

FGW-FER: Lightweight Facial Expression Recognition with Attention

Huy-Hoang Dinh¹, Hong-Quan Do^{1*}, Trung-Tung Doan¹, Cuong Le²,
Ngo Xuan Bach³, Tu Minh Phuong³, and Viet-Vu Vu⁴

¹FPT University, Hanoi, Vietnam

[e-mail: hoangdhgch190440@fpt.edu.vn, quandh13@fe.edu.vn, tungdt27@fe.edu.vn]

²Electric Power University, Hanoi, Vietnam

[e-mail: lecuongbk@gmail.com]

³Posts and Telecommunications Institute of Technology, Hanoi, Vietnam

[e-mail: bachnx@ptit.edu.vn, phuongtm@ptit.edu.vn]

⁴VNU Information Technology Institute, Vietnam National University, Hanoi, Vietnam

[e-mail: vuvietvu@vnu.edu.vn]

*Corresponding author: Hong-Quan Do

*Received June 11, 2023; revised August 17, 2023; accepted August 28, 2023;
published September 30, 2023*

Abstract

The field of facial expression recognition (FER) has been actively researched to improve human-computer interaction. In recent years, deep learning techniques have gained popularity for addressing FER, with numerous studies proposing end-to-end frameworks that stack or widen significant convolutional neural network layers. While this has led to improved performance, it has also resulted in larger model sizes and longer inference times. To overcome this challenge, our work introduces a novel lightweight model architecture. The architecture incorporates three key factors: Depth-wise Separable Convolution, Residual Block, and Attention Modules. By doing so, we aim to strike a balance between model size, inference speed, and accuracy in FER tasks. Through extensive experimentation on popular benchmark FER datasets, our proposed method has demonstrated promising results. Notably, it stands out due to its substantial reduction in parameter count and faster inference time, while maintaining accuracy levels comparable to other lightweight models discussed in the existing literature.

Keywords: Attention, Depth-wise separable convolution, Facial expression recognition, Lightweight deep learning model, Residual block.

1. Introduction

Facial expression is a fundamental way for humans to convey emotions without verbal communication. In 1992, Ekman defined seven universal emotions [1], including Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise which have been widely adopted by researchers (as shown in Fig. 1). The area of Facial Expression Recognition (FER) has garnered significant attention in recent decades, as it plays a crucial role in enabling computers to interact with humans in a more natural and intuitive manner. FER applications can include customer service, smart education, and satisfaction assessment of subjects in public services. In customer service, FER helps to improve the ability to interact and respond to customers in a better way, improve the service experience and create higher satisfaction. In smart education, FER can help teachers and managers understand the level of participation and interaction of students, thereby creating the best conditions to enhance the learning process and meet the needs of each student. FER can also be applied in satisfaction assessment of citizens in public services, such as customer service, health services or public services. By understanding the emotions of users, agencies and organizations can adjust and improve services to better meet the needs and desires of users. Fig. 2 showcases a sample of our own ongoing FER system, elucidating the general process of a FER system. Taking image or video data as input, the first task of the complete FER system is to detect faces in the image/video. Then, it classifies the extracted face into one of the 7 basic emotions. It's crucial to underline that this paper exclusively focuses on the classification phase. This indicates that our input data has undergone preprocessing to incorporate the extracted facial visuals.

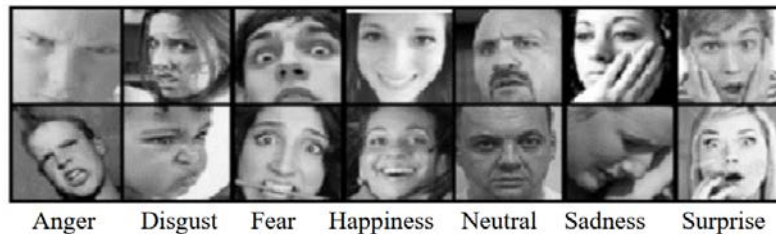


Fig. 1. Seven basic emotions in FER-Plus dataset [2].

Recently, deep learning techniques have been the primary method employed in the extensive research on Facial Expression Recognition [3]. The majority of these approaches rely heavily on convolutional operators to extract key visual features from input images. With numerous studies proposing end-to-end frameworks that stack or widen significant convolutional neural network layers, deep learning methods have become highly effective in analyzing and interpreting visual data. However, this has also led to larger model sizes and longer inference times. Recent advancements in lightweight models, such as MobileNets [4], have highlighted the potential of depth-wise separable convolution. This approach not only preserves the intrinsic space of the image but also reduces the considerable number of learnable parameters, allowing for real-time processing on mobile devices. The use of depth-wise separable convolution dates back to 2013 when Laurent Sifre developed it during his internship at Google Brain. Initially incorporated in AlexNet [5], this results in slight improvements in accuracy, significant enhancements in convergence speed, and notable reductions in model size. Following its success in AlexNet, depth-wise separable convolution was adopted as the initial layer in Inception model [6]. Andrew Howard, while at Google, further explored the potential of depth-wise separable convolution by developing efficient

mobile models called MobileNets [4]. Finally, in 2017, F. Chollet presented Xception [7], a specialized model aimed at reducing the computational cost and overall size of convolutional neural networks by leveraging the benefits of separable convolutions.

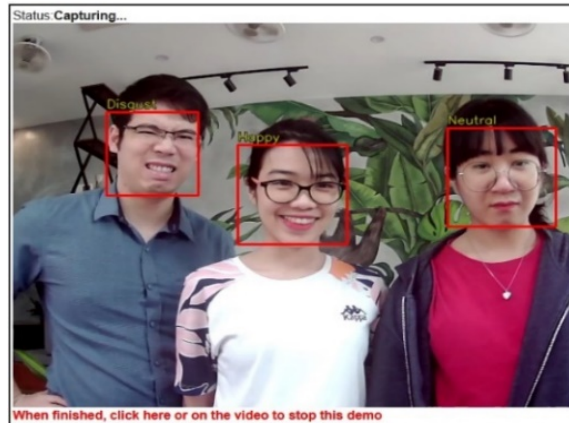


Fig. 2. An example of a facial expression recognition system that automatically detects faces in an image and recognizes the expressed emotion.

There's a recognition that constructing a lightweight model can yield substantial advantages, including the potential for deployment on resource-constrained devices and the reduction of both time and storage space overhead. As a result, our central research revolves around devising a lightweight deep learning model with a minimal parameter count, rapid training and inference times, all while maintaining the ability to effectively recognize facial expressions. We named our proposal method FGW-FER. It leverages three factors: Depth-wise Separable Convolution, Residual Block, Spatial and channel-wise attention modules. This distinguishes our approach from prior studies that often rely on a single factor or a combination of two of these factors. Depth-wise separable convolution layers are employed to reduce the model size and inference time, addressing the computational constraints. Residual blocks are integrated to mitigate the vanishing gradient problem that arises in deep models, thereby improving training efficiency. Additionally, we acknowledge that not all regions of a facial image contribute equally to the recognition of specific emotions. In many cases, only specific facial regions are crucial for understanding the underlying emotion. To address this, we incorporate an attention mechanism inspired by CBAM [8], which applies spatial and channel-wise attention to focus on the important regions of the face for learning. Moreover, it is important to note that we do not entirely replace all traditional convolution layers with depth-wise separable convolution layers; instead, we intermix both types to harness their respective benefits.

To gauge the efficacy of our method, we perform experiments on benchmark FER datasets, including FER2013 [9], CK+ [10], FER-Plus [2], and RAF-DB [11]; and compared the results with closely related works. The proposal method effectively enhances both the training and inference speed, while also achieving a significant reduction in the number of parameters and overall model size. Remarkably, despite these improvements, our model maintains a level of accuracy that is comparable to other lightweight models discussed in the existing literature. This highlights the capability of our approach to strike a balance between speed, model size, and accuracy, making it a promising solution for FER tasks.

The structure of the upcoming paper is outlined as follows: Section 2 will present an extensive review of relevant literature for our research. Section 3 will elaborate on our proposed FGW-FER architecture, while Section 4 will provide an overview of our experimental setups. In Section 5, the results obtained from our experiments will be presented and discussed. The paper will conclude with Section 6.

2. Related Works

2.1 Traditional Deep learning methods for FER

The impressive accomplishments of convolutional neural networks (CNN) or deep convolutional neural networks (DCNN) [12] in tasks such as image classification [13] have further expanded to encompass facial expression recognition as well. The concept of convolutional neural networks was initially introduced in the late 1980s, as they are particularly well-suited for processing matrix-shaped data like images or vectors. A typical architecture of a DCNN model comprises multiple convolutional layers, pooling layers, and fully connected layers. Over the years, numerous new DCNN architectures have been developed and refined, some of which have been applied to address facial expression recognition challenges. For instance, the traditional CNN model developed by Khorrami et al. in 2015 [14] achieved a high level of accuracy in emotion recognition. By utilizing zero-bias parameters, it attained a remarkable 95.1% accuracy on the Cohn-Kanade dataset (CK+) and 88.6% accuracy on the Toronto Face Dataset (TFD). Additionally, several other variations have made significant contributions to facial expression recognition. Examples include well-known models like AlexNet [5], VGGNet [15], Inception [6] [16], Residual Neural Network (ResNet) [13], and 2-Channel CNN [17]. The 2-Channel CNN architecture incorporates a standard CNN network in one channel while training a Convolutional Autoencoder in the other channel. Another noteworthy approach, as proposed in [18], involves training different facial parts in each channel.

There were also works that combine automatic features learned by deep learning models and hand-crafted features to solve FER. For example, in [19] they used three versions of VGG network (fine tunes VGG-Face and VGG-f, train from scratch VGG-13) for deep features and SIFT algorithm for handcrafted features. As a result, they achieved an exceptional performance that surpassed the previous approaches by more than 1%.

However, as previously mentioned, these models tend to be computationally demanding and resource-intensive, which presents obstacles when it comes to deploying them on devices with constrained computational capabilities, such as mobile phones or embedded systems. Therefore, there is a growing need for lightweight models for FER that are more efficient in terms of computational requirements and model size.

2.2 Attention mechanisms for FER methods

In recent years, the attention mechanism has gained considerable attention as a research topic due to its strong interpretability. It has become increasingly popular in both natural language processing and computer vision, with a diverse range of mechanisms and implementations. The attention mechanisms can be classified into two main types: one that focuses on enhancing the most important aspects of the data and another that utilizes the relationships between those aspects to produce a more meaningful representation.

In the context of facial expression recognition, Hu et al. proposed the SE-Net [20], which has a module that focuses on important features and has become a key module in the Efficient-

Net [21]. Woo et al. proposed the CBAM [8], which has two sub-modules Channel and Spatial Attention Modules to refine high-level feature maps and can be integrated into any CNN models without impacting their size and speed significantly. In 2020, W. Cao et al. [22] combined the VGG network with CBAM to achieve a recognition accuracy of 92% on the CK+ dataset. Pecoraro et al. proposed a Local Multi-Head Channel Self-Attention (LHC) [23] that uses channel-wise attention and global attention to overcome the limitations of convolution. The LHC achieved SOTA performance of 74.42% on the FER2013 dataset.

2.3 Lightweight models for FER

Lightweight models for FER typically involve reducing the number of layers and parameters in the model architecture while maintaining a good level of accuracy in recognizing facial expressions. These models can be developed using various techniques, such as pruning and quantization [24], knowledge distillation [25], or designing custom model architectures that are specifically optimized for the FER task. Pruning removes redundant weights that have minimal impact on the model's behavior, thus significantly reducing its size. On the other hand, quantization is a method that reduces computations by decreasing the precision of the datatype used for weights, biases, and activations. In contrast, knowledge distillation, introduced in 2015 [26], is a technique that transfers knowledge from a large, complex model (known as a teacher model) to a smaller model (referred to as a student model).

In the last-mentioned category, a recent study in 2021 [27] introduces a lightweight attention-based DCNN called LA-Net for facial expression recognition (FER). LA-Net incorporates squeeze-and-excitation (SE) modules and network slimming techniques to effectively reduce the model's size and computational demands. The SE modules play a crucial role by assigning weights to feature channels, enabling the network to concentrate on learning important facial features while filtering out redundant information. Additionally, network slimming contributes to further minimizing the model's size with minimal impact on accuracy. Subsequently, in 2022, there was a particular interest in compact network models, with several proposed models of this type published. [28] proposed a combination of Depth-wise Separable Convolutions, residual blocks, and Squeeze-and-Excitation Block, achieving accuracies of 66.29% on the FER2013 dataset with a model containing only 54,900 parameters. J. Zhi et al. [29] proposed a combination of the ResNet18 model with CBAM and frame-level attention mechanism to address facial expression recognition in image sequences (video-based FER). This proposed model had 13.39 million parameters and achieved recognition accuracies of 89.52% on the CK+ dataset and 88.33% on the eNTERFACE'05 dataset. Also in the same year, Y. Nan et al. [30] proposed a combination of Depth-wise Separable Convolutions and CBAM with a model containing 3.4 million parameters, resulting in recognition accuracies of 88.11% on the FER-Plus dataset and 84.49% on the RAF-DB dataset.

Compared to prior studies, our approach involves directly designing a lightweight model architecture for the Facial Expression Recognition (FER) task. Drawing inspiration from popular models such as MobileNets [4], ResNet [13], and Convolutional Block Attention Module (CBAM) [8], our model incorporates three key factors: depth-wise separable convolution modules, residual blocks, and Channel and Spatial attention modules. This sets our approach apart from many previous works that typically utilize only one or a combination of two of these factors (e.g., [22] uses only attention modules, [29] omits the use of depth-wise separation convolution, and [34] neglects residual blocks). Furthermore, our approach does not replace all traditional convolution blocks with depth-wise separable convolution. Instead, we carefully interleave both types of convolutions in the proposed architecture. This design choice aims to strike a balance between speed, model size, and accuracy, ensuring that

our model achieves proper performance for the FER task.

3. Materials and Methods

3.1 Architecture of the proposal FGW-FER model

Taking inspiration from the aforementioned deep learning models, we introduce a lightweight model that can maintain high performance on benchmark datasets. The proposed model comprises of an expansion part, Depth-wise part, and a classifier part. In addition, it utilizes regularization techniques, such as Batch Normalization and Dropout to enhance the learning efficiency of the model. Further details of each part are outlined below (Fig. 3).

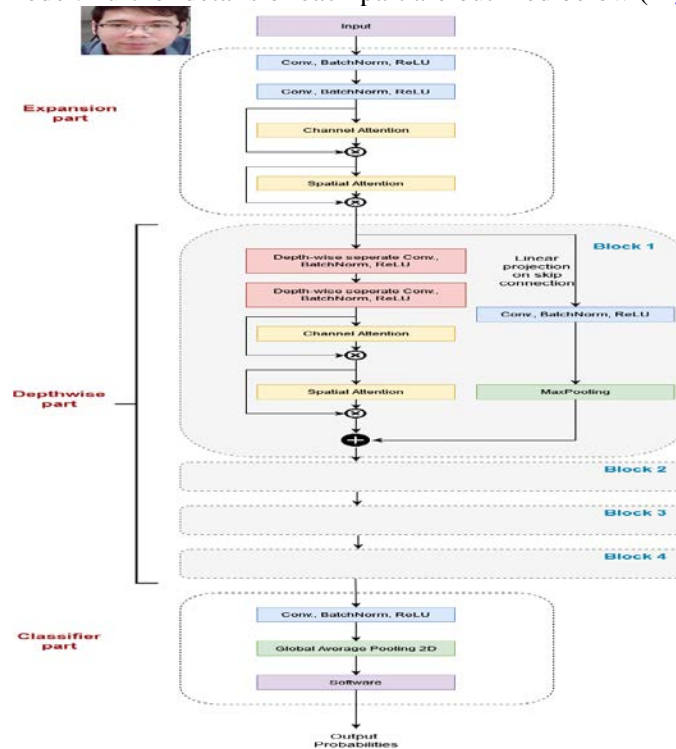


Fig. 3. Architecture of the proposal FGW-FER method.

- The first part is the expansion part, consisting of two traditional convolution layers with eight 3x3 filters and a stride of 1. This part extracts features from input images and passes them through Channel and Spatial modules to obtain a better representation of feature maps.
- The second part is the Depth-wise part that includes four blocks. Each block consists of two different branches: projection skip connection in the first branch and Depth-wise Separable Convolution layers performed in the second branch. The first branch contains [16, 32, 64, 128] traditional convolution 3x3 filters with a stride of 1 and a max-pooling layer. The second branch contains [16, 32, 64, 128] Depth-wise Separable Convolution filters with a stride of 2, and then passed through Channel and Spatial modules to obtain a set of feature maps. After processing the two branches, an additional operation is performed between their respective outputs. This addition merges the information from both branches, combining their feature maps. Furthermore, the output of the previous

block serves as the input to the subsequent block. This sequential flow allows for the propagation of features and information through the network, contributing to the overall learning process.

- The third part is the classifier part, which includes 3 layers: Convolution, Global Average Pooling 2D, Softmax to synthesize the information learned from the previous section and make predictions based on the probability of each label. Instead of using Fully Connection classes, this block uses Global Average Pooling 2D to reduce the parameters and number of calculations that the model has to perform.

Finally, in our proposed method, a cross-entropy loss function is used when adjusting model weights during training. It encourages the model to assign high probabilities to the true class and low probabilities to other classes. By minimizing the cross-entropy loss function, the model learns to predict the correct class with high probability. It is defined as:

$$L_{CE} = - \sum_{i=1}^n t_i \cdot \log_2(p_i) \tag{1}$$

, where t_i is the truth label and p_i is the Softmax probability for the i^{th} class.

The subsequent sections will detail the core layers of our proposed network.

3.2 Depth-wise Separable Convolution

Depth-wise Separable Convolution is a type of convolutional neural network layer that decomposes a standard convolution into two distinct layers: depth-wise convolution and pointwise convolution (Fig. 4). The depth-wise convolution performs a spatial convolution separately on each input channel using a relatively small kernel size, usually 3x3. This produces a set of output feature maps with the same channel count as the input. On the other hand, the pointwise convolution, also known as a 1x1 convolution, then applies a linear combination of these output feature maps across all channels to produce the final output. This combination effectively enhances the dimensionality of the output feature map while introducing non-linearity to the network.

Mathematically, let's consider an input tensor characterized by its dimensions ($H \times W \times C$). Here, H denotes the height, W signifies the width, and C represents the number of input channels. The depth-wise separable convolution consists of the following two steps.

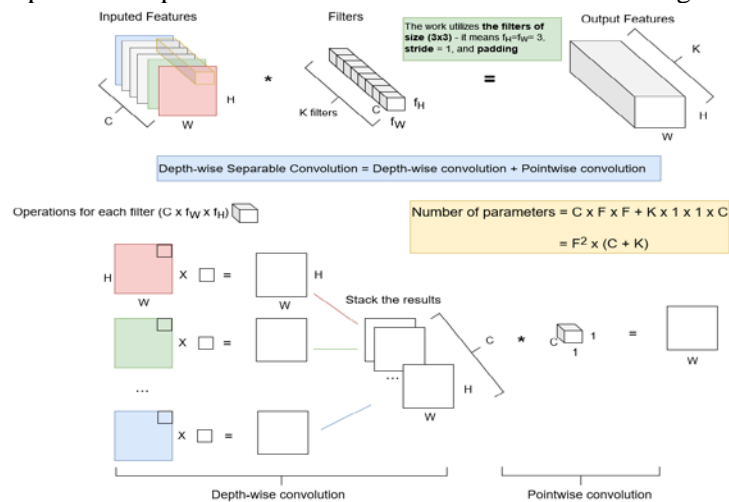


Fig. 4. Depth-wise Separable Convolution decomposes a standard convolution into two parts: depth-wise convolution and pointwise convolution.

Depth-wise convolution: In depth-wise convolution, instead of using a single filter that traverses all channels of the input as in a standard convolution, a separate filter is used for each channel of the input. This implies that for an input with C channels, we will have C depth-wise filters. This convolution gathers spatial features independently within each channel. Given an input tensor X of size $(H \times W \times C)$ and a depth-wise filter F of size $(f_H \times f_W \times C)$. For each channel c in the C input channels, the depth-wise convolution is applied as follows:

$$Y_c = X_c * F_c \quad (2)$$

Here, $*$ denotes the convolution operation, and Y_c is the corresponding output for channel c . The output tensor Y has dimensions $(H' \times W' \times C)$, where H' and W' depend on the stride, padding size, and filter size. In the project, a filter size of (3×3) , a filter stride of 1, and padding size of 1 are applied. This generates an output tensor Y with dimensions $(H \times W \times C)$, matching the size of the input tensor.

Pointwise convolution: Pointwise convolution, also known as (1×1) convolution, is performed after depth-wise convolution. It applies a set of (1×1) filters to the depth-wise outputs across channels to produce the final result. Pointwise convolution helps to combine and linearly transform the features learned from depth-wise convolution.

Given an input Y that is the output from the depth-wise convolution, with dimensions $(H \times W \times C)$, and a pointwise filter P with dimensions $(1 \times 1 \times C \times K)$ – where the final size K represents the number of filters – pointwise convolution is applied as follows:

$$Z = Y * P \quad (3)$$

Here, $*$ represents the convolution operation, and Z is the final output tensor. The resulting tensor Z has dimensions equal to that of Y , which is $(H \times W \times C)$.

In summary, a depth-wise separable convolution involves 2 steps:

1. Firstly, applying depth-wise convolution to each individual input channel.
2. Then, stacking these outputs from depth-wise convolution and applying pointwise convolution.

For the same input size, a standard convolution would require $f_H \times f_W \times (C \times K)$ parameters, while depth-wise separable convolution has a parameter count of $f_H \times f_W \times (C + K)$, where $(f_H \times f_W)$ is the size of the filters, C is the number of channels, and K is the number of filters. Thus, it can be said that depth-wise separable convolution uses fewer parameters and computations compared to standard convolution, making it an efficient alternative.

3.3 Residual Block with Projection Skip Connection

Residual blocks are the key components of ResNet, a deep neural network architecture that emerged victorious in the 2015 ImageNet Large Scale Visual Recognition Challenge [13]. A residual block consists of two main components: a main branch, which applies a series of convolutional layers and nonlinear activation functions to the input, and a shortcut connection, which passes the input through an identity mapping. The output of the main branch and the shortcut connection are then added, producing the final output of the block. However, to handle scenarios where the dimensions of the identity mapping and the stacked layers may differ, a linear projection can be performed. This projection involves using a convolutional neural network (CNN) to transform the identity function and align its dimensions with those of the stacked layers. By applying this linear projection, the dimensions can be matched, ensuring compatibility between the identity mapping and the subsequent layers (Fig. 5)

Mathematically, the residual block with a projection skip connection utilized in the proposed method can be represented as follows:

$$y = F(x) + H(x) \tag{4}$$

In this equation, x represents the input to the block, y represents the output of the block, F denotes the transformation carried out by the main branch, H represents the projection skip connection, and the '+' symbol signifies element-wise addition.

The inclusion of a skip connection between the input and output in the residual block enables the network to learn the residual mapping, which represents the difference between the input and output of the block. This relationship can be expressed as:

$$y - H(x) = F(x) \tag{5}$$

By focusing on learning the residual mapping rather than the complete mapping, the network can converge more efficiently and achieve improved performance, particularly in deeper networks. Furthermore, the skip connection addresses the issue of vanishing gradients by facilitating the direct flow of gradients from the output to the input of the block.

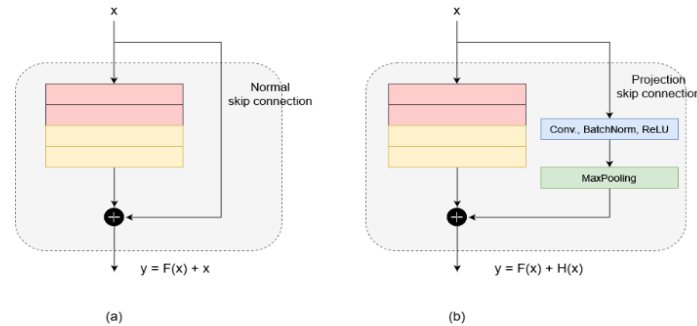


Fig. 5. Residual Block with Normal skip connection in (a) and Projection skip connection in (b).

3.4 Channel and Spatial Attention Modules

In this study, we will leverage the power of attention mechanisms to improve the performance of our model. Specifically, we will utilize two types of attention modules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). These modules, inspired by the pioneering work in [8], have shown promising results in various computer vision tasks. In Fig. 6, we provide a visual representation of how the CAM and SAM modules are applied within our network architecture. This visualization highlights the sequential nature of these attention mechanisms and their impact on the feature maps.

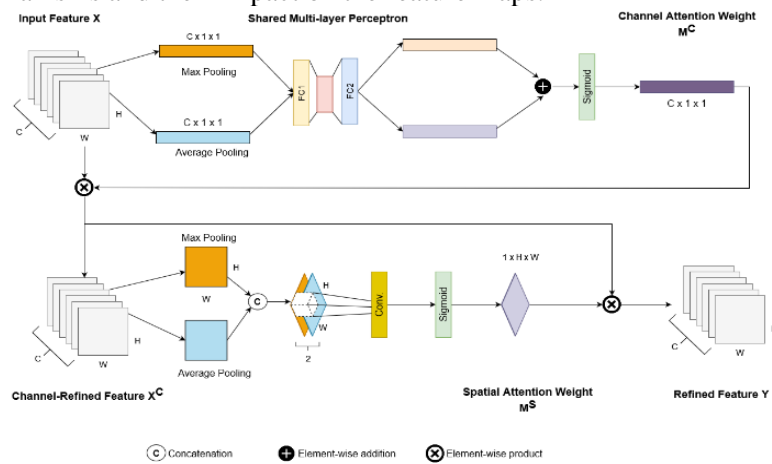


Fig. 6. The detailed structure of channel and spatial attention modules.

Let's first define the input feature of size $(H \times W \times C)$, where H and W are the height and width of the map, and C represents the number of channels. The two attention modules mathematically operate as follows.

The Channel Attention Module (CAM) aims to capture the interdependencies between channels and to learn a weighting coefficient for each channel that represents its importance in the feature map. It firstly computes the channel attention weights $M^C \in \mathbb{R}^{C \times 1 \times 1}$ as follows:

$$M^C = \sigma \left(MLP(X_{avg}) + MLP(X_{max}) \right) \quad (6)$$

where MLP is a shared network with ReLU activations that operate on the global average X_{avg} and maximum pooled feature maps X_{max} respectively, and σ is the sigmoid activation function. It should be noted that the MLP (Multi-Layer Perceptron) consists of a multi-layer perceptron with a single hidden layer. To minimize the parameter overhead, the size of the hidden activation is set to $\mathbb{R}^{\frac{C}{r} \times 1 \times 1}$, where r represents the reduction ratio. As a result, the expression for M^C can be reformulated in the following manner:

$$M^C = \delta \left(W_1(W_0(X_{avg})) + W_1(W_0(X_{max})) \right) \quad (7)$$

where $W_0 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_1 \in \mathbb{R}^{C \times \frac{C}{r}}$

Finally, the channel attention is applied to the input feature map X to achieve the channel-refined feature X^C :

$$X^C = M^C \otimes X \quad (8)$$

where \otimes represents the element-wise multiplication operator.

The Spatial Attention Module (SAM) aims to capture the spatial interdependencies between different regions of the feature map and to learn a weighting coefficient for each spatial location. Mathematically, firstly the input feature map X^C is concatenated with 1×1 convolution layers:

$$M^S = \sigma \left(MLP \left(\left[X_{avg}^C; X_{max}^C \right] \right) \right) \quad (9)$$

where $[. ; .]$ denotes the concatenation operation, MLP is a two fully connected layer with ReLU activations that operate on the concatenation of the spatial average and maximum pooled feature maps, respectively. The output is fed through a sigmoid activation function σ resulting in the spatial attention mask $M^S \in \mathbb{R}^{1 \times H \times W}$.

The final output denoted as Y , is the element-wise multiplication between the channel-wise and spatial attention maps:

$$Y = M^S \otimes (M^C \otimes X) \quad (10)$$

To summarize, by sequentially incorporating the CAM and SAM modules within our model, we can effectively refine feature maps from both the channel and spatial dimensions. This enables the model to focus on informative channels and relevant spatial regions, leading to improved discrimination and spatial awareness.

4. Experimental Setup

In this section, we delve into the details of our experimental setups. To start, we provide a succinct overview of the datasets employed in this study, highlighting their key characteristics and relevant details. Moving forward, we present the experimental configuration and parameter settings that were applied in our experiments. Lastly, we elaborate on the evaluation metrics that were utilized to assess the performance of our model, ensuring a comprehensive analysis of its effectiveness.

4.1 Datasets

In this work, an overview of datasets will be presented, including the extended Cohn-Kanade (CK+), FER2013, FER Plus, and RAF.

CK+: The extended Cohn-Kanade (known as CK+) [10] serves as a public dataset for both action units and emotion recognition. This dataset encompasses a variety of expressions, including both posed and non-posed (spontaneous) ones. In total, CK+ consists of 593 sequences involving 123 subjects. Previous studies have commonly used the last frame of these sequences for image-based facial expression recognition. Our experiments focus on seven specific expressions: Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Contempt. Notably, the distribution of classes in CK+ exhibits slight variations (as depicted in Fig. 7).

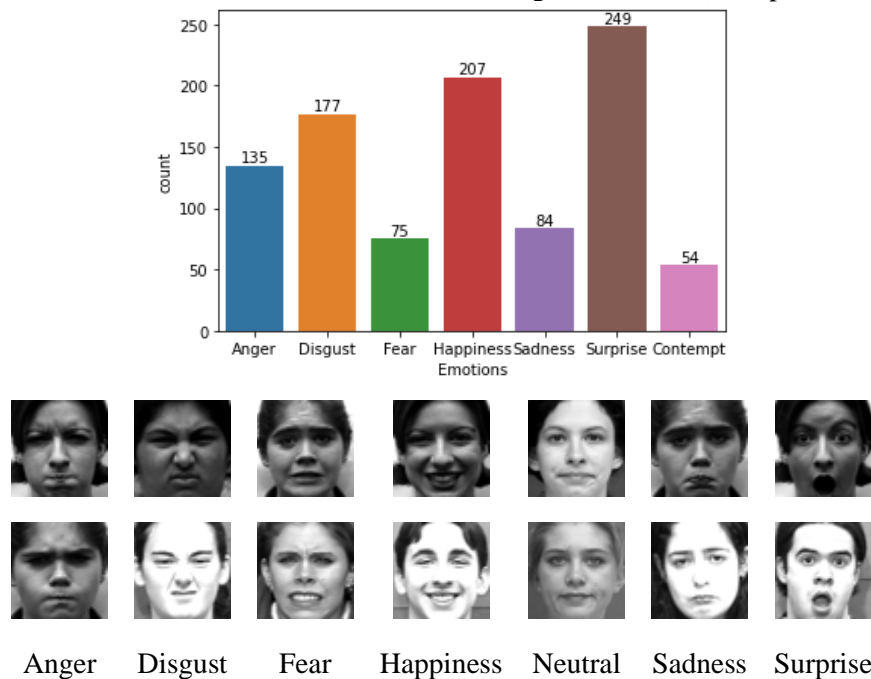


Fig. 7. Distribution classes and sample images in CK+ [10].

FER2013: The dataset was initially presented in the ICML 2013 Challenges in Representation Learning [9] to enable facial expression recognition. The dataset comprises 35,887 images with a resolution of 48x48, mostly captured in uncontrolled environments. The dataset is partitioned into three parts: a training set consisting of 28,709 images, a public test set of 3,589 images, and a private test set comprising 3,589 images. FER exhibits more variability in its images compared to other datasets, with examples of low-contrast images, partial faces, face occlusion, eyeglasses, and some images even lacking faces altogether. Fig. 8 showcases a few examples from the FER2013 dataset. In addition, the distribution between the number of classes is a huge difference. The label imbalance leads the model to only converge on the majority of data and ignore the rest. This is also the problem that the FER2013 dataset is not easy to achieve high results.

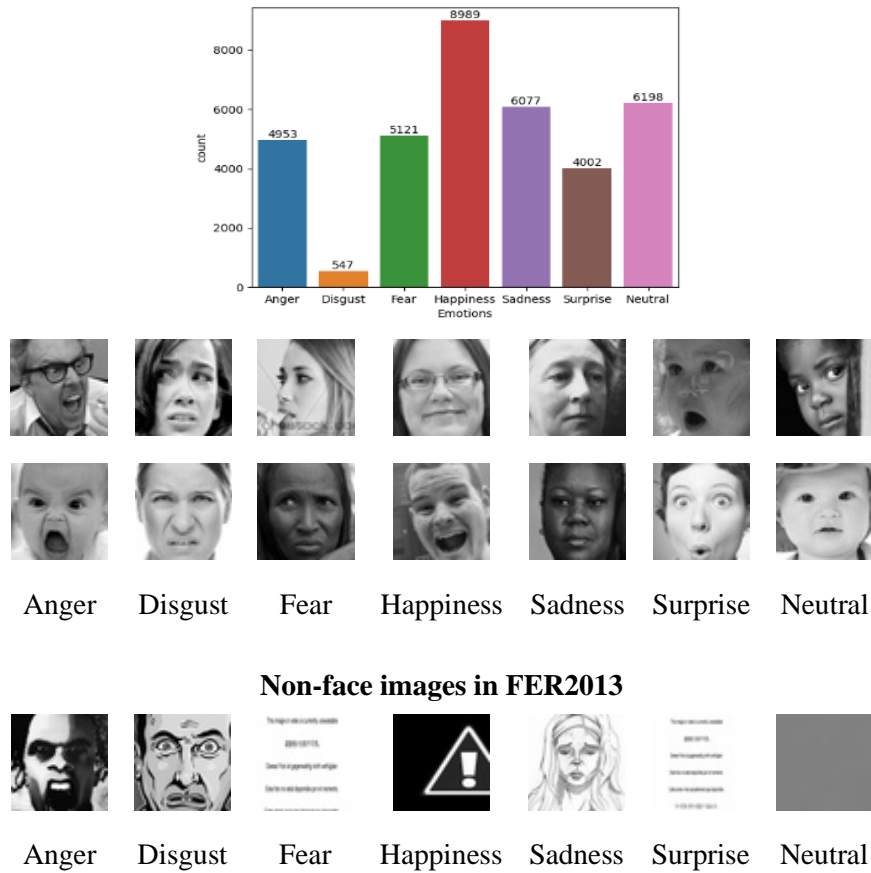


Fig. 8. Distribution classes and sample images in FER2013 [9]. This imbalance dataset includes several images without human faces, yet they are still labeled.

FER-Plus: is an extension of FER2013 dataset introduced by [2], comprises the same 35,483 grayscale images with a 48x48-pixel resolution as FER2013. Unlike FER2013, FER-Plus offers more comprehensive and accurate emotion labels. One notable distinction is that FER-Plus encompasses eight expression categories, including the addition of a contempt expression. However, it should be noted that the number of contempt expressions in the dataset is limited. To ensure consistency with other datasets, only the degree of presence for the seven basic emotions—anger, disgust, fear, happiness, sadness, surprise, and neutral—is included for each image. Each image is annotated by ten human labelers, and their annotations are aggregated using a probability distribution to generate the final labels. **Fig. 9** illustrates sample images from the FER-Plus dataset and their respective class distributions.

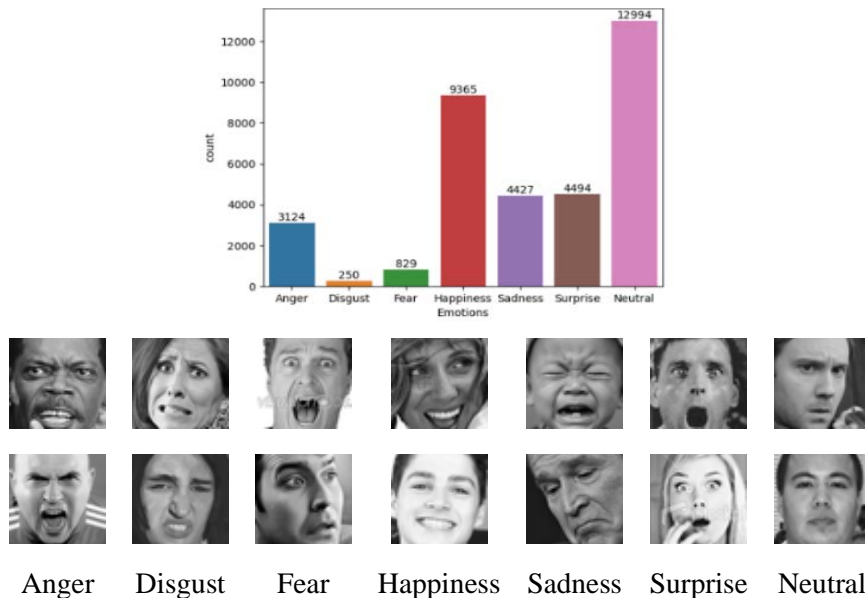


Fig. 9. Distribution classes and sample images in FER-Plus [2].

RAF-DB: a real-world emotional faces database [11] containing 29,672 RGB images of faces with a resolution of 100 x 100 pixels. The dataset is divided into two subsets: a single-label subset encompassing seven fundamental emotions, and a two-label subset comprising twelve types of emotions. For our study, we focused on the single-label subset, which includes 12,271 training images and 3068 test images. It's worth mentioning that the contempt expression is not included in the RAF-DB dataset. The images in RAF-DB exhibit varying sizes, ranging from small to large, which poses a challenge for deep learning models. Sample images from the RAF-DB dataset are displayed in Fig. 10.

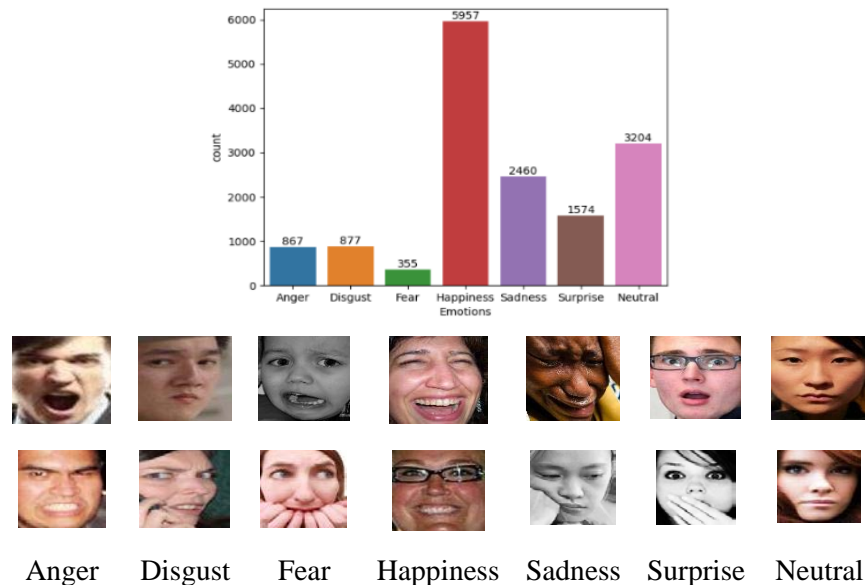


Fig. 10. Distribution classes and sample images in RAF-DB [11].

4.2 Experimental configuration and parameter settings

The experimental results in this study were conducted using a specific set of hardware and software specifications. The GPU used was an NVIDIA GTX 3060 with 12GB VRAM, while the CPU was an Intel Core i5 12600K with 16GB RAM. The operating system used was Ubuntu 20.04.5, and the deep learning framework used was PyTorch 1.10.2 and Torchvision 0.11.3. The version of Python used in this study was 3.6.13.

The FER2013 dataset comes pre-partitioned into training, validation, and testing sets at an 8-1-1 split ratio, which is publicly disclosed. Given that FER-Plus is a subsequent expansion of FER2013, its distribution adheres to the same 8-1-1 split. In contrast, for the CK+ and RAF-DB datasets, we segment them into training, validation, and testing sets at a 7-1-2 ratio during our experiments. The validation set is used during the training for parameter tuning. After adjusting and refining the model's parameters, the testing set is used to evaluate the final results. The results from the test data set are used to compare and evaluate the proposed model with other related works.

For each dataset used in the study, the proposed model mentioned above was trained using specific hyper-parameters. The training process was performed from scratch for a total of 300 epochs. During each iteration, a batch of 64 samples was processed simultaneously. To initialize the weights of the model, random Gaussian variables were employed with zero mean and a standard deviation of 0.05 for linear layers. For the convolutional layers, we applied the Kaiming initialization method [31], which specifically designed to help propagate gradients effectively, leading to improved training convergence. During the training process, we employed the AdamW optimization algorithm. The learning rate for the AdamW optimizer was set to 0.001, striking a balance between fast convergence and avoiding overshooting the optimal solution. Furthermore, a learning rate scheduler in the form of ReduceLROnPlateau was implemented.

4.3 Evaluation metrics

The obtained experimental results will be evaluated using several parameters, including the following:

- **Model Parameters and Model size:** Model parameters refers to the total number of learnable parameters, and it directly influences the model's complexity, memory usage, training time and inference time. On the other hand, the model size refers to the storage space required to store the model on disk. It is typically measured in megabytes (MB) and includes not only the parameters but also the model's architecture, configurations, and other associated metadata. The model size is an important consideration, especially in resource-constrained environments.
- **Overall classification accuracy:** This is a straightforward measure of the classification problem that involves dividing the number of correct predictions by all predictions. It can be expressed mathematically as:

$$CA = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (11)$$

- **Confusion Matrix:** is a tabular representation that summarizes the performance of a classification model. It is particularly useful when the dataset is imbalanced, where one class has significantly more samples than another. In such cases, accuracy alone can be misleading, and the confusion matrix provides a more detailed view of the classifier's performance. The confusion matrix allows us to compute various performance measures

that provide deeper insights into the classifier's effectiveness. These measures include precision, recall, and F1-score.

5. Experiment Results and Discussions

5.1 Overview of the experiment results

Table 1 presents a comprehensive overview of the experimental results obtained from our proposed method. The table includes various performance metrics such as the number of model parameters in the 5th column, the model size in the 6th column, and the inference time calculated by second per image. The table also provides the overall classification accuracy on CK+, FER2013, FER-Plus, and RAF-DB datasets, respectively. As mentioned earlier, the FER2013 dataset is particularly more challenging than other FER datasets due to class imbalance, posing difficulties during training. Our proposed model achieved an accuracy rate of approximately 69.38% on the FER2013 test set. Furthermore, we achieved an accuracy of 98.98% on the CK+ dataset, 79.36% on FER-Plus, and 80.75% on RAF-DB. These results demonstrate the effectiveness of our model across multiple datasets. One notable advantage of our model is its substantial reduction in parameter count, resulting in a memory footprint of only 0.32 MB. Additionally, our model exhibits faster inference times, achieving an impressive speed of up to 0.004 seconds per image. These factors contribute to the efficiency and practicality of our model for real-time facial expression recognition tasks.

Table 1. Overview of experiment results

| | #Params. | Size (Mb) | Inference time (s/image) | Overall Accuracy | | | |
|---|---------------|-------------|--------------------------|------------------|--------------|--------------|--------------|
| | | | | CK+ | FER-2013 | FER-Plus | RAF-DB |
| Replace all DSC* layers with Conv** in the proposed model | 321,816 | 1.4 | 0.005 | 97.6 | 68.73 | 79.24 | 79.55 |
| Our proposal FGW-FER | 64,176 | 0.32 | 0.004 | 98.98 | 69.38 | 79.36 | 80.75 |

* DSC: Depth-wise Separable Convolution; ** Conv: Traditional Convolution

In addition, when we substitute all Depth-wise Separable Convolution layers with Traditional convolution layers in the proposal model, the outcomes presented in **Table 1** demonstrated a decline across all evaluation metrics. Despite the model's parameter count increasing by over 5 times, recognition accuracy witnessed a decline of approximately 0.5% to 1%. This serves as additional substantiation, endorsing our assumption that an all-out replacement of traditional convolutional layers with Depth-wise Separable Convolution in the proposed model is not advisable. Instead, the model synergistically amalgamates both types to harness the distinct advantages each offers. Depth-wise Separable Convolution layers significantly curtail parameter count while upholding diversity and coherence among the Traditional convolutional layers, allowing the proposed model to mitigate loss.

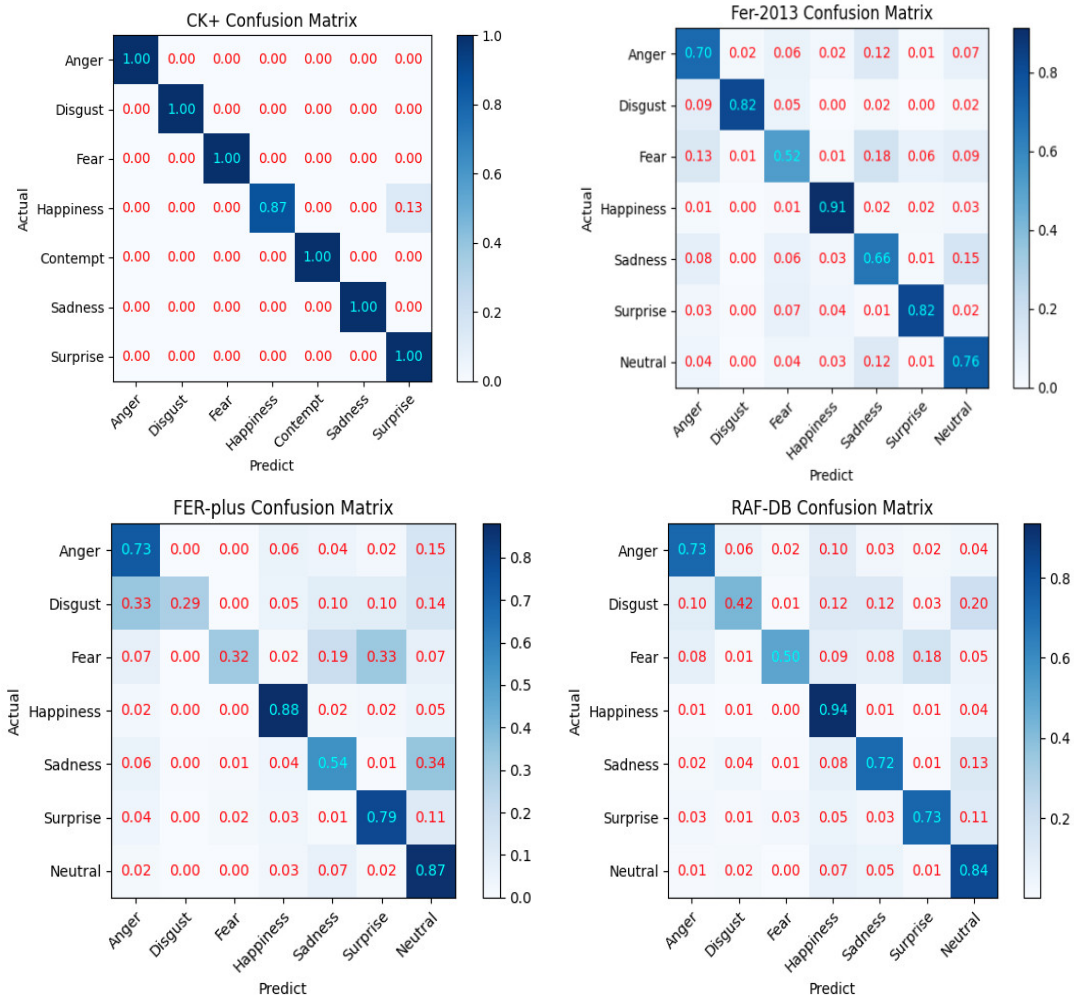


Fig. 11. Confusion matrices of our FGW-FER model on CK+, FER2013, FER-PLUS, and RAF-DB datasets.

Furthermore, **Fig. 11** includes the confusion matrices for the test set of the four datasets, providing valuable insights into the proposed method's recognition performance. The model showcased exceptional proficiency in recognizing the "happy" expression, achieving an accuracy rate of over 87% across all four datasets for this category. Specifically, in the CK+ dataset, the model achieved 100% accuracy for all expressions, except for the "happy" class. Moreover, in the FER2013 dataset, the "disgust" and "fear" classes also demonstrated high accuracy. However, in the FER-Plus and RAF-DB datasets, the limited availability of samples for the "disgust" and "fear" classes increased the likelihood of misclassification.

5.2 Comparison results with related works

In our experiments, we compared the performance of our proposed method with other related lightweight models from the literature. However, due to the unavailability of code for some of these models, we were unable to retrain them ourselves. Instead, we relied on the accuracy information provided in the literature to evaluate their performance on the test set. We

specifically selected these comparison models based on the criterion that their model sizes were published and relatively small, with fewer than 5 million parameters.

It is important to note that for the sake of maintaining consistency, certain related works will not be included in the comparison with our proposed method. This is because these works were conducted in different experimental setups. For example, DenseNet-1 [32] is not considered due to its reported recognition accuracy on the validation set rather than the test set, and A-MobileNet [30] is not included as it was pretrained on the ImageNet [33] dataset.

Table 2 to **Table 5** provide valuable insights for analyzing and interpreting the effectiveness of our proposed method in comparison to existing methods on the four FER datasets. In short, our model demonstrates superior performance compared to many common lightweight networks, including MobileNet V1 [4], MobileNet V2 [36], MobileNet V3 [35], and SqueezeNet [39]. It also showcases competitive performance compared to related lightweight models, making it an excellent choice for real-time expression recognition applications.

Table 2. Performance comparison on the CK+ dataset

| # | Ref. | Method used | Year | #Params. | Accuracy rate on CK+ |
|----------|------|------------------------------------|------|---------------|----------------------|
| 1 | [34] | Based on MobileNetV2 and Inception | 2019 | 2,639,239 | 92.4 |
| 2 | [4] | MobileNet v1 | 2017 | 3,213,575 | 95.0 |
| 3 | [35] | MobileNet v3 | 2019 | 4,210,711 | 96.0 |
| 4 | [36] | MobileNet v2 | 2018 | 2,232,263 | 98.0 |
| 5 | [37] | Deep-Emotion* | 2021 | 66,877 | 98.0 |
| 6 | [38] | MBCC-CNN** | 2021 | 4,384,175 | 98.48 |
| 7 | | Our proposal FGW-FER | | 64,176 | 98.98 |

* Deep-Emotion: Attention mechanism is added through spatial transformer network

** MBCC-CNN: multiple branch cross-connected convolutional neural network

To delve into the details, the results presented in **Table 2** indicate that our proposed model achieves the highest accuracy on the CK+ dataset compared to other works, while ranking third on the FER2013 dataset in **Table 3**, FER-Plus in **Table 4**, and RAF-DB dataset in **Table 5**. However, what sets our model apart is its impressive performance despite having fewer parameters than the majority of recent related works. With only 64,176 parameters, our model is surpassed in parameter count by only two works: [28] with 54,000 parameters and [40] with 58,423 parameters. However, our model achieves significantly higher accuracy than both of these works. While they report an average accuracy of 66.29% and 67% on the FER2013 dataset respectively, our model achieves an impressive accuracy of 69.38%. Furthermore, compared to the top-1 recognition accuracy [38] in FER2013, FER-Plus, and RAF-DB, their models have approximately 70 times more parameters than ours. This significant reduction in model size is noteworthy. Another noteworthy work is Deep-Emotion [37], which has a relatively small model size similar to ours. It achieves better accuracy on the FER2013 dataset while slightly lower accuracy on the CK+ dataset compared to our work.

Table 3. Performance comparison on the FER2013 dataset

| # | Ref. | Method used | Year | #Params. | Accuracy rate on FER2013 |
|----------|------|---------------------------------|------|---------------|--------------------------|
| 1 | [28] | DSC + RE + SE* | 2022 | 54,900 | 66.29 |
| 2 | [35] | MobileNet v3 | 2019 | 4,210,711 | 66.15 |
| 3 | [36] | MobileNet v2 | 2018 | 2,232,263 | 66.47 |
| 4 | [4] | MobileNet v1 | 2017 | 3,213,575 | 67.08 |
| 5 | [40] | MTCNN + RE + DSC** | 2021 | 58,423 | 67.00 |
| 6 | [37] | Deep-Emotion | 2021 | 66,877 | 70.02 |
| 7 | [38] | MBCC-CNN | 2021 | 4,384,751 | 71.52 |
| 8 | | Our proposal FGW-FER | | 64,176 | 69.38 |

* DSC: Depth-wise Separable Convolution, RE: Residual blocks, SE: Squeeze-and-Excitation Block

** MTCNN: multi-task cascaded convolutional network

Table 4. Performance comparison on the FER-PLUS dataset

| # | Ref. | Method used | Year | #Params. | Accuracy rate on FER-Plus |
|----------|------|---------------------------------|------|---------------|---------------------------|
| 1 | [35] | MobileNet v3 | 2019 | 4,210,711 | 71.42 |
| 2 | [36] | MobileNet v2 | 2018 | 2,232,263 | 79.26 |
| 3 | [39] | SqueezeNet | 2016 | 740,000 | 80.13 |
| 4 | [4] | MobileNet v1 | 2017 | 3,213,575 | 80.04 |
| 5 | [41] | ShuffleNet v2 | 2018 | 1,260,000 | 80.44 |
| 6 | [38] | MBCC-CNN* | 2021 | 4,384,751 | 88.10 |
| 7 | | Our proposal FGW-FER | | 64,176 | 80.36 |

* MBCC-CNN: multiple branch cross-connected convolutional neural network

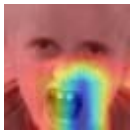
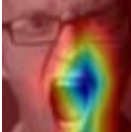
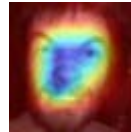
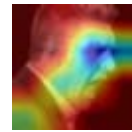
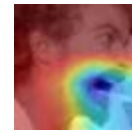
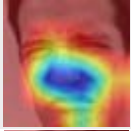
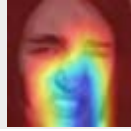
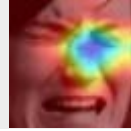
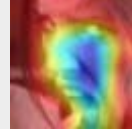
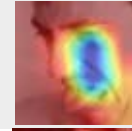
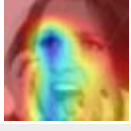
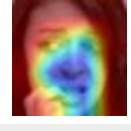
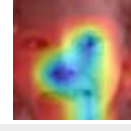
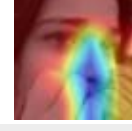
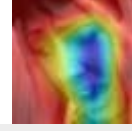

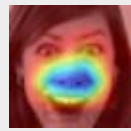

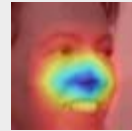
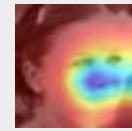
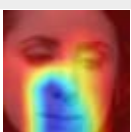
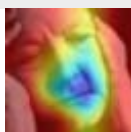
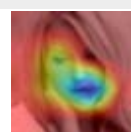
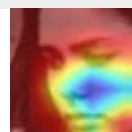
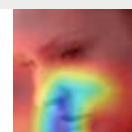
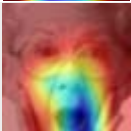
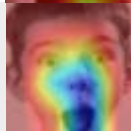
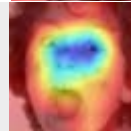
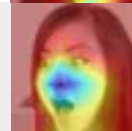
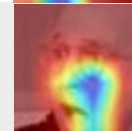

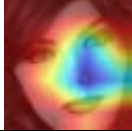
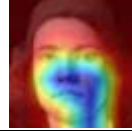
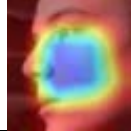
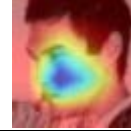
Table 5. Performance comparison on the RAF-DB dataset

| # | Ref. | Method used | Year | #Params. | Accuracy rate on RAF-DB |
|----------|------|---------------------------------|------|---------------|-------------------------|
| 1 | [4] | MobileNet V1 | 2017 | 3,213,575 | 79.92 |
| 2 | [36] | MobileNet V2 | 2018 | 2,232,263 | 70.13 |
| 3 | [35] | MobileNet V3 | 2019 | 4,210,711 | 78.6 |
| 4 | [27] | LA-Net (70% Pruned) | 2021 | 1,010,000 | 85.89 |
| 5 | [38] | MBCC-CNN | 2021 | 4,384,751 | 87.34 |
| 6 | | Our proposal FGW-FER | | 64,176 | 80.75 |

5.3 Attention Visualization

Table 6 and **Table 7** illustrate the attention maps depicting various emotions in FER2013. The tables consist of seven rows, with each row representing one of the seven expression categories. From top to bottom, the categories are anger, disgust, fear, happiness, sadness, surprise, and neutral. In **Table 6**, we have organized a set of five columns to showcase correctly labeled examples, where the first three columns display images of frontal faces, and the next two columns illustrate images of horizontally positioned faces. On the other hand, **Table 7** represents seven columns of images with incorrectly labeled emotions.

Table 6. Attention visualization of different expressions with true prediction on FER2013 dataset

| # | | True prediction | | | | |
|---|-----------|---|---|---|--|---|
| | | 1 | 2 | 3 | 4 | 5 |
| 1 | Anger |  |  |  |  |  |
| 2 | Disgust |  |  |  |  |  |
| 3 | Fear |  |  |  |  |  |
| 4 | Happiness |  |  |  |  |  |
| 5 | Sadness |  |  |  |  |  |
| 6 | Surprise |  |  |  |  |  |
| 7 | Neutral |  |  |  |  |  |

Examining the first row of **Table 6**, labeled as "Anger," we observe that the attention maps primarily focus on the mouth and nose regions for the frontal faces. In the case of horizontally positioned images, the attention maps concentrate on the informative areas such as the mouth or eyes. Moving to the second row, labeled as "Disgust," we find that the nose plays a significant role in determining this expression. In the horizontally positioned images, the nose alone is insufficient to convey the emotion, resulting in the attention maps spreading across the nose, eyes, and mouth regions. In the third row, to recognize the "Fear" expression, the attention maps concentrate on a relatively large area of the face. This indicates that when expressing fear, multiple regions of the face undergo noticeable changes, including the eyes, nose, and mouth. The attention maps presented in the fourth row for the "Happiness" expression provide a surprising insight. Initially, we anticipated that the mouth region would be the most crucial area for conveying happiness. However, in most samples labeled as "Happiness," the attention maps actually appear in the nose region. This suggests that the nose area is crucial in distinguishing the "Happiness" label from others. This understanding becomes more evident when we analyze the "Surprise" label. For the "Surprise" expression, most samples exhibit an expanded mouth, while elongated mouths without expansion can be found in several samples labeled as "Happiness." Lastly, for the "Sadness" and "Neutral" labels,

the attention maps focus on both the mouth and nose regions.

When analyzing images with mislabeled emotions in [Table 7](#), it is evident that the corresponding attention maps often fail to accurately emphasize the key regions observed in correctly labeled images. Nevertheless, a subset of these mislabeled images raises concerns regarding the accuracy of their assigned labels. Notably, in row 1, column 4, the original label indicating "Anger" contradicts our prediction of "Happiness." Similarly, in row 1, column 6, the original "Anger" label conflicts with our prediction of "Surprise." Likewise, in row 3, column 4, the original "Fear" label is at odds with our prediction of "Happiness," and in row 3, column 6, the original "Fear" label contrasts with our prediction of "Surprise." These instances underscore the inherent challenge in evaluating facial expressions within images, highlighting the propensity for confusion in both human and machine-based analyses.

Table 7. Attention visualization of different expressions with false prediction on FER2013 dataset

| # | Ground truth | False Prediction | | | | | | |
|---|--------------|------------------|--------------|-----------|----------------|--------------|---------------|--------------|
| | | 1 Anger | 2 Disgust | 3 Fear | 4 Happiness | 5 Sadness | 6 Surprise | 7 Neutral |
| 1 | Anger | | | | | | | |
| 2 | Disgust | | | | | | | |
| 3 | Fear | | | | | | | |
| 4 | Happiness | | | | | | | |
| 5 | Sadness | | | | | | | |
| 6 | Surprise | | | | | | | |
| 7 | Neutral | | | | | | | |

6. Conclusion

In the work, we presented a novel lightweight model architecture for Facial Expression Recognition (FER) task. The architecture integrates three essential factors: Depth-wise Separable Convolution, Residual Block, and Attention Modules. The objective of incorporating these factors is to achieve a balance between model size, inference speed, and accuracy in FER tasks. It opens up opportunities for efficient and effective deployment of real-time FER applications in various devices with a limited memory. Comparative analysis with existing lightweight models discussed in the literature demonstrates the advantages of our proposed approach. Despite its reduced parameter count and faster inference speed, our model achieves accuracy levels that are on par with or comparable to other state-of-the-art lightweight models. Specifically, the proposed model attains the highest accuracy on the CK+ dataset compared to other studies, while ranking third on the FER2013, FER-Plus, and RAF-DB datasets. Additionally, when visualizing attention maps for different expressions, they reveal the potential for misinterpretations in both human-based and machine-based analyses. This indeed poses a significant challenge in solving the facial expression recognition problem.

Moving forward, we plan to evaluate our approach on additional datasets, adapt it for video, and integrate it into web and mobile applications. Moreover, we intend to find solutions to develop more effective attention mechanisms on lightweight FER models.

Acknowledgement

This research has been done under the research project QG.21.58: "Researching and developing clustering integrating constraints and deep learning algorithms" of Vietnam National University, Hanoi.

References

- [1] P. Ekman, "Facial expressions of emotion: an old controversy and new findings," *Philos Trans R Soc Lond B Biol Sci.*, vol. 335, no. 1273, pp. 63-69, 1992. [Article \(CrossRef Link\)](#)
- [2] E. Barsoum, C. Zhang, C. Ferrer and Z. Zhang, "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution," in *Proc. of the 18th ACM International Conference on Multimodal Interaction*, New York, NY, USA, pp. 279-283, 2016. [Article \(CrossRef Link\)](#)
- [3] V. Dang, H. Do, V. Vu and B. Yoon, "Facial Expression Recognition: A Survey and its Applications," in *Proc. of 23rd International Conference on Advanced Communication Technology (ICACT)*, 2021. [Article \(CrossRef Link\)](#)
- [4] H. AG, Z. M, C. B, K. D, W. W, W. T and e. al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *CoRR*, vol. abs/1704.04861, 2017. [Article \(CrossRef Link\)](#)
- [5] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017. [Article \(CrossRef Link\)](#)
- [6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going Deeper with Convolutions," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015. [Article \(CrossRef Link\)](#)
- [7] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [Article \(CrossRef Link\)](#)
- [8] S. Woo, J. Park, J. Lee and I. Kweon, "CBAM: Convolutional Block Attention Module," in *Proc. of European Conference on Computer Vision*, pp. 3-19, 2018. [Article \(CrossRef Link\)](#)

- [9] I. Goodfellow, D. Erhan, P. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee and et al., "Challenges in Representation Learning: A Report on Three Machine Learning Contests," *Neural Networks*, vol. 64, pp. 59-63, 2015. [Article \(CrossRef Link\)](#)
- [10] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, San Francisco, CA, USA, 2010. [Article \(CrossRef Link\)](#)
- [11] S. Li, W. Deng and J. Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017. [Article \(CrossRef Link\)](#)
- [12] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998. [Article \(CrossRef Link\)](#)
- [13] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [Article \(CrossRef Link\)](#)
- [14] P. Khorrami, T. L. Paine and T. S. Huang, "Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition?," in *Proc. of IEEE International Conference on Computer Vision Workshop (ICCVW)*, Santiago, Chile, 2015. [Article \(CrossRef Link\)](#)
- [15] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proc. of International Conference on Learning Representations*, 2015. [Article \(CrossRef Link\)](#)
- [16] A. Mollahosseini, D. Chan and M. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Placid, NY, USA, 2016. [Article \(CrossRef Link\)](#)
- [17] D. Hamester, P. Barros and S. Wermter, "Face expression recognition with a 2-channel Convolutional Neural Network," in *Proc. of International Joint Conference on Neural Networks (IJCNN)*, 2015. [Article \(CrossRef Link\)](#)
- [18] L. Nwosu, H. Wang, J. Lu, I. Unwala, X. Yang and T. Zhang, "Deep Convolutional Neural Network for Facial Expression Recognition Using Facial Parts," in *Proc. of IEEE International Symposium on Dependable, Autonomic and Secure Computing*, 2017. [Article \(CrossRef Link\)](#)
- [19] M. -I. Georgescu, R. T. Ionescu and M. Popescu, "Local Learning With Deep and Handcrafted Features for Facial Expression Recognition," *IEEE Access*, vol. 7, pp. 64827-64836, 2019. [Article \(CrossRef Link\)](#)
- [20] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018. [Article \(CrossRef Link\)](#)
- [21] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. of the 36th International Conference on Machine Learning*, 2019. [Article \(CrossRef Link\)](#)
- [22] W. Cao, Z. Feng, D. Zhang and Y. Huang, "Facial Expression Recognition via a CBAM Embedded Network," *Procedia Computer Science*, vol. 174, pp. 463-477. [Article \(CrossRef Link\)](#)
- [23] P. R, B. V and B. V, "Local Multi-Head Channel Self-Attention for Facial Expression Recognition," *Information*, vol. 13, no. 9, 2022. [Article \(CrossRef Link\)](#)
- [24] T. Liang, J. Glossner, L. Wang, S. Shi and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370-403, 2021. [Article \(CrossRef Link\)](#)
- [25] A. A., A. F. and A. I., "Knowledge distillation in deep learning and its applications," *PeerJ Comput Sci.*, 2021. [Article \(CrossRef Link\)](#)
- [26] H. G., V. O. and D. J., "Distilling the knowledge in a neural network," in *Proc. of NIPS Deep Learning and Representation Learning Workshop*, 2015. [Article \(CrossRef Link\)](#)
- [27] H. Ma, T. Celik and H.-C. Li, "Lightweight attention convolutional neural network through network slimming for robust facial expression recognition," *Signal, Image and Video Processing*, vol. 15, pp. 1507-1515, 2021. [Article \(CrossRef Link\)](#)

- [28] X. Xu, J. Cui, X. Chen and C.-L. Chen, "A Facial Expression Recognition Method based on Residual Separable Convolutional Neural Network," *Journal of Network Intelligence*, vol. 7, no. 1, pp. 59-69, 2022. [Article \(CrossRef Link\)](#)
- [29] J. Zhi, T. Song, K. Yu, F. Yuan, H. Wang, G. Hu and H. Yang, "Multi-Attention Module for Dynamic Facial Emotion Recognition," *Information*, vol. 13, no. 5, 2022. [Article \(CrossRef Link\)](#)
- [30] Y. Nan, J. Ju, Q. Hua, H. Zhang and B. Wang, "A-MobileNet: An approach of facial expression recognition," *Alexandria Engineering Journal*, vol. 61, no. 6, pp. 4435-4444, 2022. [Article \(CrossRef Link\)](#)
- [31] K. He, X. Zhang, S. Ren and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015. [Article \(CrossRef Link\)](#)
- [32] G. Zhao, H. Yang and M. Yu, "Expression recognition method based on a lightweight convolutional neural network," *IEEE Access*, vol. 8, p. 38528–38537, 2020. [Article \(CrossRef Link\)](#)
- [33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. of IEEE conference on computer vision and pattern recognition*, 2009. [Article \(CrossRef Link\)](#)
- [34] L. Ale, X. Fang, D. Chen, Y. Wang and N. Zhang, "Lightweight Deep Learning Model for Facial Expression Recognition," in *Proc. of 18th IEEE International Conference on Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference*, 2019. [Article \(CrossRef Link\)](#)
- [35] A. H. e. al., "Searching for MobileNetV3," in *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019. [Article \(CrossRef Link\)](#)
- [36] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018. [Article \(CrossRef Link\)](#)
- [37] S. Minaee, M. Minaei and A. Abdolrashidi, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network," *Sensors*, 21(9), 2021. [Article \(CrossRef Link\)](#)
- [38] C. Shi, C. Tan and L. Wang, "A facial expression recognition method based on a multibranch cross-connection convolutional neural network," *IEEE Access*, vol. 9, pp. 39255-39274, 2021. [Article \(CrossRef Link\)](#)
- [39] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," *ArXiv*, 2016. [Article \(CrossRef Link\)](#)
- [40] N. Zhou, R. Liang and W. Shi, "A Lightweight Convolutional Neural Network for Real-Time Facial Expression Detection," *IEEE Access*, vol. 9, pp. 5573-5584, 2021. [Article \(CrossRef Link\)](#)
- [41] N. Ma, X. Zhang, H. Zheng and J. Sun, "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design," in *Proc. of the European Conference on Computer Vision (ECCV)*, pp. 122-138, 2018. [Article \(CrossRef Link\)](#)



Huy-Hoang Dinh is presently pursuing a Bachelor of Science degree in computing at the FPT campus of Greenwich University in Hanoi, Vietnam. His research primarily centers on cutting-edge Deep Learning models, including Transformers, GANs, and attention mechanisms, with a primary emphasis on their applications within the field of Computer Vision. Currently, he is actively engaged in the development of a product project with the objective of implementing license plate detection and integration solutions for public parking facilities in Vietnam.



Hong-Quan Do received a double M.S. degree in Information and Communication Technology from University of Science and Technology of Hanoi, Vietnam and The University of Rennes 1, France in 2015. He is a lecturer at FPT University, Hanoi, Vietnam. His research concentrates primarily on Clustering, Semi-supervised Clustering, Image processing and Recommender Systems. At present, he has also been involved in various E-Government projects, and E-Commerce Recommendation applications.



Trung-Tung Doan is a highly qualified computer scientist with a PhD degree from the University Blaise Pascal in France. He earned his PhD in 2012 and has since been working as a lecturer at FPT University. His research domains include Grid computing, Cloud computing, and Machine Learning. Currently, he is focusing his research on the fields of Data Science and Business Intelligence.



Cuong Le received his PhD from the Hanoi University of Sciences and Technology (HUST). He taught at School of Applied Mathematics and Informatics (SAMI), Hanoi University of Sciences and Technology (HUST) from 1998 to 2016. He is now a lecturer at Electric Power University. His research primarily focuses on information security, mathematical foundations for computer sciences and scientific computation as well as Quaternion and Clifford analysis in solving PDE.



Ngo Xuan Bach received his B.Sc. degree in Computer Science from University of Engineering and Technology (UET), Vietnam National University (VNU) Hanoi (2006), M.Sc. and Ph.D. degrees in Information Science from Japan Advanced Institute of Science and Technology (JAIST) Japan (2011 and 2014). He is now an Associate Professor of Computer Science, Vice Dean of Faculty of Information Technology, Posts and Telecommunications Institute of Technology (PTIT), Vietnam. His research interests include natural language processing, machine learning (deep learning), and recommender systems.



Tu Minh Phuong is professor of computer science and Chairman of University Council at Posts and Telecommunications Institute of Technology. His research interests include machine learning, recommender systems, natural language processing, and computer vision. He received Ph.D. degree in control in technical systems from the National Academy of Sciences, Uzbekistan, in 1995.



Viet-Vu Vu received the B.S. degree in Computer Science from Ha Noi University of Education in 2000, a M.S. degree in Computer Science from Hanoi University of Technology in 2004, and a Doctor Degree in Computer Science from Paris 6 University in 2011. He is a researcher at Information Technology Institute, Vietnam National University, Hanoi. His research interests include clustering, active learning, semi-supervised clustering, and E-government applications.