorithm Based on Convolutional Neural Network. In: 2023 26th ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter). pp. 166–170 IEEE, Taiyuan, Taiwan (2023). https://doi.org/10.1109/SNPD-Winter57765.2023.10224044.

[38] Joseph, J.L., Mathew, S.P.: Facial Expression Recognition for the Blind Using Deep Learning. In: 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON). pp. 1–5 IEEE, Kuala Lumpur, Malaysia (2021). https://doi.org/10.1109/GUCON50781.2021.9574035.

[39] Ju, L., Zhao, X.: Mask-Based Attention Parallel Network for in-theWild Facial Expression Recognition. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2410–2414 IEEE, Singapore, Singapore (2022). https://doi.org/10.1109/ICASSP43922.2022.9747717.

[40] Kim, S., Kim, H.: Deep Explanation Model for Facial Expression Recognition Through Facial Action Coding Unit. In: 2019 IEEE International Conference on Big Data and Smart Computing (BigComp). pp. 1–4 IEEE, Kyoto, Japan (2019). https://doi.org/10.1109/BIGCOMP.2019.8679370.

[41] Kumar, R.: A Deep Learning Approach To Recognizing Emotions Through Facial Expressions. In: 2023 Global Conference on Wireless and Optical Technologies (GCWOT). pp. 1–5 IEEE, Malaga, Spain (2023). https://doi.org/10.1109/GCWOT57803.2023.10064654.

[42] Lee, G.-C. et al.: Ensemble Algorithm of Convolution Neural Networks for Enhancing Facial Expression Recognition. In: 2022 IEEE 5th International Conference on Knowledge Innovation and Invention (ICKII). pp. 111–115 IEEE, Hualien, Taiwan (2022). https://doi.org/10.1109/ICKII55100.2022.9983573.

[43] Li, H. et al.: Differential Diagnosis for Pancreatic Cysts in CT Scans Using Densely-Connected Convolutional Networks, http://arxiv.org/abs/1806.01023, (2018).

[44] Li, Y. et al.: Deep Learning for Micro-Expression Recognition: A Survey. IEEE Trans. Affective Comput. 13, 4, 2028–2046 (2022). https://doi.org/10.1109/TAFFC.2022.3205170.

[45] Liliana, D.Y. et al.: Geometric Facial Components Feature Extraction for Facial Expression Recognition. In: 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS). pp. 391–396 IEEE, Yogyakarta (2018). https://doi.org/10.1109/ICACSIS.2018.8618248.

[46] Liu, H. et al.: Adaptive Multilayer Perceptual Attention Network for Facial Expression Recognition. IEEE Trans. Circuits Syst. Video Technol. 32, 9, 6253–6266 (2022). https://doi.org/10.1109/TCSVT.2022.3165321.

[47] Liu, K.-C. et al.: Facial Expression Recognition Using Merged Convolution Neural Network. In: 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE). pp. 296–298 IEEE, Osaka, Japan (2019). https://doi.org/10.1109/GCCE46687.2019.9015479.

[48] Liu, W., Fang, J.: Facial Expression Recognition Method Based on Cascade Convolution Neural Network. In: 2021 International Wireless Communications and Mobile Computing (IWCMC). pp. 1012–1015 IEEE, Harbin City, China (2021). https://doi.org/10.1109/IWCMC51323.2021.9498621.

[49] Liu, Y.: Facial Expression Recognition Model Based on Improved VGGNet. In: 2023 4th International Conference on Electronic Communication and Artificial Intelligence (ICECAI). pp. 404–408 IEEE, Guangzhou, China (2023). https://doi.org/10.1109/ICECAI58670.2023.10177007.

[50] Lu, H.: AF-Transformer: Attention Fusion Transformer for Facial Expression Recognition. In: 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA). pp. 939–942 IEEE, Changchun, China (2022). https://doi.org/10.1109/CVIDLICCEA56201.2022.9824452.

[51] Luo, Y. et al.: Design of Facial Expression Recognition Algorithm Based on CNN Model. In: 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA). pp. 580–583 IEEE, Shenyang, China (2023). https://doi.org/10.1109/ICPECA56706.2023.10075779.

[52] Meena, G. et al.: Sentiment analysis on images using convolutional neural networks based Inception-V3 transfer learning approach. International Journal of Information Management Data Insights. 3, 1, 100174 (2023). https://doi.org/10.1016/j.jjimei.2023.100174.

[53] Muhamad, M. et al.: Recognizing Human Emotion Using Computer Vision. In: 2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS). pp. 1–4 IEEE, IPOH, Malaysia (2021). https://doi.org/10.1109/AiDAS53897.2021.9574411.

[54] Munasinghe, M.I.N.P.: Facial Expression Recognition Using Facial Landmarks and Random Forest Classifier. In: 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS). pp. 423–427 IEEE, Singapore (2018). https://doi.org/10.1109/ICIS.2018.8466510.

[55] N, M.: Squeeze aggregated excitation network, http://arxiv.org/abs/2308.13343, (2023).

[56] NV, M.: Variations of Squeeze and Excitation networks, http://arxiv.org/abs/2304.06502, (2023).

[57] Nwosu, L. et al.: Deep Convolutional Neural Network for Facial Expression Recognition Using Facial Parts. In: 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech). pp. 1318–1321 IEEE, Orlando, FL (2017). https://doi.org/10.1109/DASCPICom-DataCom-CyberSciTec.2017.213.

[58] Poux, D. et al.: Dynamic Facial Expression Recognition Under Partial Occlusion With Optical Flow Reconstruction. IEEE Trans. on Image Process. 31, 446–457 (2022). https://doi.org/10.1109/TIP.2021.3129120.

[59] Shiomi, T. et al.: Facial Expression Intensity Estimation Considering Change Characteristic of Facial Feature Values for Each Facial Expression. In: 2022 23rd ACIS International Summer Virtual Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Summer). pp. 15–21 IEEE, Kyoto City, Japan (2022). https://doi.org/10.1109/SNPDSummer57817.2022.00012.

[60] Taini, M. et al.: Facial expression recognition from near-infrared video sequences. In: 2008 19th International Conference on Pattern Recognition. pp. 1–4 IEEE, Tampa, FL, USA (2008). https://doi.org/10.1109/ICPR.2008.4761697.

[61] Tiwari, T. et al.: Facial Expression Recognition Using Keras in Machine Learning. In: 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N). pp. 466–471 IEEE, Greater Noida, India (2021). https://doi.org/10.1109/ICAC3N53548.2021.9725756.

[62] Vinutha, K. et al.: A Machine Learning based Facial Expression and Emotion Recognition for Human Computer Interaction through Fuzzy Logic System. In: 2023 International Conference on Inventive Computation Technologies (ICICT). pp. 166–173 IEEE, Lalitpur, Nepal (2023). https://doi.org/10.1109/ICICT57646.2023.10134493.

[63] Yang, J. et al.: Facial Expression Recognition Based on Facial Action Unit. In: 2019 Tenth International Green and Sustainable Computing Conference (IGSC). pp. 1–6 IEEE, Alexandria, VA, USA (2019). https://doi.org/10.1109/IGSC48788.2019.8957163.

[64] Yang, B. et al.: Facial Expression Recognition Using Weighted Mixture Deep Neural Network Based on DoubleChannel Facial Images. IEEE Access. 6, 4630–4640 (2018). https://doi.org/10.1109/ACCESS.2017.2784096.