THE UNIVERSITY OF HONG KONG

DEPARTMENT OF

COMPUTER SCIENCE

Department of Computer Science

The University of Hong Kong

Final Year Project

Facial Expression Recognition without One-hot Labels (Group 1)

Final Individual Report

Duan Ruxiao

Team Members: DUAN Ruxiao, FU Yuming, SHANDILYA Eeshaanee

Project Supervisor: Dr. CHAN, Kwok Ping

TABLE OF CONTENTS

Acknowledgments

It is our utmost pleasure to express our sincere gratitude towards our supervisor and course coordinator, Dr. Kwok Ping CHAN for his guidance and constant supervision, as well as for providing necessary support regarding the project. We would further thank Dr. Chan for answering our numerous questions about facial expression recognition technology, computer vision, and deep learning with unfailing patience.

The completion of the project would also not be possible without the resources of the Department of Computer Science, the University of Hong Kong, particularly the HKU GPU Farm Phase 1 and Phase 2. Model training and result collection in this project are mostly conducted based on these facilities. We are grateful for the resources and the support team who helped us resolve occasional configuration issues with the GPU Farm.

List of Figures

List of Tables

Abstract

Facial expression recognition (FER), a technology that focuses on the prediction of human emotions based on their facial images, has become a hot topic of research in computer vision and deep learning. Convolutional neural network (CNN) is a deep learning model architecture which is commonly employed to handle FER-related tasks, and a facial image dataset is usually applied for model training. However, most existing facial image databases have a weakness: each of their images has only one single emotion label. This might induce information loss in model training, since human expressions are often more complex, and a single-label emotion might not be sufficient to represent an expression. This project investigates the advantages of multi-label learning of FER technology and attempts to construct a FER model with high prediction accuracy. Several model optimization techniques are carried out and their effectiveness is evaluated. FERPlus, a multi-label facial image dataset, is selected in this project for model training and testing.

Experiments are conducted to test model performance under various settings. Four models, VGG16, RAN, DAN, and MTEN respectively achieve a test accuracy of 86.06%, 86.82%, 88.51%, and 88.60%. Data augmentation and face alignment are both effective in some scenarios and can contribute to an accuracy increase of about 1% and 0.5% respectively. MSE is tested to be the best loss function for model training, and ImageNet appears to be a better option for model pre-training than AffectNet. The models derived from multi-label training outperform those trained by single-label data by 2% to 4%, depending on the model tested. When evaluating complex facial images, the accuracy increase brought by multi-label training rises to 10% for VGG16 model. The experiment results demonstrate the superiority of multi-label training, suggesting that future researchers can enhance their FER model performance by training with multi-label data.

1 Introduction

This chapter provides an overview of the project. Some background information is briefly introduced in section 1.1, and the project objectives are listed in section 1.2. Section 1.3 discusses the contributions of this project, and section 1.4 gives an outline of the remaining chapters of the report.

1.1 Background

The focus of this project is the **facial expression recognition** (FER) technology, which is currently a hot topic in computer vision. Facial expressions are a powerful tool to express one's emotion as a means of non-verbal communication, and being able to identify one's facial emotions can potentially aid in Human-Computer Interaction research. FER is a technology that aims to accurately predict people's emotions based on their facial images [1].

With the emergence of deep learning techniques, computer vision and FER technology also advance rapidly. A class of deep learning networks called **convolutional neural network** (CNN) has been invented to solve problems related to image analysis. Many tasks including image recognition and object detection can be properly handled by employing CNN model architecture, thanks to its capability to extract the feature information from the image.

However, with regard to the prediction of human emotions based on facial images in the wild, there is still a long way to go to achieve high accuracy: the best emotion prediction accuracy on AffectNet [2], SFEW [3], FER2013 [4], three well-known databases of wild facial expressions, are merely 66.3%, 56.4%, 76.9% respectively [5, 6, 7], implying that this technology, on some scenarios, is still immature and can hardly be applied under many circumstances where accurate predictions are required.

What factors can have limited the development of FER technology? One of the possible answers should be the complexity of human expressions. Research in the field of facial expression has illustrated that facial images of human beings are more complex than a single facial expression or emotion: more likely than not, the expression of a facial image tends to be a depiction of multiple emotions [8]. But most existing facial image datasets still adopt single-label annotation, i.e., only one type of emotion is assigned to each facial image. This might lead to information loss, if a large proportion of image samples actually reveal more than one emotion, and as a consequence, the model will learn much less from the training images, and test accuracy can hardly be further improved.

It has been augured that a singly labeled facial expression is an oversimplified approach to dealing with the issue: complex species like human beings, for example, may experience the emotions of surprise and fear at the same time in situations of unexpected fright [9]. While significant research of FER with one-hot label indeed contributes to important progress in the field, the limited applicability and practicality of a single hot label restrict researchers' efforts to further optimize the model [8]. Thus, this makes research on multiple hot labels crucial, which is the main topic this project aims to investigate.

1.2 Objectives

The project has two main objectives. First, deep learning models for accurate emotion prediction have to be constructed and optimized based on existing model architectures, and several optimization techniques should be carried out to test their effectiveness in the model learning process.

Second, the advantages of applying multi-label facial image data to model training must be assessed, and how much the model accuracy can be increased by adopting a multi-label learning scheme has to be measured.

1.3 Applications

The technology of FER promotes industries like medicine, employment, security, business, marketing, etc. In behavioral medicine, understanding facial expressions at the onset of a behavioral condition can contribute to the timely detection and prevention of the disease [1, 10]. In marketing, customer sentiment analysis may help companies evaluate customer preferences and get an insight into the current market trends. In security mechanisms, deceit detection also relies on the analysis of human emotions. Understanding human beings and their emotions is becoming increasingly more crucial in today's society, thus this project can potentially aid in some of the real-world challenges one faces today.

This project not only has the potential for applications in various industries of the modern world, but also pushes the boundaries of the field of deep learning and computer vision. Numerous kinds of optimization techniques are applied in this project on different kinds of models, and their effect on the model performance is scrutinized. The statistics in this project provide insights into some methods which future researchers can utilize to construct more powerful deep learning models for image analysis.

1.4 Report Outline

The report is divided into six chapters, and the subsequent chapters are arranged as follows. Chapter 2 further introduces some background information about this project in detail and encapsulates some previous work relevant to FER technology. Chapter 3 expounds on the project methodology, including data processing procedures, model evaluation metrics, model architecture selection, and model optimization approaches.

The experiments conducted in this project and the test results are demonstrated in Chapter 4, and the challenges and limitations encountered during the project, along with the proposed future work, are discussed in Chapter 5. Chapter 6 completes this report by summarizing the major aspects and drawing a conclusion.

2 Project Background and Literature Review

This chapter presents detailed background information and summarizes some past research relevant to the project topic.

2.1 Detailed Project Background

Some important concepts about the convolutional neural network model and how the model can be employed to perform facial expression recognition are explicated in section 2.1.1. Section 2.1.2 introduces some widely adopted facial image databases, and section 2.1.3 points out a potential issue that most existing databases have, which is a motivation of this project.

2.1.1 Convolutional Neural Network for Emotion Prediction

Convolutional neural network (**CNN**) is a class of deep learning models that is commonly used to solve problems including but not limited to image analysis, such as image classification and image object detection, and has been shown to be more than qualified to handle vision-related tasks [11]. Fig. 1 illustrates an example of CNN model architecture.

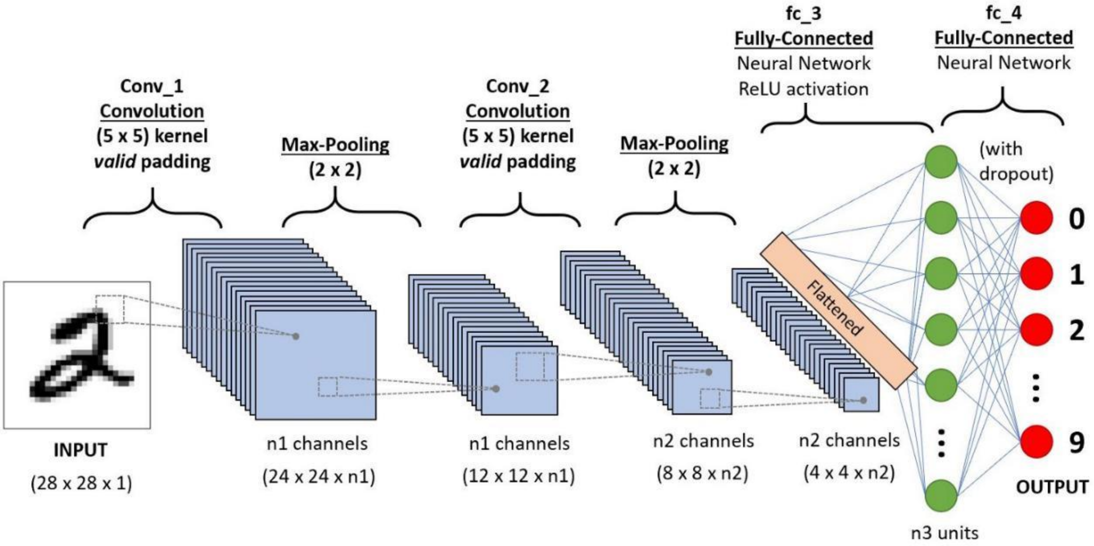


Fig. 1. The model architecture of an example CNN for digit recognition [12]. Each entry of the 10-dimensional output vector represents the input image's probability of being a specific digit. For the case shown in this figure, if the model is well-trained, the value of the entry that corresponds to the number "2" in the output vector should be the greatest among the ten.

The input of a CNN model is usually an image represented by a tensor of pixel values, each value within the range of [0, 255]. This image tensor passes through several convolution layers and pooling layers to get its features fully extracted and simplified. Subsequently, the feature map derived is flattened into a feature vector, which is then forwarded to a fully connected neural network. The final output vector stands for the probability distribution of the target object type [12]. In FER-specific tasks, the input to the CNN model is an image of a human face, and the output should be an emotion vector, with its dimension equal to the number of emotion categories. The emotion corresponding to the entry with the highest predicted value is chosen to be the predicted emotion.

Emotions can be classified into several main categories based on the classification scheme. Ekman & Friesen defined six basic emotions: **happiness**, **sadness**, **fear**, **disgust**, **surprise**, and **anger**, and it has become widely acknowledged by FER researchers [1, 8]. Most existing FER-oriented CNN models adopt emotion labels of either seven or eight types: seven types include the six basic emotions plus a **neutral**, and eight types further include one more emotion, **contempt**. Hence, the typical dimension of the output emotion vector in these FER models is 7 or 8.

A FER model must be trained by sufficient image samples in order to learn which kind of feature indicates which emotion, so that it can perform accurate predictions itself. Therefore, a large-scale dataset of facial images annotated with emotions is required for model training.

2.1.2 Single-label Facial Image Databases

There are many publicly available facial image databases with images annotated with emotion labels, and these datasets serve as the foundation for researchers to train and test their FER models. Three widely used facial expression databases which assign each image a single-label emotion are introduced below.

AffectNet [2] is a massive-scale dataset consisting of more than one million facial images collected from the Internet, and about one-half of these image samples are annotated manually into one of eight expression categories. Apart from the emotions, the intensity of valence and arousal is also recorded for the images, allowing researchers to construct continuous dimensional models for affective computing. The best test accuracy of AffectNet is currently 66.3% [5].

Static Facial Expressions in the Wild (**SFEW**) [3] is another database of facial expressions. SFEW is created based on static frames of a database named AFEW, and it contains only more than one thousand image samples, each tagged as one of seven emotion categories. The highest prediction accuracy of SFEW is 56.4% [6].

FER2013 [4] is a collection of about 35,000 facial grayscale facial images with the image size set to 48×48 pixels. The human face in each image is approximately located at the center of the image, and the size of each face is roughly the same. This adjustment eases the feature extraction and model learning process, thus might be one of the reasons why the test accuracy on this dataset is higher than the previous two. The state-of-the-art test accuracy of FER2013 is 76.9% [7].

2.1.3 A Flaw of Single-label Datasets

It is not hard to imagine a human face that betrays more than a single kind of feeling, and it should be common sense that people may often experience and show multiple emotions. Some past research, such as [8] and [9], has also supported this theory.

For instance, the upper image in Fig. 2 shows an expression that might indicate a mixed feeling of sadness and disgust, thus an emotion distribution, rather than only a single label, should be the more proper way to summarize the emotion information. Some emotions like anger in this image, although not being the primary feeling shown by the person, may still correspond to some crucial facial features (e.g., the downward corners of the mouth), thus it is not appropriate to eliminate this important information.

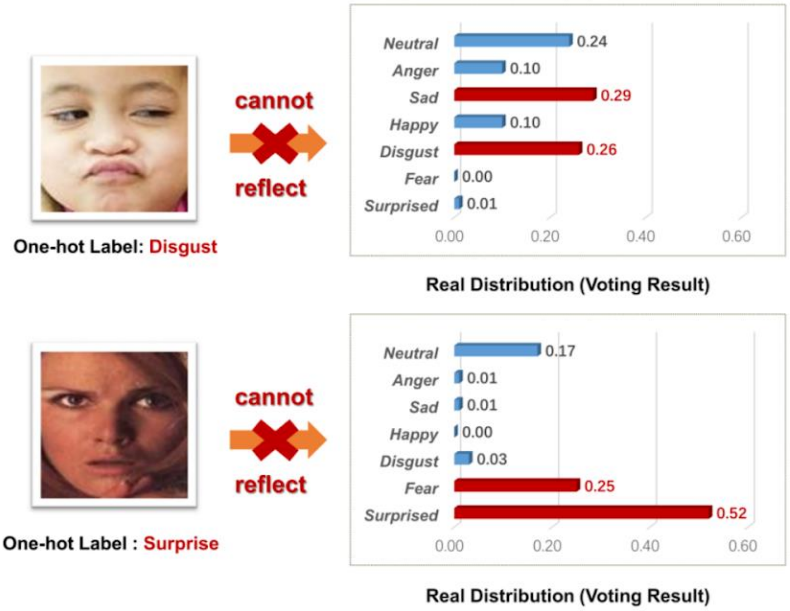


Fig. 2. A single-label emotion is insufficient to represent the real emotion distribution of a complex facial expression [13].

However, most existing facial image datasets annotate their image samples with a single-label emotion, which might impede researchers from training their models to higher accuracy. Since only one label instead of a complete distribution is given to the image, information loss is inevitable, which could significantly affect the performance of the models trained by these singly labeled data.

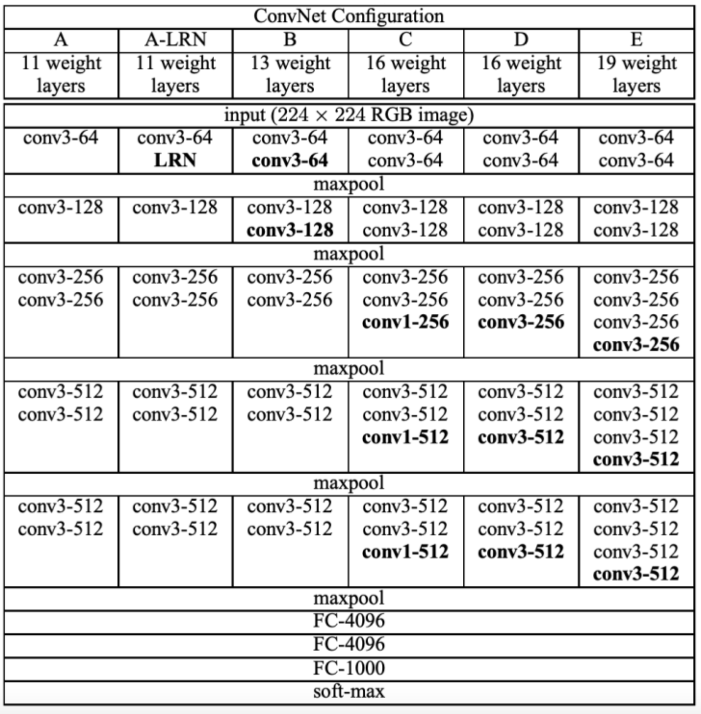
2.2 Related Work

This section provides an overview of some of the past research attempts concentrated on CNN, FER, and multi-label learning.

2.2.1 Classical Convolutional Neural Networks

Model performance is directly determined by its architecture. A CNN model has many hyper-parameters, including the number of layers, the size and number of convolutional filters, the type of pooling layers, etc. Previous researchers have devised various model architectures with excellent performance, and the following are two of the distinguished CNN architectures.

By adopting the conventional CNN structure, **VGG** model architecture was invented by the Visual Geometry Group from the University of Oxford in 2015 for large-scale image recognition tasks [14]. Small filters of size 3×3 were applied and model depth was increased to 19 layers. The model hyperparameters were tuned to let the model acquire maximum learning capability, and the model helped the development team achieve the state-of-the-art result in the ImageNet Challenge 2014. Several VGG models of different depths, including VGG11, VGG13, VGG16, and VGG19 were invented, and their configurations are illustrated in Fig. 3.



ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Fig. 3. Configurations of VGG model architecture [14].

The team stopped appending more layers to VGG model after the 19th layer, mainly because the model performance suffered from the increasing depth. This was a common problem in deep learning networks: although more layers in the neural networks enabled the model to analyze and process the obscure image features, they led to new issues such as gradient explosion and vanishing, which negatively affected the model's learning process. Therefore, conventional network architecture could no longer satisfy researchers' wish to obtain deeper and more powerful CNN models.

In 2015, He and his team proposed a new model architecture called **Residual Network** (**ResNet**) [15], which was regarded as a benchmark of the CNN development history. A residual block (Fig. 4) was introduced and integrated into the conventional CNN, which allowed low-level features to be retained and transferred to higher levels in the neural network without passing through intermediate feature extraction layers. This structure significantly alleviated the issues of gradient explosion and vanishing, thus it acted as a foundation for the model depth to be further increased. ResNet showcased outstanding performance and secured the first place in ImageNet detection, ImageNet localization in 2015. The network depth of the model was increased to 152 layers, and ResNets of various depths were constructed (Fig. 5).

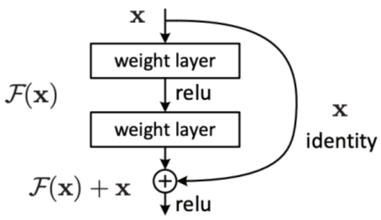
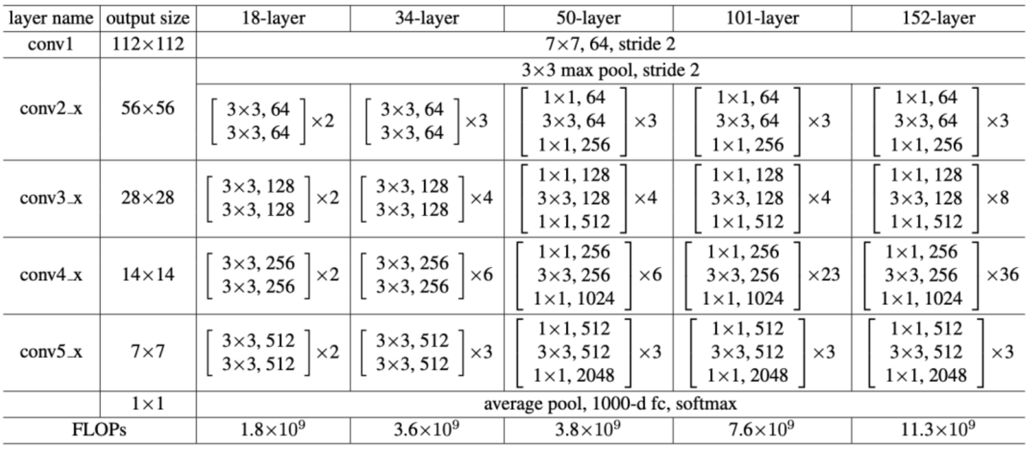

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Fig. 4. Residual block architecture [15].



layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		[3×3, 64 3×3, 64] ×2	[3×3, 64 3×3, 64] ×3	[1×1, 64 3×3, 64 1×1, 256] ×3	[1×1, 64 3×3, 64 1×1, 256] ×3	[1×1, 64 3×3, 64 1×1, 256] ×3
conv3_x	28×28	[3×3, 128 3×3, 128] ×2	[3×3, 128 3×3, 128] ×4	[1×1, 128 3×3, 128 1×1, 512] ×4	[1×1, 128 3×3, 128 1×1, 512] ×4	[1×1, 128 3×3, 128 1×1, 512] ×8
conv4_x	14×14	[3×3, 256 3×3, 256] ×2	[3×3, 256 3×3, 256] ×6	[1×1, 256 3×3, 256 1×1, 1024] ×6	[1×1, 256 3×3, 256 1×1, 1024] ×23	[1×1, 256 3×3, 256 1×1, 1024] ×36
conv5_x	7×7	[3×3, 512 3×3, 512] ×2	[3×3, 512 3×3, 512] ×3	[1×1, 512 3×3, 512 1×1, 2048] ×3	[1×1, 512 3×3, 512 1×1, 2048] ×3	[1×1, 512 3×3, 512 1×1, 2048] ×3
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Fig. 5. Configurations of ResNet model architecture [15].

2.2.2 Advanced Facial Expression Recognition Models

Most classical CNN models are designed for general image analysis problems (e.g., image classification, object detection, etc.), so their performance on FER-specific tasks may not be equally satisfying. Researchers focusing on the field of FER have proposed several new model structures based on CNN to help them better interpret facial features, and region attention is one of the approaches designed for FER.

Wang and his team introduced a CNN-based FER model architecture in 2019, called **Region Attention Network** (**RAN**) [6], in which an image cropping scheme was applied before forwarding the input face picture to backbone CNN, and an attention weight was assigned to each crop in the process to derive model output. The crop in which the face was at the crop center and was of appropriate size would receive a higher weight and would be interpreted as more significant than other crops. This mechanism enabled the model to more or less neglect those unimportant facial regions, such as the regions with occlusions (e.g., somebody in the image would wear glasses or a mask, or put their hands on the face), thus the more important region which contains decisive indications of emotion would be given more attention. The RAN model achieved the start-of-the-art test accuracy on the SFEW facial expression dataset (56.4%).

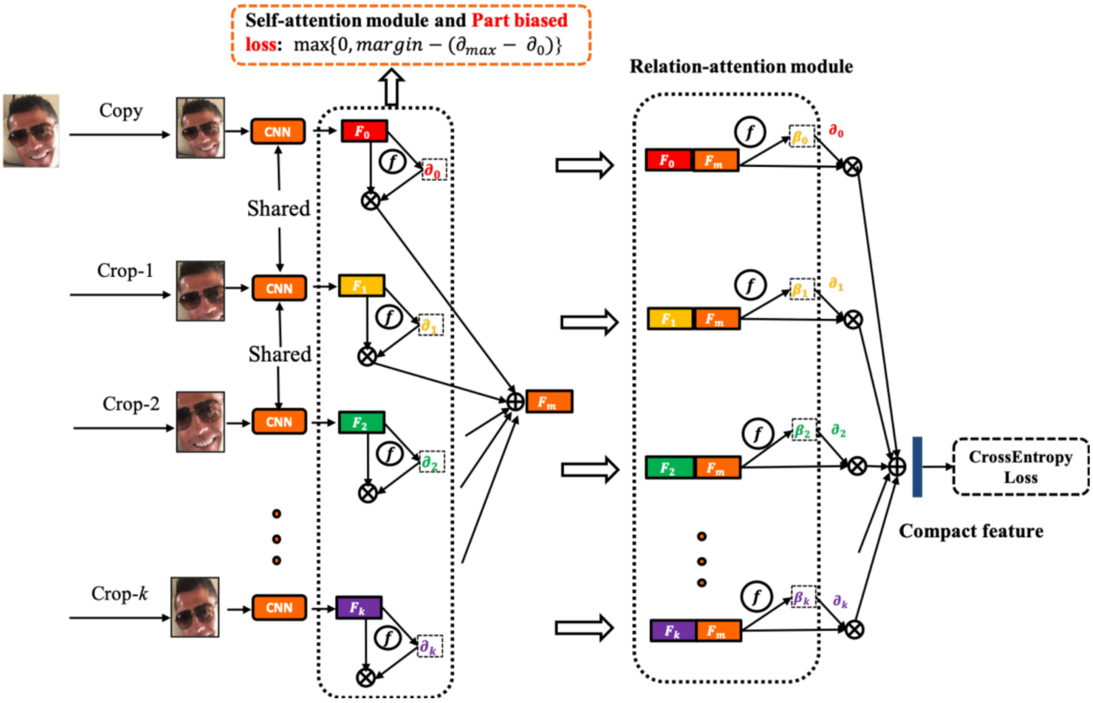


Fig. 6. Region attention network architecture [6].

The concept of multi-task learning neural network was put forward by Savchenko in 2021, who aimed to construct a single lightweight CNN to perform multiple tasks including the prediction of people's emotions, age, gender, and ethnicity [5]. A backbone CNN of this model was used to extract facial features, and these features were processed by multiple modules respectively to generate the predictions of various attributes of the person, such as emotion, age, and race (Fig. 7).

Several lightweight CNN architectures, including MobileNet [16], EfficientNet [17], and RexNet [18], were selected to be the backbone of this model, and their performance was measured based on some facial image databases. After the optimization process,

Multi-task EfficientNet-B2 turned out to be the best model for emotion prediction. The model achieved the highest test accuracy of AffectNet dataset (66.3%), which still has not been surpassed up to now.

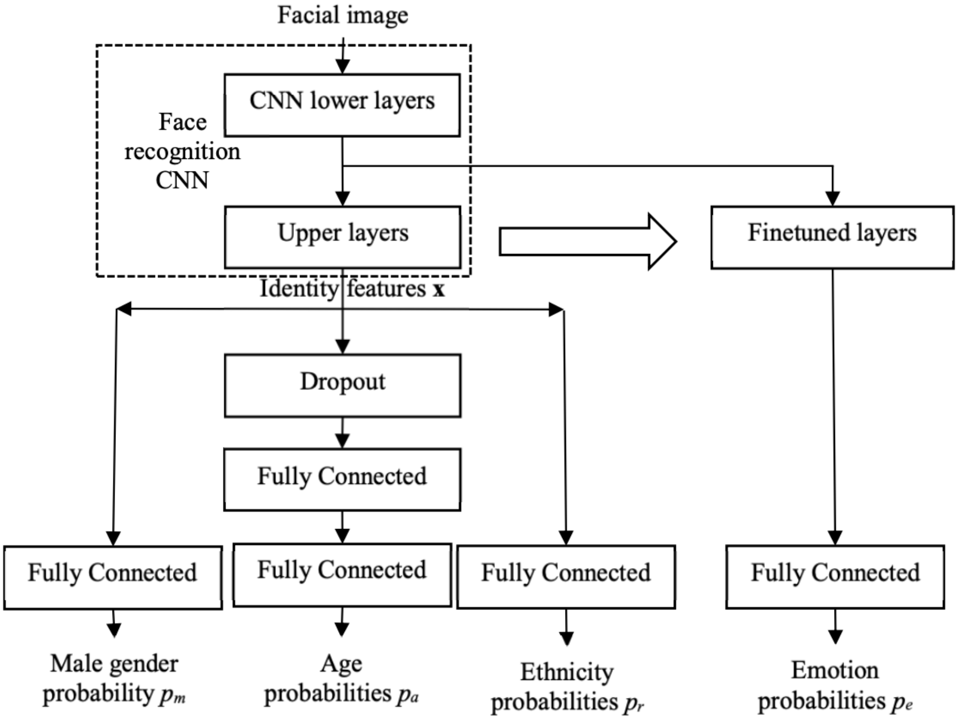


Fig. 7. The model architecture of multi-task CNN [5].

Another FER model called **FER-VT** [19] was proposed in 2021, which demonstrated remarkable performance in accurate emotion prediction. Traditional convolutional filters heavily depended on the spatial locality, thus had limited learning capability, especially when learning the long-range inductive biases between various sections of the facial image. FER-VT addressed this issue by designing two attention mechanisms for low-level feature extraction and the high-level semantic representation.

To lean the low-level image features, the long-range dependencies of different facial regions were modeled by the grid-wise attention mechanism. The feature map of the image, with attention weights attached to different grids of the face, was forwarded to a backbone CNN for feature extraction. The features from the middle layers and upper layers were further transformed into individual semantic tokens, and the global semantic representation was learned by adopting a technique called visual transformer [20]. The FER-VT model is currently the best model for the evaluation of facial images in the FERPlus dataset [21], with a test accuracy of 90.0%.

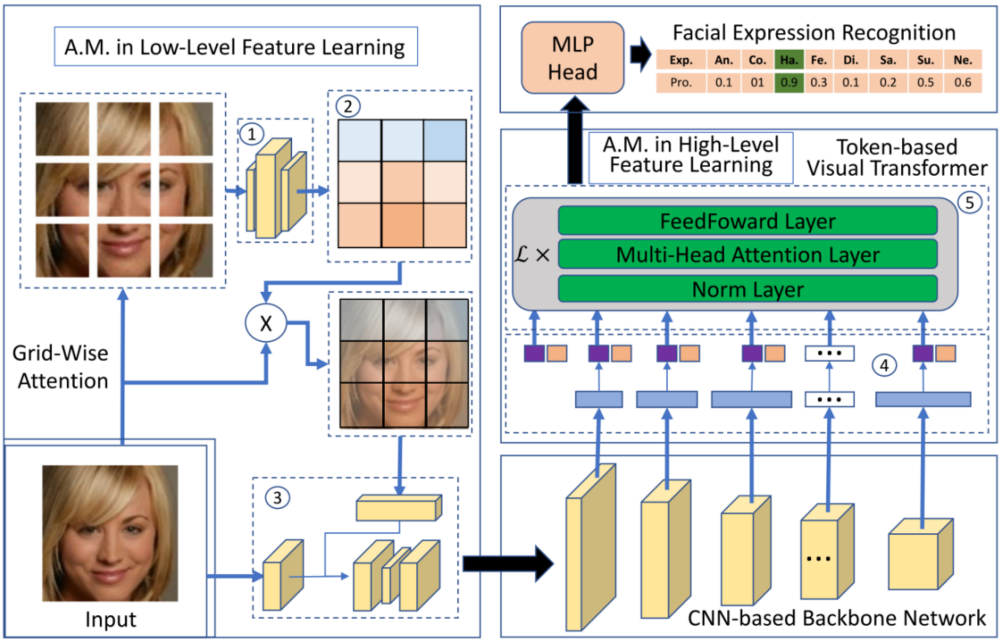

A.M. in Low-Level Feature Learning
1
2
Grid-Wise Attention
X
3
Input
MLP Head
Facial Expression Recognition
Exp.
An.
Co.
Ha.
Fe.
Di.
Sa.
Su.
Ne.
Pro.
0.1
01
0.9
0.3
0.1
0.2
0.5
0.6
A.M. in High-Level Feature Learning
Token-based Visual Transformer
5
FeedFoward Layer
$\mathcal{L}$ ×
Multi-Head Attention Layer
Norm Layer
4
CNN-based Backbone Network

Fig. 8. The framework of FER-VT model [19].

2.2.3 Multi-label Datasets

Many researchers realized the insufficiency of emotion representation by a single label, and the limitation that singly labeled datasets had on the FER-related study. Thus measures were put into effect to create facial image datasets with multi-label emotions to facilitate the research attempts for the community.

Li and Deng proposed the idea of compound emotion category when they created the **Real-world Affective Faces Database** (**RAF-DB**) [22], a large-scale image database containing 30 thousand facial expression images with great diversity. The images were collected from the Internet and labeled by 40 annotators. An emotion distribution vector of 7 dimensions was assigned to each image, and the image samples in RAF-DB were divided into different categories (Fig. 9). Some categories contained simple expressions that were considered as showing a single kind of emotion (e.g., happy, surprised, disgusted, etc.), while some categories were a collection of complex facial expressions, with images showing fixed feelings (e.g., fearfully surprised, sadly angry, happily surprised, etc.).

It was emphasized by Li and Deng that, since human facial expressions were, more often than not, a combination of various basic emotions, a single-label emotion might not be enough to represent the blended feeling of the person in the image. Hence, another facial image dataset called **Real-world Affective Faces Multi Label** (**RAF-ML**) was invented by them to further stress the issue of compound emotions. RAF-ML was a collection of about 5 thousand facial images with complex emotions from the Internet. The images in the dataset were labeled by 315 well-trained taggers, and a 6-dimensional emotion probability vector was assigned to each image sample

(Fig. 10). Only images with multi-peak label distribution were retained in the RAF-ML, thus this dataset mainly consisted of complex facial expressions.

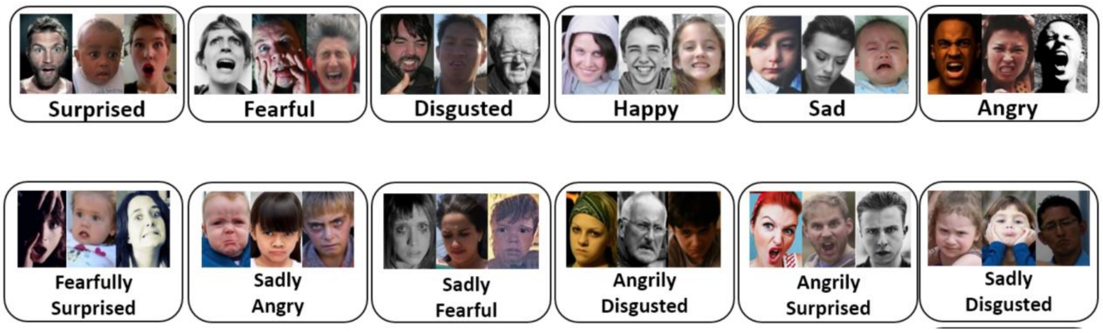


Fig. 9. Image categories in RAF-DB [22]. The top row is image categories of a single emotion, and the next two rows are the categories containing images of compound emotions.

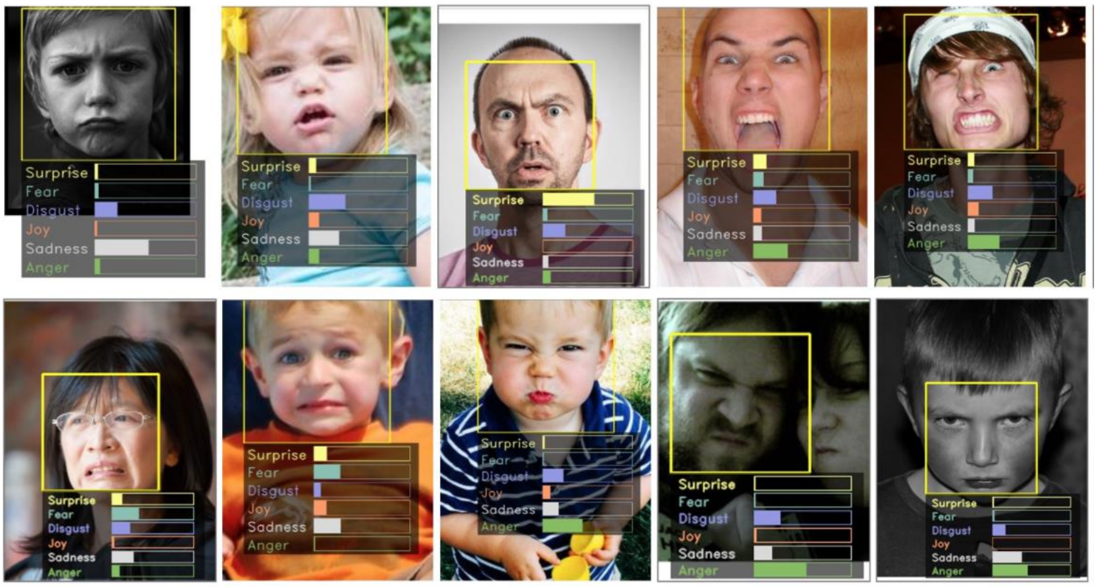


Fig. 10. Image samples in RAF-ML [9]. Each image has an expression distribution attached to represent the intensity of each emotion.

Based on the previous FER2013 dataset, Barsoum et al. [21] implemented a new dataset called **FERPlus**, which contained exactly the same image samples as FER2013, but with the emotion label of each image re-tagged by 10 annotators. Thus, images in FERPlus each obtained an expression distribution. (Some details of FERPlus to be further introduced in the next chapter.)

It was noticed by researchers that multiple taggers of an image could result in a more accurate interpretation of a facial expression [21]. For instance, among the two labels given to each facial image shown in Fig. 11, the upper one is from the FER2013 dataset, which is the opinion of the single tagger, while the lower one is the majority voting result from the ten taggers of FERPlus. It could be observed that the majority voting result turned out to be a more accurate representation of the emotion, since the new label given by FERPlus did receive more common agreement (Fig. 12).

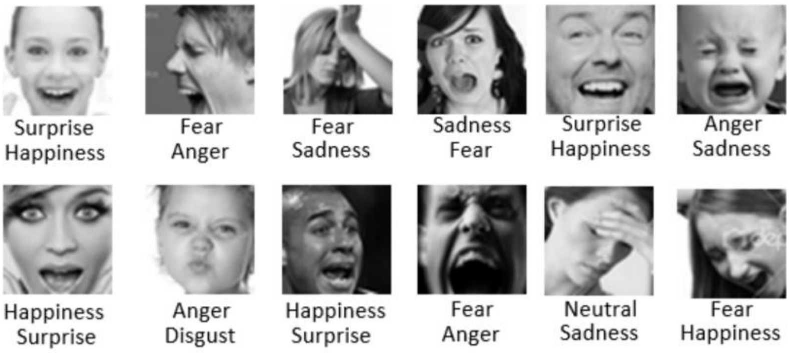


Fig. 11. FER2013 (upper label) and FERPlus (lower label) labeling examples [21]. The majority voting result given by FERPlus tends to be a more accurate interpretation of the person's emotion.

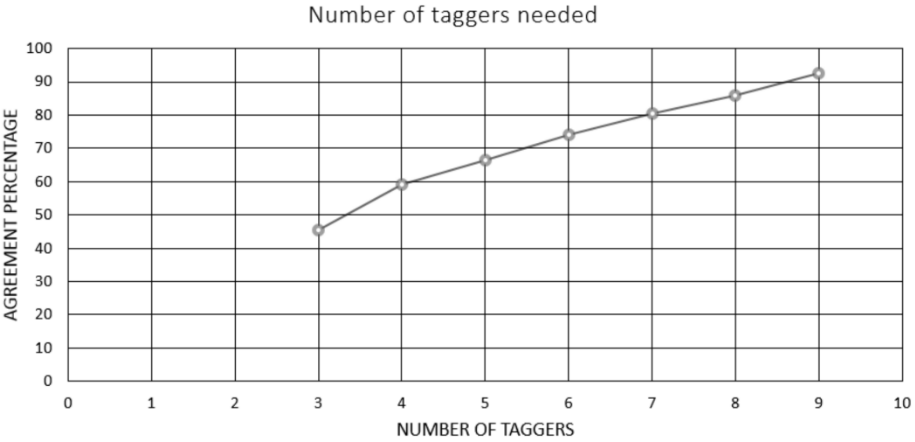


Fig. 12. Tagger count v.s. label quality [21]. 10 taggers should be sufficient to generally represent the opinion of the majority about the dominant emotion of a face.

2.2.4 Multi-label Learning

Based on these multi-label facial expression datasets, research was carried out to explore whether the multi-label data could be constructive for the development of FER technology.

Du and his team [8] emphasized the significance of using compound facial expressions to understand human cognition and proposed computational models to stress the complexity of facial expressions. With frameworks like Action Unit Analysis, facial expressions were predicted for images forwarded to the system [8].

Li and Deng devised a model architecture called Deep Bi-Manifold CNN and trained their model by the multi-label data in RAF-ML [9]. They implemented the model by learning the discriminative features of facial images while keeping the multi-fold structure of the facial image intact [9].

Barsoum et al. also attempted multi-label learning based on the multi-label data of FERPlus [21]. A VGG13 model was trained by different training schemes, including majority voting scheme (training by single-label data), and the label distribution training scheme (training by multi-label data) respectively, and the model performances were compared to each other. It was observed that multi-label training could result in a model with higher test accuracy. Several loss functions such as cross entropy loss and the probabilistic label drawing approach were applied to the multi-label model training process, and statistics showed that model trained by probabilistic label drawing seemed to have a slightly better performance than cross entropy loss. But both results were significantly superior to the model trained by single-label data (i.e., majority voting scheme).

3 Methodology

This chapter elucidates the methodology of the project. The dataset for model training and testing is firstly introduced in section 3.1, and the complexity level of a facial image is defined in section 3.2. Data pre-processing procedures are clarified in section 3.3, and the adopted accuracy metrics to evaluate the model performance are presented in section 3.4. Section 3.5 and section 3.6 describe the selected models for emotion prediction and several techniques to optimize the models, respectively. And finally, the software and hardware employed in the project are listed in section 3.7.

3.1 Dataset Details

The dataset applied for model training and testing is FERPlus, which consists of approximately 35,000 grayscale facial expression images with the size of 48×48 pixels [23]. The main reason that FERPlus is chosen for this project is that, unlike many existing single-label facial image datasets, the image samples in FERPlus are each evaluated by ten crowd-sourced taggers, thus it obtains an expression distribution of each image. Furthermore, compared to its counterparts such as RAF-DB and RAF-ML which also contain multi-label facial image data, FERPlus expands more emotion categories and owns more image samples: RAF-DB consists of 30 thousand images with only 7 emotion categories, while RAF-ML consists of only 5 thousand images with 6 emotion categories.

Every annotator of FERPlus is asked to label each image into one of the 10 emotion categories: neutral, happiness, surprise, sadness, anger, disgust, fear, contempt, unknown, and not a face (N.F.) [21]. Since the samples in FERPlus have multiple emotion labels, it facilitates researchers' attempts to estimate the probability distribution of the emotions behind a human face. For this project, training data with

N-hot labels can be directly obtained from FERPlus (details to be elaborated in section 3.3.3), thus this dataset perfectly meets the research objective of the project.

3.2 Facial Expression Complexity

The facial expressions in FERPlus images have diverse complexities (Fig. 13). Some expressions are trivial and can be represented by simply one or two emotion types (e.g., images at the top row of Fig. 13), but some expressions betray complicated feelings and can hardly be described by just one or two emotions (e.g., images at the bottom row of Fig. 13). Annotators' opinions hardly vary when evaluating trivial expressions, but their votes are usually given to different categories when they assess images with higher complexity.

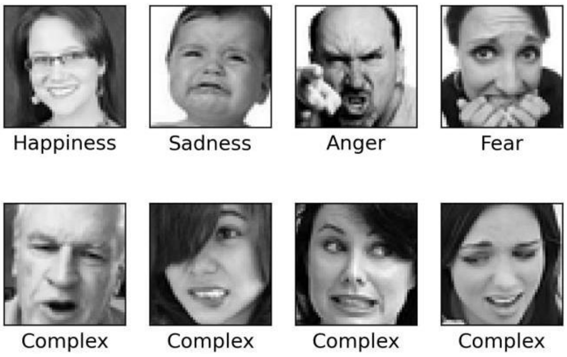


Fig. 13. Images in FERPlus are of various complexities.
Top images are trivial examples (with all the votes given to exactly 1 emotion category) and bottom images are extremely complex examples (votes allocated to 7 different categories).

In this project, the **facial expression complexity level** (or simply **complexity level**) X is defined for each image sample in FERPlus, where X equals the number of emotion categories on which the image receives votes. (E.g., X = 1 stands for images receiving votes on exactly 1 emotion, so these images should be easy to interpret. A higher value of X stands for higher complexity, since the taggers' opinions vary on these images.)

3.3 Data Pre-processing

This section explains the data processing methods applied in this project. A framework of data pre-processing procedures is exhibited in Fig. 14. The details of data cleaning, test data derivation, and training data derivation are further illustrated in the subsequent sections.

3.3.1 Data Cleaning

In the original FERPlus dataset, each image can be assigned labels of at most ten types [23]. However, two classes among these ten, i.e., unknown emotion (Un.) and not a face (N.F.), can negligibly contribute to the accurate prediction of a natural emotion, since they either are not a face image at all, or do not refer to a specific kind of emotion, keeping the target emotion unclear (Fig. 15).

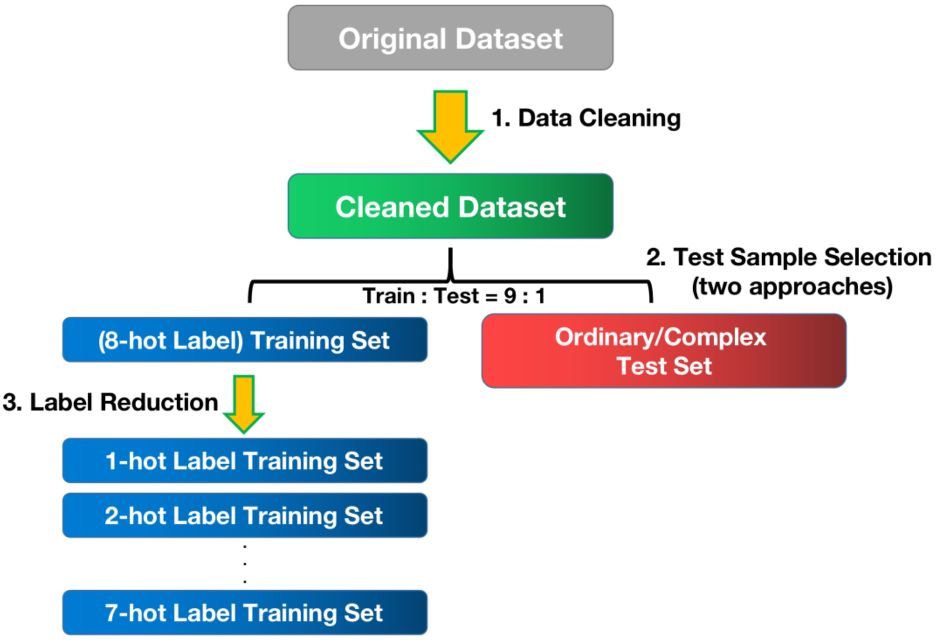


Fig. 14. Data pre-processing procedures. The bad samples in the dataset are eliminated by data cleaning. 10% of the cleaned dataset is extracted to obtain a test set, and the remaining data form the training set. Label reduction operation transforms the training set into different versions. Details of the procedures are to be elaborated in the following sections.

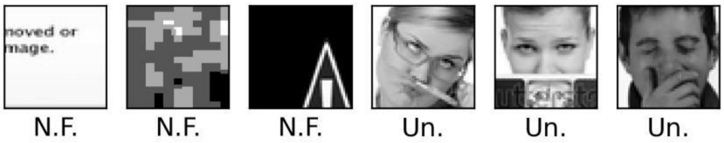


Fig. 15. Eliminated FERPlus image samples. These six selected images receive more than 3 votes on either N.F. or Un. category. (N.F.: not a face. Un.: unknown.) Unknown emotions are sometimes a consequence of image occlusion, thus the person's emotion is ambiguous.

Therefore, in this project, the images that are labeled by more than three taggers as unknown or N.F. are regarded as noisy samples and are eliminated from the dataset (e.g., Image A in Fig. 16). For all the remaining images, their emotion vectors are normalized after removing the unknown and N.F. columns (e.g., Image B and Image C in Fig. 16). Thus eventually, each image is assigned a corresponding emotion vector with eight entries, each representing the probability of having that specific emotion.

After data cleaning and the normalization of the emotion vector, the complexity level of each image has to be reassessed. Now, the level of complexity X should equal the number of emotion categories among the remaining 8 classes with positive estimated probability. Thus, the maximum complexity level X of an image should be 8 after data pre-processing.

The number of images at different levels of complexity in FERPlus after data cleaning is summarized in Fig. 17. It can be calculated that about 90% of the samples in FERPlus have complexity level X ≤ 3 (i.e., receive votes on no more than three emotions), which is predictable since few expressions can be so complex that more than three emotions are needed to describe them.

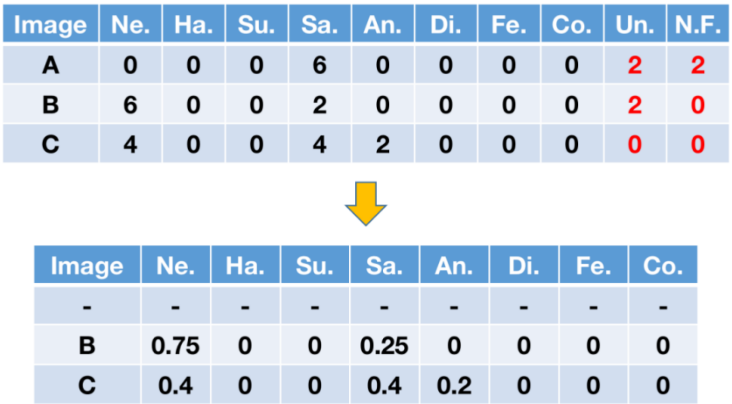


Image	Ne.	Ha.	Su.	Sa.	An.	Di.	Fe.	Co.	Un.	N.F.
A	0	0	0	6	0	0	0	0	2	2
B	6	0	0	2	0	0	0	0	2	0
C	4	0	0	4	2	0	0	0	0	0

Image	Ne.	Ha.	Su.	Sa.	An.	Di.	Fe.	Co.
-	-	-	-	-	-	-	-	-
B	0.75	0	0	0.25	0	0	0	0
C	0.4	0	0	0.4	0.2	0	0	0

Fig. 16. Data cleaning and expression distribution normalization. (Ne., Ha., etc., are the abbreviations for the emotion categories.)

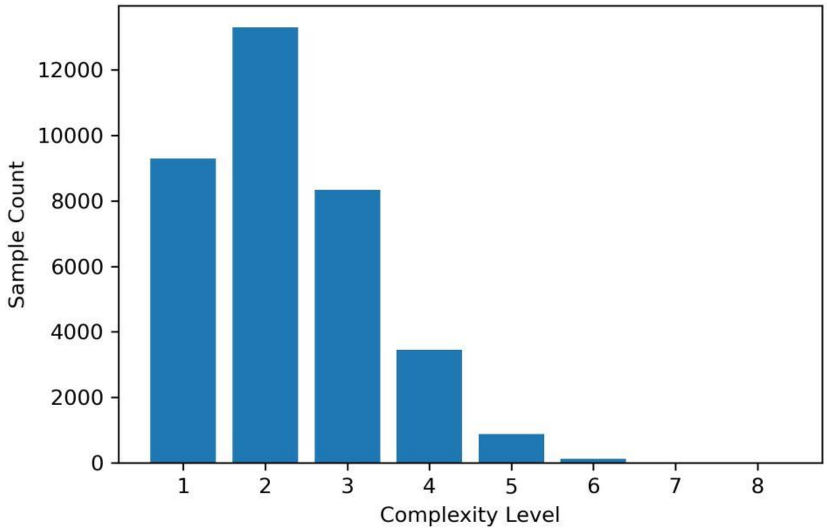


Fig. 17. The number of images of different complexities in FERPlus after data cleaning.

3.3.2 Ordinary Test Set and Complex Test Set

After data cleaning, the dataset is divided into a training set and a test set, with a train-test ratio of approximately 9:1. The complexity of the test set can be manually determined by applying different approaches to split the dataset.

If about one-tenth of the image samples are randomly drawn from the dataset, the test set obtained is regarded as an **ordinary test set**.

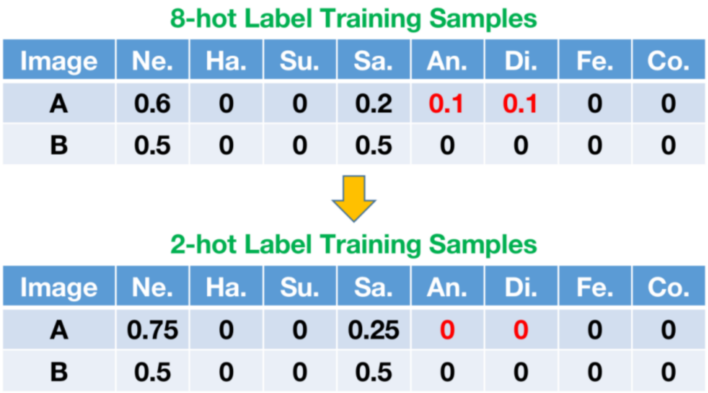
To acquire a test set that contains images with more complex emotions, the following procedures are adopted. Firstly, the images in FERPlus with complexity level X > 3 are defined as complex samples. These faces tend to be more difficult for models to analyze, and the prediction results generated for these expressions are often more inaccurate. To test the model's capability to evaluate these tricky images, it is desired that this new test set should consist of these images. Test samples are randomly drawn from these complex images, and the number of test images still remains roughly one-tenth of the entire dataset. The test set obtained by this approach is regarded as a **complex test set**. It is expected that model accuracy would be lower when tested by the complex test set.

Whichever scheme is adopted to derive the test set, the rest of the dataset is gathered to make up the training set, which consists of approximately 90% of the data samples.

3.3.3 N-hot Label Training Sets

Based on the training set, minor attributes in the emotion vector of each image are partially eliminated (label reduction) to obtain the N-hot label versions of the training data (N = 1, 2, ..., 7).

N-hot label data is defined to be the data with at most N labels (N = 1, 2, ..., 8), i.e., at most N entries of the emotion vector for each image can have a positive value. Theoretically, each image can receive votes on any of the eight emotion categories, hence the training data before processing is 8-hot label.



8-hot Label Training Samples

Image	Ne.	Ha.	Su.	Sa.	An.	Di.	Fe.	Co.
A	0.6	0	0	0.2	0.1	0.1	0	0
B	0.5	0	0	0.5	0	0	0	0

2-hot Label Training Samples

Image	Ne.	Ha.	Su.	Sa.	An.	Di.	Fe.	Co.
A	0.75	0	0	0.25	0	0	0	0
B	0.5	0	0	0.5	0	0	0	0

Fig. 18. The derivation of 2-hot label training samples. Image A has a complexity level of 4, thus it should have its minor attributes (angry & disgust) deleted and the resulting vector normalized. Image B has a complexity level of only 2, so no further processing is required.

To produce the N-hot label training dataset for each N < 8, images with complexity level no more than N do not require further processing (since the number of positive entries in their emotion vector is X, which does not exceed N). It suffices to sort the 8 emotions of each of the images with X > N according to the descending order of their probabilities (random order is applied to a tie), and then set the last 8-N entries to 0 and normalize the resulting emotion vector. The ultimate emotion vector has at most N entries holding positive values and the sum of entries is still 1. An example case is shown in Fig. 18.

3.4 Accuracy Metrics

The model output is an expression distribution vector of eight dimensions, indicating the corresponding probabilities of the eight emotions. The emotion with the highest probability is considered as the predicted emotion.

The predicted emotion is compared with the actual emotion vector of the image. If the predicted emotion matches with any of the emotions that have the highest value in the actual emotion vector, the prediction is considered accurate. The model test accuracy is calculated by dividing the number of accurate predictions by the number of test samples.

3.5 Prediction Models

In this project, some classic CNNs and three other advanced models are selected for emotion prediction.

The architecture of a model is directly related to its performance, thus model selection is a vital part of this project. Some classical CNN models including VGG [14], ResNet [15], DenseNet [24], MobileNet [16], etc., have proved to be highly qualified for tasks including image classification by previous experiments, so they are promising models to handle the task of FER. The architecture of VGG and ResNet is already introduced in section 2.2.1. DenseNet applies dense connections between layers to construct a relatively heavyweight model [24], and MobileNet utilizes depthwise separable convolutions to establish a relatively lightweight model [16]. All these four models have proved to be excellent in certain types of image analysis tasks.

Apart from classical CNN models, some CNN-based models designed to handle FER-related tasks should also be taken into account for the purpose of this project. Three advanced models, Region Attention Network (RAN) [6], Distract your Attention Network (DAN) [25] and Multi-task EfficientNet-B2 (MTEN) [5], are selected in this project to predict facial emotions. They have respectively achieved the state-of-the-art result for the evaluation of other three popular facial expression datasets, SFEW [3], RAF-DB [22], and AffectNet [2].

These models adopt special structures to stress common problems faced in FER tasks, such as occlusion-robust and pose-invariant issues, and how to relate different facial areas from a global perspective. Considering their performance on other datasets, it is expected that they can have a much higher prediction accuracy on FERPlus, compared to classical CNN models.

3.6 Optimization Approaches

This section introduces four techniques applied in the project to enhance the model performance.

3.6.1 Data Augmentation

By applying certain types of processing such as horizontal flip and rotation to the original training images, and appending the processed images with exactly the same emotion vector to the training dataset, data augmentation increases the diversity of training samples, thus helping the model learn more image features in the training process.

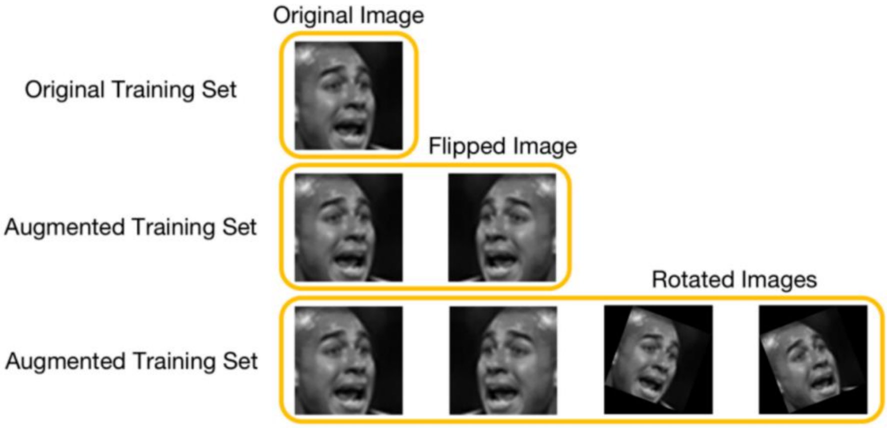


Fig. 19. The flipped image and rotated images are appended to augment the training set.

However, the best angle of rotation in data augmentation still remains to be explored. A small rotation angle makes the rotated image too similar to the original one, thus can hardly help the model acquire new informative features. On the contrary, a large rotation angle is likely to convey wrong features to the model. Therefore, the most appropriate angle of rotation should be determined by experiments.

3.6.2 Face Alignment

Face alignment is an alternative to data augmentation with a similar effect, but it does not require the augmentation of the training set. Since no more training samples are introduced, the time for model training can be greatly reduced, and the overfitting problem can also be alleviated.

Face alignment is an image pre-processing technique applied to sample images to crop and align the faces, which can be achieved by using the CNN detector and the ERT [26] based face alignment technique provided by an external library Dlib [27]. After alignment, the corresponding landmarks such as the eyes, nose and mouth of the faces will approximately be located at the same position in the image (Fig. 20). Therefore, the models can capture facial features more easily when analyzing these images and may attain a higher test accuracy.

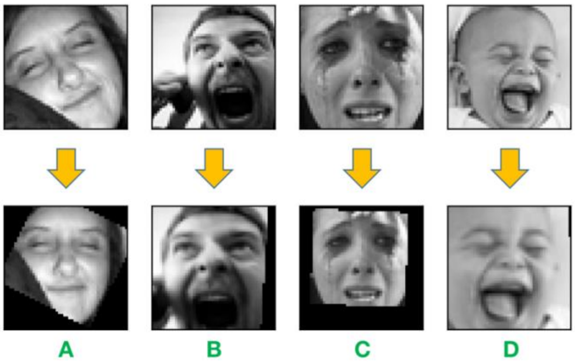


Fig. 20. The effect of face alignment. Image rotation (A), image translation (B) and image scaling (C&D) can be performed to align the faces.

3.6.3 Loss Function

To measure how close the predicted value is to the actual value, various loss functions can be applied in the model training process. Categorical cross-entropy (CCE) is a frequently adopted loss function for multi-label training. Some other popular loss functions including mean squared error (MSE), KL divergence (KLD), and cosine similarity (CS) are also implemented to train the models.

Probabilistic label drawing (PLD) is another training scheme which has shown to have higher accuracy than categorical cross-entropy loss, as indicated by a previous study [21]. This approach assigns a randomly drawn single label to each image in each epoch based on the expression distribution of the image, and the label is reassigned in each of the later training epochs. Over the long training process with numerous training epochs, the accumulated single label assigned to an image can gradually approach the true expression distribution, thus the PLD is supposed to have a similar effect as multi-label learning with ordinary loss functions.

3.6.4 Model Pre-training

Well-selected model initial weights can contribute to higher prediction accuracy, thus should be carefully determined. Appropriate initial weights can be obtained by pre-training the model using other image datasets before training and testing by FERPlus.

ImageNet is a large-scale dataset containing 14 million images of all types [28], while AffectNet is a facial image dataset containing 1 million samples [2]. These two datasets are both chosen for model pre-training. Due to memory restrictions of the device, ImageNet cannot be downloaded to train the models, and only about 300,000 samples in AffectNet can be used for training. Fortunately, Keras provides models with weights obtained from ImageNet pre-training, thus those models can be applied directly in this project.

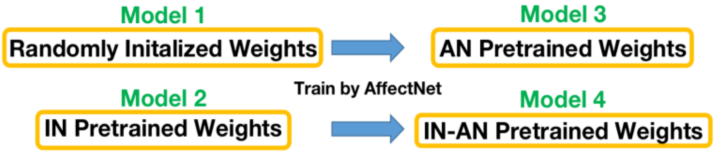


Fig. 21. The model pre-training process. A model with randomly initialized weights (Model 1) and the same model with ImageNet pre-trained weights (Model 2) are both trained by partial samples of AffectNet to generate Model 3 and Model 4. (IN: ImageNet. AN: AffectNet.)

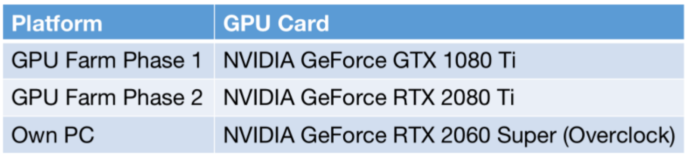
The four models with different initial weights shown in Fig. 21 will be evaluated later on to determine which dataset is the best for model pre-training.

3.7 Tools

TensorFlow, Keras and Pytorch are adopted because of their abundant build-in methods to construct and optimize the deep learning model, and to test the model performance. Matplotlib is applied for plotting. Dlib toolbox [27] is employed to implement the face alignment.

GPU is adopted for model training and testing in this project. The HKU GPU Farm Phase 1 and Phase 2 supported by the CS department and the PC of a group member are all utilized, and Table 1 illustrates their configurations.

Table 1. Hardware configurations.



Platform	GPU Card
GPU Farm Phase 1	NVIDIA GeForce GTX 1080 Ti
GPU Farm Phase 2	NVIDIA GeForce RTX 2080 Ti
Own PC	NVIDIA GeForce RTX 2060 Super (Overclock)

4 Experiments and Results

This chapter delineates the experiments conducted in this project. The experiment settings are explicated in section 4.1, and the test result of various experiments are demonstrated in the subsequent four subsections.

4.1 Experiment Setup

In the experiments in section 4.2 and section 4.3, various models were trained with different optimization techniques, and test results were collected and analyzed. The experiments in this part were based on default training and test settings, i.e., by using the original (8-hot label version) training data to train the model, and the ordinary (the randomly drawn) test data to test the model performance. Experiments in section 4.4 changed the training dataset to the N-hot label version, while the testing in section 4.5 concentrated on model performance analysis on a complex test set.

The input of the model was the pixel tensor of the facial image, with the pixel value range rescaled from [0, 255] to [0, 1], and the image size 48×48 (except DAN, which is 224×224). Adam was chosen as the optimizer for model training because of its capability to adaptively change the learning rate. The initial learning rate for most experiments was set to 1e-4 (some of the experiments required a slightly lower learning rate depending on other parameter settings). The model in most experiments was trained for 40 epochs, but DAN needed 60 epochs before reaching its highest test accuracy.

4.2 Model Selection

This section focuses on the selection of appropriate model architecture to handle FER-related tasks. The classical CNN models and the advanced models will be evaluated in the following subsections.

4.2.1 Test Result of Classical Models

Experiments were firstly conducted to find the best classical CNN models, and Fig. 22 demonstrates the test result.

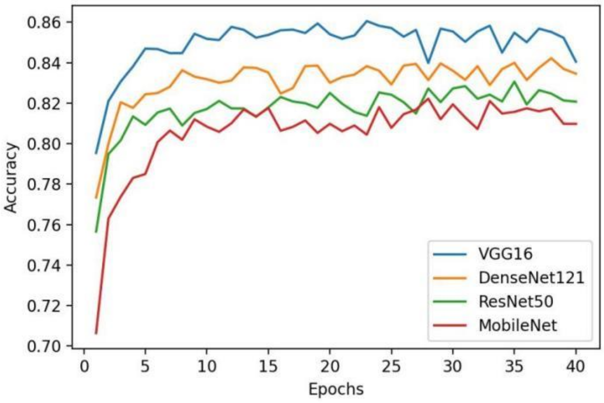


Fig. 22. Test accuracy of four classical models.

VGG16 was apparently more accurate than the other three models. Apart from the models in Fig. 22, other VGG models (VGG13 and VGG19), ResNet models

(ResNet18, ResNet34, ResNet50, etc.), DenseNet models (DenseNet169 and DenseNet201), and all of the EfficientNet models (from EfficientNet-B0 to EfficientNet-B7) were also evaluated, but none of them reached higher test accuracy than VGG16, though the accuracy of other two VGG models was close to with that of VGG16.

It was out of expectation that ResNet did not achieve similarly high accuracy as VGG did in the experiment. A possible explanation was that the images of human faces were relatively simple and were less complicated than images of other objects. Therefore, 20 layers were more than enough to fully extract the important facial features and perform the analysis. The residual blocks in ResNet and its extra layers provided little contribution to the emotion prediction, thus the performance of ResNet was not as good as that of VGG.

This outcome showcased VGG model structure’s capability to analyze facial features and generate reliable predictions of emotion. Since VGG16 turned out to be the best classical CNN model among those tested in the experiment, it was selected to be the benchmark for evaluating the advanced models.

4.2.2 Test Result of Advanced Models

Based on the same training and test setup, further experiments were implemented to investigate the performance of the three advanced models. The resulting test accuracy is presented in the bar chart shown in Fig. 23.

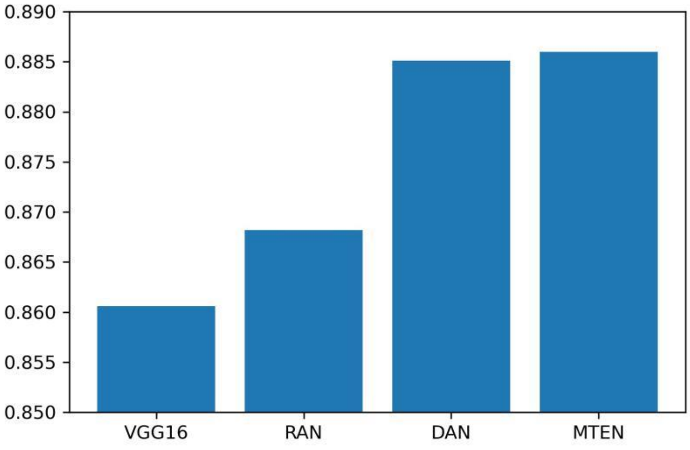


Fig. 23. Test accuracy of three advanced models compared to VGG16.

DAN and MTEN achieved remarkable test accuracy (both of which reached up to about 88.6%), while RAN was inferior to them. But noticeable improvement of these three advanced models compared with VGG16 could be observed, which was reasonable since advanced models for FER tasks employed specific structures to facilitate the analysis of facial regions, instead of targeting the general object recognition.

For the remaining part of the experiments, the three advanced networks, i.e., RAN, DAN and MTEN, together with the baseline model VGG16 would be applied to explore the effect of various factors (including the optimization approaches and the training and test settings) on different model architectures.

4.3 Model Optimization

The experiments in this section were designed to figure out whether several optimization techniques were effective or not in the FER model improvement.

4.3.1 Test Result of Data Augmentation

The positive effect of the data augmentation approach was impressive on the VGG16 model, which could be seen from the increment of test accuracy after the augmentation of the training dataset (Fig. 24). Adding mirror images to the training set dramatically increased the accuracy, and adding rotated ones further enhanced model performance. The result could be justified by the new useful features brought by the flipped and rotated training images.

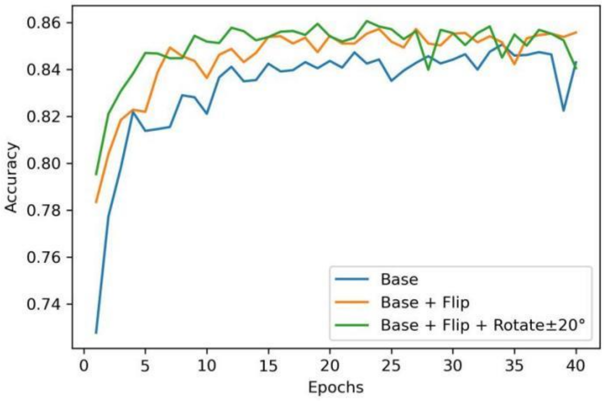


Fig. 24. The test accuracy of VGG16 trained by different training sets.
(Base: the original training image set.
Flip: horizontally flipped images.
Rotate±20°: images after rotation of 20° counterclockwise and clockwise.)

To find the best angle of rotation, test statistics of VGG16 trained by augmented training sets with diverse rotation angles, i.e., 10°, 20°, 30°, 40°, and randomly selected angles between 10° and 30° were inspected (Fig. 25). It is slightly difficult to clearly differentiate the lines in Fig. 25, thus some other statistics including the highest test accuracy in the entire training process, the 75 percentile of test accuracies of the 40 training epochs, and the number of epochs that achieved a test accuracy of higher than 85% were collected and summarized in Table 2. All these three criteria supported that 20° was the best rotation angle. The result showed that the rotation angle of human facial images could bring few new features when under 20°, but might distort the features when beyond 20°.

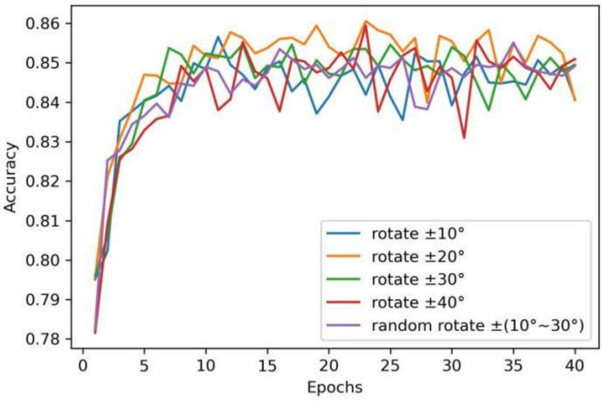
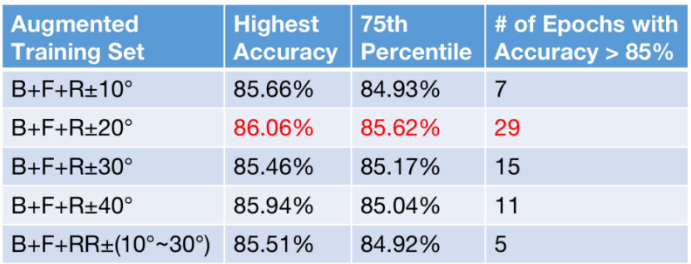


Fig. 25. The test accuracy of VGG16 trained by augmented training set (Base+Flip+Rotate) with different rotation angles.

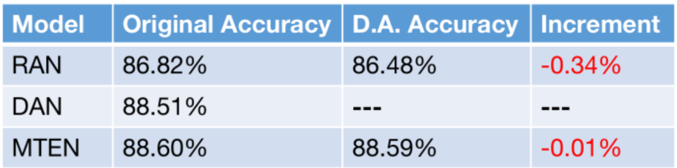
Table 2. Test statistics of VGG16 trained by augmented training set with different rotation angles. (B: Base. F: Flip. R: Rotate. RR: Random Rotate.)



Augmented Training Set	Highest Accuracy	75th Percentile	# of Epochs with Accuracy > 85%
B+F+R±10°	85.66%	84.93%	7
B+F+R±20°	86.06%	85.62%	29
B+F+R±30°	85.46%	85.17%	15
B+F+R±40°	85.94%	85.04%	11
B+F+RR±(10°~30°)	85.51%	84.92%	5

However, on the advanced model architectures, data augmentation had only a negligible effect. Table 3 displays the increment of model accuracy after the training set is augmented to Base + Flip + Rotate±20°. Since DAN had already integrated the data augmentation step into its model, there was no need to further augment its training data. The little increase (the diminutive negative increase, to be exact) of RAN's and MTEN's test accuracy could possibly be explained by their model structures. It might be because their network architectures enabled them to learn sufficient facial features from the original training set, thus the additional information introduced by augmented samples could hardly be more constructive.

Table 3. Test accuracy of advanced models trained by original training set and augmented training set. (D.A.: data augmentation.)



Model	Original Accuracy	D.A. Accuracy	Increment
RAN	86.82%	86.48%	-0.34%
DAN	88.51%	---	---
MTEN	88.60%	88.59%	-0.01%

4.3.2 Test Result of Face Alignment

The effect of face alignment failed to meet the expectation. On VGG16 model, face alignment did improve accuracy by about 0.5%, but still did not exceed what had been achieved by using data augmentation (about 1% increase). On DAN and MTEN, face alignment was not instrumental at all. RAN already applied face alignment in its model, so no additional testing on this model was required.

A possible reason that could lead to the ineffectiveness of face alignment was that, instead of augmenting the training set and forcing the model to learn various kinds of facial features, face alignment provided FER models a “comfort zone”. It aligned the input faces for the model, thus the model could lose its capability to analyze the unaligned images. However, due to some reason (to be explained in section 5.2.3), not all faces could be successfully detected and aligned by this technique. Hence, the model found it difficult to evaluate the unaligned images in the test set, and only got a moderate test accuracy.

4.3.3 Test Result of Loss Function

The choice of loss function has an obvious impact on the training process, and Fig. 26 demonstrates a summary of the test accuracy of VGG16 trained using ten selected loss functions.

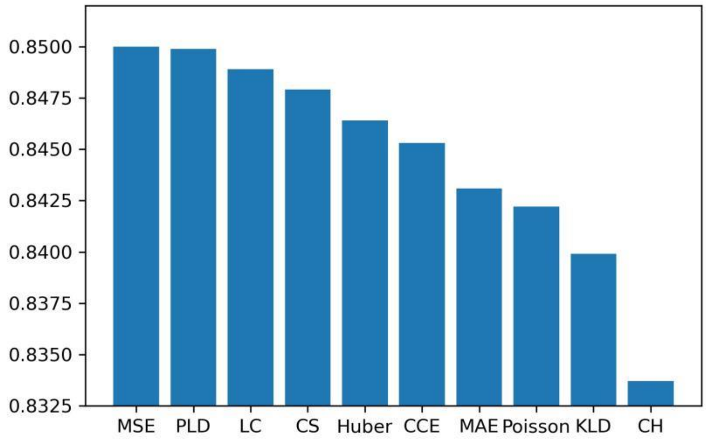


Fig. 26. Test accuracy of VGG16 trained using ten different loss functions. (MSE: mean squared error. PLD: probabilistic label drawing. LC: log cosh. CS: cosine similarity. CCE: categorical cross-entropy. MAE: mean absolute error. KLD: KL divergence. CH: categorical hinge.)

From the experiment results, MSE achieved the best accuracy (85.00%) among the ten selected functions. Although PLD was extremely close to this value, its performance on the augmented training set was not that satisfactory: when training the model using the augmented training set, MSE helped VGG16 attain 86.17% accuracy, while PLD only let the model accuracy reach 85.01%. This might be because that PLD required many epochs of training to make the accumulated single labels approach the actual label distribution, thus was sometimes incapable of rapidly learning complex features. Nonetheless, PLD still outperformed most loss functions, especially the widely-adopted loss function CCE, and this result was consistent with the findings of the previous study [21].

Considering the performance of all the selected loss functions, MSE was regarded as the best option and would be applied for the remaining experiments.

4.3.4 Test Result of Model Pre-training

As illustrated in Fig. 27, the test accuracy of VGG16 pre-trained by ImageNet surpassed that of the model without pre-training by ImageNet by approximately 2%. However, AffectNet pre-training had only a negligible effect on the model accuracy, probably due to the insufficiency of training samples (since only less than 30% of AffectNet samples were used for training as a consequence of memory restrictions).

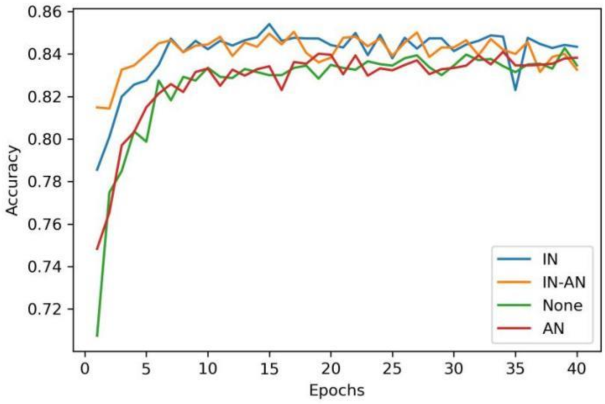


Fig. 27. Test accuracy of VGG16 pre-trained by different datasets.
(IN: ImageNet. AN: AffectNet.
IN-AN: pre-trained by ImageNet first and then AffectNet. None: randomly initialized weights)

Since deep learning software such as Keras and Pytorch provides model weights pre-trained using ImageNet only for those classical models, the advanced models in this project could not be pre-trained by ImageNet. Since AffectNet showed limited improvement to model performance on VGG16, it was not applied to pre-train the advanced models.

4.4 Test Result of N-hot Label Training

Four selected models, VGG16, RAN, DAN, and MTEN were trained respectively by the N-hot label training set (N from 1 to 8), and their highest accuracy was measured (Fig. 28).

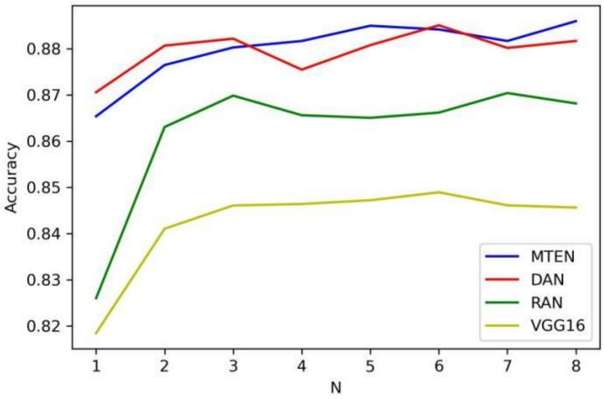


Fig. 28. Test accuracy of four selected models trained by N-hot label training set.

For the four models, the accuracy increments from training by 1-hot label training data to 2-hot label, and from 2 to 3 were all relatively significant. This phenomenon confirmed that the minor attributes in the emotion vector were still somehow related to certain features of facial expression, thus the inclusion of these additional labels did contribute to the model learning.

However, the increase of model accuracy was not obvious when N grew larger than three. This was also anticipated, since 90% of FERPlus image samples had the complexity level no more than 3 (as indicated in Fig. 17), thus most sample data in 3-hot label training set to 8-hot label training set should be the same. Therefore, the information gain by increasing N from 3 to 8 was relatively limited. The subtle fluctuation of accuracy at this range could be explained by the randomness of the training process.

Due to the great time and tremendous human effort required to manually label the facial images, few existing datasets attempted to assign more than a single label to the images. Notwithstanding, the increase in accuracy of models trained by multi-label dataset was still notable compared to models trained by single-label data (an accuracy increase of about 2% to 4% could be observed for the four models). This test result suggested that future researchers could enhance their FER model performance if they could train their models by facial image datasets with multi-label emotions.

4.5 Test Result of Complex Tests

Finally, experiments were conducted to investigate how the complexity of test data affects model performance (Fig. 29).

For each model trained by each of the training sets (N from 1 to 8), the test accuracy dramatically decreased (for about 20% each) when tested by the complex test set, which was consistent with what had been anticipated.

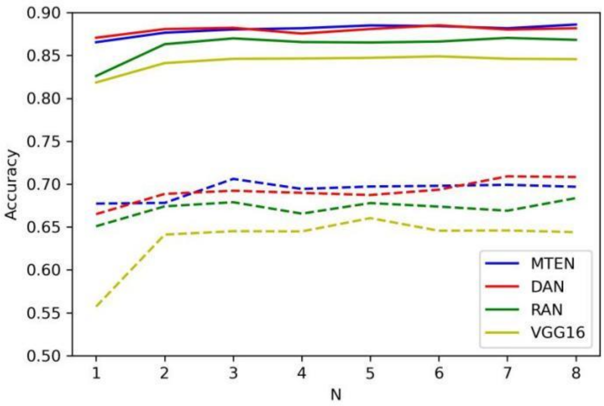


Fig. 29. Test accuracy of four selected models trained by N-hot label training set, and tested by ordinary test set (solid lines) and complex test set (dashed lines) respectively.

It could be observed that the test accuracy of VGG16 is inordinately low (only about 55%) when trained by single-label data and tested by complex samples. This could serve as convincing proof that some traditional model architectures like VGG were incapable of generating accurate emotion predictions of complicated facial expressions, especially when the model was trained by single-label data, which lacked the vital information the model must learn in order to know more about complex faces. Since many recent FER models are still developed based on classical structures such as VGG and ResNet, the model performance might suffer if the model is trained by single-label training data.

The three advanced models, however, did not suffer from single-label training in complex tests as much as VGG16 did (with their test accuracy when N = 1 maintained at about 65% to 70%). This result illustrated that these model architectures had somehow managed to extract sufficient information about facial features even when given single-label training data, thus their performance did not suffer as much as VGG16. The rigorous explanations for this phenomenon might require further investigation.

5 Discussion

This chapter lists the difficulties encountered and the limitations of the project in section 5.1 and 5.2. Some future work based on the findings of this project is elaborated in section 5.3.

5.1 Difficulties

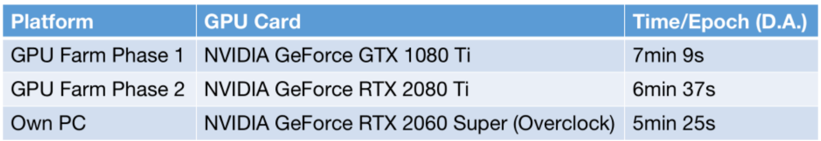
Three main difficulties of the project are discussed in the subsequent sections: the enduring training process, limited memory space, and optimization hardships.

5.1.1 Time-consuming Training Process

The primary difficulty of the implementation of the experiments derives from the time-consuming training process. The experiments of this project are mainly run on three sets of hardware: HKU GPU Farm phase 1 and phase 2, and a group member's personal computer. Despite the use of advanced graphics cards on three sets of hardware, the time-consuming training process is still a problem.

For instance, given 4 times the size of the original training set as the training sample size (for the data augmentation purpose), all three hardware platforms need about 6 minutes to complete each epoch to train a VGG16 model (Table 4). Some other model architectures, such as DenseNet, even require a much longer time. In this case, it would take weeks to test all combinations of different factors, even if the equipment keeps running all day. To get the results and compare the model performance more efficiently, only factors that may most likely result in the highest accuracy are selected and tested. Furthermore, due to the variation of test accuracy in each experiment on the same model, finding the best model is extremely challenging. Experiments of each model are conducted only once to save time.

Table 4. Training time per epoch with augmented training set
(4 times original training sample size) on VGG16 model. (D.A.: data augmentation.)



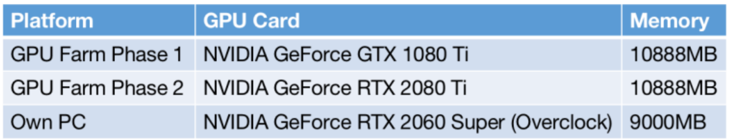
Platform	GPU Card	Time/Epoch (D.A.)
GPU Farm Phase 1	NVIDIA GeForce GTX 1080 Ti	7min 9s
GPU Farm Phase 2	NVIDIA GeForce RTX 2080 Ti	6min 37s
Own PC	NVIDIA GeForce RTX 2060 Super (Overclock)	5min 25s

5.1.2 Limited GPU Memory

Apart from time, space limitation is another issue encountered in this project. Due to the memory restrictions of GPU Farm and the PC (Table 5), some experiment approaches become impossible to carry out in this project. For example, large-scale image datasets including ImageNet (150 GB) and AffectNet (120 GB) cannot be entirely loaded for model pre-training. As a compromise, only partial images from AffectNet are loaded to pre-train the model. However, pre-training using insufficient samples seems to be ineffective, as illustrated in section 4.3.4.

Furthermore, the memory constraint also prevents batch size from being further increased, which hampers the attempt to save the model training time by increasing the size of a batch.

Table 5. GPU memory of the hardware platforms.



Platform	GPU Card	Memory
GPU Farm Phase 1	NVIDIA GeForce GTX 1080 Ti	10888MB
GPU Farm Phase 2	NVIDIA GeForce RTX 2080 Ti	10888MB
Own PC	NVIDIA GeForce RTX 2060 Super (Overclock)	9000MB

5.1.3 Hardships in Model Optimization

During the implementation of some previously developed models, some model training details such as parameter settings are not published, likely leading to worse performance of the model. For instance, although the code for RAN implementation is publicly available, some essential sections such as the data processing details are not given by the authors. As a consequence, the test accuracy of RAN (86.82%) is significantly lower than what the authors have declared (89.16%) [6].

Some other advanced FER models, like FER-VT and LResNet50E-IR [29] which have achieved the first and the second highest test accuracy on FERPlus dataset (90.04% and 89.26% respectively), do not have their code shared with the community, thus the implementation and optimization of these model architectures are indeed challenging.

5.2 Limitations

Some limitations and potential areas of improvement of this project are specified in the following six subsections.

5.2.1 Validation Bias

The selection of the best set of hyperparameters is based on the validation accuracy in each experiment, and the validation bias becomes a problem. For all previously mentioned experiments with test results, the validation samples are either the ordinary test set or the complex test set, either of which is selected before all the tests are carried out and never altered again, thus the best model settings are likely to be biased towards these two fixed test sets. Although methods like cross validation can possibly mitigate this issue, training and testing by multiple sets of data take too much time, thus this approach is not adopted.

Also due to the limited amount of time and GPU memory, model training and testing are only conducted based on a single dataset, i.e., FERPlus, thus the robustness of the models might compromise. First, the universality of the model may suffer as a result of sample insufficiency. Lack of diversity is one of the major shortcomings of the model, since extra data from other sources including facial images of people expanding all regions and ages can hardly all be fed to the model. Second, incompatibility with colored or high-resolution pictures is another concern. The FER system developed by the project may underperform in recognizing colored or high-resolution emotions, because FERPlus dataset contains only grayscale and photos with low resolution (48×48 pixels). Both these factors cause the developed model to be less reliable regarding its real-life performance.

5.2.2 Deficiency of Data Augmentation

Data augmentation is actually an expansive approach for model training. Although the positive effect it brings is noticeable (e.g., it increases the accuracy of VGG16 by more than 1%), the extra time that is required for the model to learn the additional features from the augmented samples is sometimes regarded as a severe issue (the training time doubles if the training set includes flipped images, and further doubles if rotated images are introduced to the training set). This is also the reason why some FER researchers would rather choose face alignment as an alternative to data augmentation.

Data augmentation sometimes brings few useful facial feature information to the dataset with regard to part of the image samples. For example, the horizontal flip does not greatly change the image, if the original image itself is already close to being symmetric (Fig. 30). Appending the flipped images of this kind of samples leads to redundancy of training data, and might result in problems like overfitting.

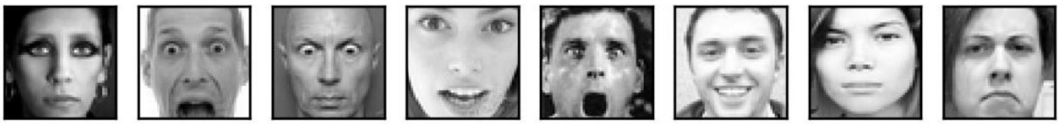


Fig. 30. Image samples in FERPlus that are nearly symmetric.
Horizontal flipping has no significant effect on these images.

Apart from horizontal flipping, image rotation sometimes is also not a proper operation on some of the image samples. Some faces in the dataset might have already been skewed, thus further image rotation, even of only 20°, might cause the face to nearly lie flat (Fig. 31). This over-rotation distorts the facial feature that the model should learn, and brings additional noise to the dataset (since the model is not supposed to evaluate these distorted faces).

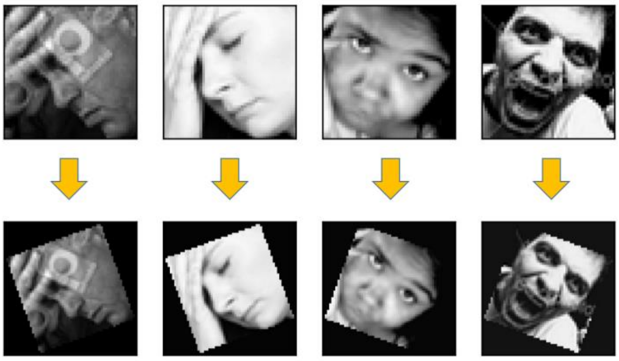


Fig. 31. Oblique faces in FERPlus before and after 20° counterclockwise (left two) or clockwise (right two) rotation. Images after rotation do not show normal facial features.

Although the experimental results indicate that 20° should be an appropriate angle of rotation, this is from the perspective of the entire general dataset. Indeed, most of the facial expressions in FERPlus are relatively upright (unlike the ones illustrated in Fig. 31), thus 20° rotation on these samples is beneficial for the data augmentation. But for those minorities with skew faces, the rotation had better not be forced on them.

Therefore, it would be better if the face skewness could be measured beforehand by applying some other model, so that the decision can be made afterward to determine whether to rotate these images or not. However, this idea is still not implemented in this project, thus the current data augmentation technique is not that intelligent in this sense.

5.2.3 Ineffectiveness of Face Alignment

Due to the aforementioned defects of the data augmentation approach, face alignment seems to be a superior alternative method: face alignment applies rotation according to the original face's skewness automatically; furthermore, this technique does not require more training data, thus the time spent for model training will not be increased.

However, face alignment also has its weakness: it cannot be applied to all the image samples. The face alignment model utilized in this project is from the Dlib toolbox. A CNN detector is applied for the model to identify the location of your facial landmarks such as eyes and mouth, based on which the alignment operation can be carried out. But a problem occurs when the landmarks cannot be detected. Some samples in FERPlus (Fig. 32), either because they are profiles with one side of the face hidden, or because of occlusions, have their facial landmarks concealed from view, hindering the detector from discovering landmark locations. Thus the landmark-based face alignment cannot be implemented on these images.

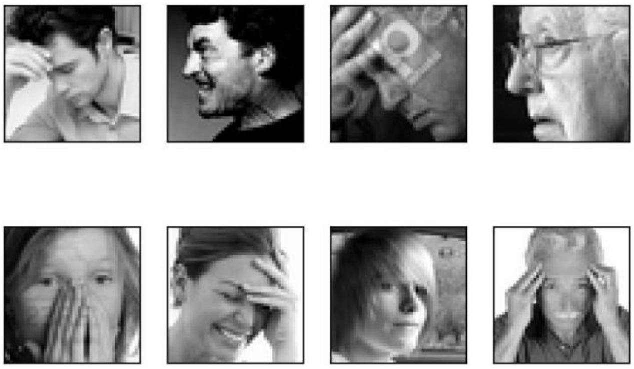


Fig. 32. Facial landmarks are hidden in some FERPlus images.
Top row: profiles of a person. Bottom row: faces with occlusion.

5.2.4 Sample Imbalance

The image samples in FERPlus are highly imbalanced with regard to their emotions. Fig. 33 illustrates the distribution of the votes assigned to different emotion categories. It can be observed that neutral and happiness constitute more than 50% of the emotion proportion of FERPlus image samples, while disgust, fear, and contempt make up merely no more than 10% of total votes. The class imbalance problem directly causes the model constructed to be incapable of accurately analyzing images with emotions such as contempt, since the training set contains few samples in these categories.

Even if some other databases rather than FERPlus are selected for model training and testing, the issue is still unsolved, since images accessible on the Internet follow a similar distribution, with only a small proportion of samples exhibiting these relatively rare expressions such as disgust and contempt.

Another sample imbalance is the uneven distribution of images with various expression complexity, which can be noticed in Fig. 17. The majority of facial images (up to 90%) have a complexity level of no more than 3, indicating that in most scenarios, three emotions are more than adequate to describe a facial expression.

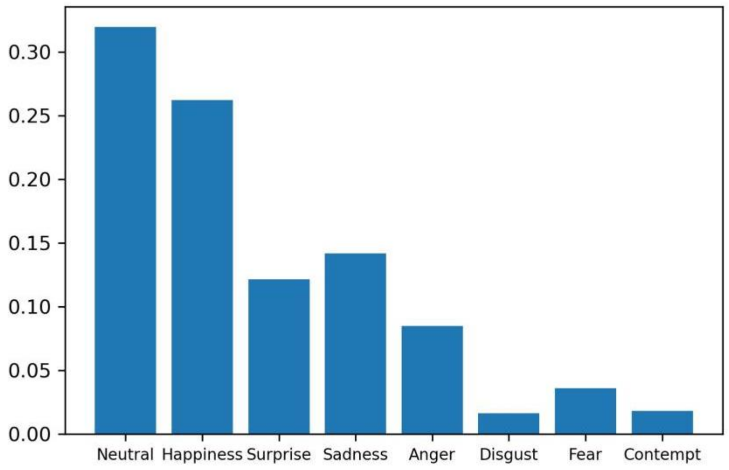


Fig. 33. Expression distribution of FERPlus after data cleaning. Calculated by adding all the normalized emotion vectors after data cleaning, and performing a normalization again.

This imbalance leads to the insufficiency of samples with complex expressions, hence the model obtained tends to be highly inaccurate when evaluating images of this type (test results already shown in section 4.5). This issue can be alleviated by adding complex expression images from external sources to the training set to let the model learn more about those complex features. However, at least two concerns still exist.

First, since most of the current datasets adopt a single-emotion labeling scheme, the complexity of expressions can hardly be assessed for images from those external sources. Therefore, there is barely any approach for the potentially complex images to be identified and collected from those datasets, not to mention to classify them according to the level of complexity.

Second, even if a few other multi-label datasets like RAF-ML are examined, only a small number of complex samples can be obtained. Theoretically speaking, the images from other datasets should also comply with a similar distribution of complexity (i.e., with few images having a high complexity level), hence, the contribution that these additional samples can provide to the model is still limited.

5.2.5 Inaccuracy of Expression Distribution

As indicated by the previous study [21], 10 taggers, in most scenarios, can be more than sufficient to give a relatively good approximation to the ground truth emotion label of a person’s facial expression (Fig. 12). However, this result only illustrates that the primary emotion can be accurately estimated by the 10 taggers. What about those secondary or tertiary emotions? It is still not clear whether those minor emotions are correctly interpreted by the merely 10 annotators.

Since all the emotion examiners are asked to assign an emotion label to the expression which they think is most likely, instead of a distribution, it can be probable in some cases that all the 10 taggers assign the dominant emotion to a facial image, while leaving those secondary emotions unassigned. This problem generally exists in most multi-label datasets, but the increase of the number of annotators can somewhat mitigate this issue. Whether 10 taggers are enough to give an accurate estimation of the true expression distribution remains to be further investigated.

It can also be possible in some scenarios that some annotator, accidentally due to some reason, gives a label that is far from the ground truth, i.e., this label is the outlier of the true emotion distribution for the image. But data cleaning on the expression distribution is an arduous task, since outliers cannot be easily recognized. For instance, if some image receives 9 votes on an emotion category, and receives 1 vote on another emotion, should this single vote be treated as an outlier? It is clear that the elimination of those outliers should be able to improve the data quality. However, the currently adopted approach is still to retain all the votes, since the opinions of the minorities still truly reflect the minor emotions of the person in most cases. But this might induce troubles since the outliers can never get removed.

5.2.6 Expression Intensity

Since FERPlus only provides the voting result and does not offer any information about the intensity of an emotion, how strong an emotion is can hardly be measured. In this project, the intensity of an emotion is approximated by the proportion of votes given to the image. This approach of estimating expression intensity has to be improved.

Fig. 34 gives an illustration of facial emotions with different levels of intensity. The 12 sample pictures are selected from FERPlus, upon whose emotion all the 10 annotators reach a consensus. That is to say, all the 10 votes are assigned to 1 category only, and their final emotion vector is one-hot label. However, it can be noticed that the images on the upper row exhibit a stronger emotion than those on the bottom row, but the emotion distributions assigned to them are identical (if their estimated emotions are of the same type) and cannot reflect the difference of intensity. If the intensity value of an emotion can be measured and taken into consideration in the model construction process, it is possible that more robust FER models can be produced.

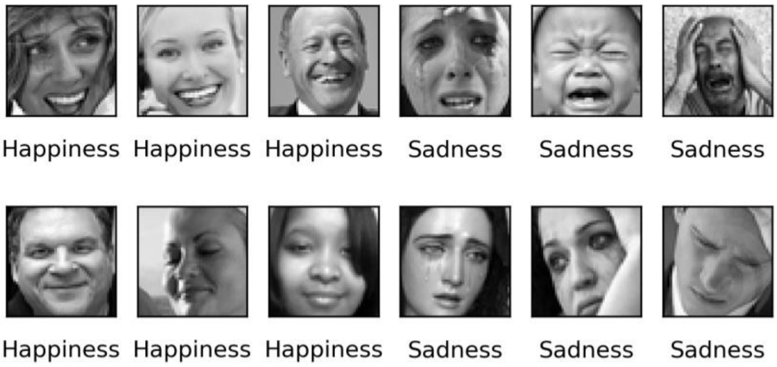


Fig. 34. Facial expressions with different intensities.
The first row shows expressions with higher intensity and the second row with lower intensity.

5.3 Future Work

Several feasible aspects for the development of FER technology are proposed in the subsequent five subsections, and these areas might enlighten the direction for further research.

5.3.1 Model Pre-training Based on AffectNet

The model pre-training by partial AffectNet in this project seems to be ineffective, probably due to the insufficiency of training samples or the limited amount of training time. Further experiments can be conducted in the future to test whether pre-training by large-scale face datasets such as AffectNet for a sufficiently long time period can improve the model performance. Since typical facial features are learned by the model in the pre-training process, it should be hopeful that the model can obtain higher accuracy.

5.3.2 Adaptive Data Augmentation

Because of the disadvantages of data augmentation mentioned in section 5.2.2, this technique is not adopted in many recent FER projects. But this method does show a significant effect on classical models like VGG16. Since some latest FER models such as RAN, DAN, and FER-VT still apply traditional architectures such as VGG and ResNet as their backbone models [6, 19, 25], data augmentation is still a method with great potential.

The data augmentation process can be further optimized using the following strategy. Since existing tools such as CNN detector [27] can capture the location of facial landmarks, they can be applied to calculate the inclination of the face. If the face is found to be nearly horizontally symmetric (i.e., with a small inclination), the horizontal flipping does not need to be carried out at all, helping to reduce the redundancy of the augmented training set (image (b) in Fig. 35). If the face has been recognized as tilted towards one direction, then there is no need to perform image rotation of that image towards the same direction, avoiding excessive rotation (image (c) in Fig. 35). If neither of the above two conditions is met, ordinary data augmentation is applied (image (a) in Fig. 35). By applying this adaptive data augmentation, the augmented training set will have a smaller size, less redundancy, and better quality, which might contribute to model training.

However, the definitions of "close to symmetric" and "being tilted" are both unclear. In theory, two thresholds of inclination should be determined: X_1 and X_2 ($X_1 < X_2$). If the inclination of the face does not exceed X_1, the face should be regarded as "close to symmetric". And if the inclination is higher than X_2, the face is "tilted". Further experiments can be conducted to determine the best values for X_1 and X_2, based on the performance of the augmented training set obtained by applying X_1 and X_2. Note that the best angle of rotation might no longer be 20° and might have to be revised if adaptive data augmentation is adopted.

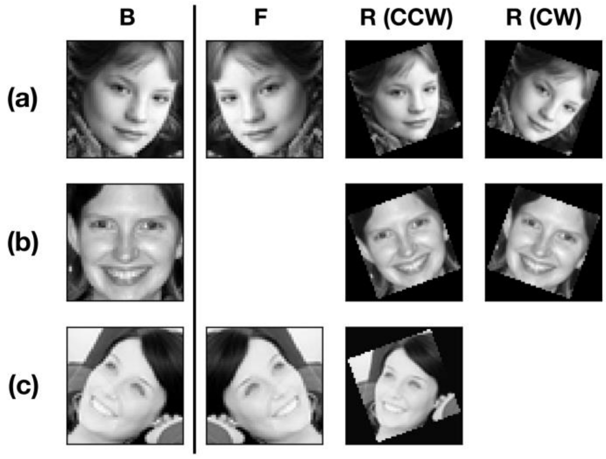


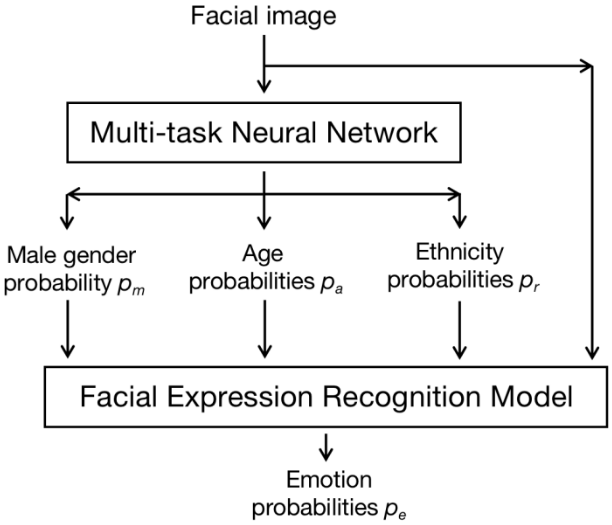
Fig. 35. Examples of adaptive data augmentation.
The original and augmented images are at the L.H.S and R.H.S. of the line respectively.
(B: base image. F: flipped image. R: rotated image. CW, CCW: clockwise, counterclockwise)
Image (a): face inclination between X_1 and X_2, perform full augmentation.
Image (b): face inclination below X_1, perform only rotation.
Image (c): face inclination above X_2, perform flipping and rotation to only one direction.

5.3.3 Emotion Learning Based on Multiple Attributes

Models like multi-task neural network [5] have already been developed to give predictions on not only people's emotion, but also other factors such as their race, gender, and age (introduced in section 2.2.2). However, limited research is available on the topic of how to utilize this newly obtained information to perform more accurate emotion predictions.

Although those additional attributes seem to be irrelevant to people's inner feelings, they might be a strong indicator of people's facial features. People from different races or of different ages may exhibit their inner emotions in a different way, due to reasons like bone structures. For example, the facial features of babies (Fig. 36) tend to be distinctive and vary from those of adults. And people in the same group (of race, age, etc.) might have a higher similarity with respect to certain facial features.

Fig. 36. Image samples of baby faces in FERPlus.



Facial image
Multi-task Neural Network
Male gender probability p_m
Age probabilities p_a
Ethnicity probabilities p_r
Facial Expression Recognition Model
Emotion probabilities p_e

Fig. 37. A proposed mode architecture to predict emotion based on additional attributes related to the person.

If the model learns the facial features together with the extra attributes of the person such as ethnicity and age, it might gain a better performance when evaluating faces from a specific race or age group. Fig. 37 demonstrates a possible model structure based on multi-task neural network which takes into consideration gender, race, and age information when predicting emotions. This extra-attribute-based emotion learning might be a potential area of study.

5.3.4 Dataset Construction with Balanced Distribution

As discussed in section 5.2.4, most currently available facial image datasets suffer from an imbalance of sample distribution: the images with complex expressions or rare expressions (disgust and contempt) constitute only a small proportion of the dataset. This makes the model predictions on those expressions extremely unreliable.

It is likely that, if the image samples are collected following a relatively even distribution, the problem could be alleviated. Therefore, future FER researchers can try to collect image data from various sources (including existing databases), and form a new dataset with an approximately similar number of samples in each emotion category, and of different levels of complexity. This could probably improve the model's performance, especially when evaluating complex expressions and expressions with minor emotions (disgust and contempt).

5.3.5 Multi-label Dataset Construction

Since only a few facial image databases such as FERPlus, RAF-DB and RAF-ML adopt a multi-emotion labeling scheme, and most of these multi-label datasets are relatively small-scale compared to other databases like AffectNet or ImageNet, the data samples that can refer to in the study related to multi-label FER are indeed limited. Although it is a laborious task to manually label a large-scale dataset, it is worthwhile if a larger facial image database of multi-label emotions can be created and used for further research.

Considering the time that it might take to gather image samples from various sources, it is more feasible to relabel some existing dataset, like AffectNet (just as how FERPlus is obtained from FER2013). The increase of sample size of multi-label data can promote the construction of better models based on multi-label learning, and accelerate the research process of multi-label FER.

During the database establishment, some alternative labeling schemes can be attempted. The traditional approach is to ask each tagger to assign a single label to a facial image. However, some minor feelings conveyed by the expression may be lost if every annotator thinks they are not the dominant emotion and fails to report them. Increasing the number of taggers is certainly a possible solution. Another way is to give each tagger multiple votes, thus each of them can assign a distribution to an image according to their perception of the expression. This will make the ultimate distribution give more weight to the emotions which originally received little attention, but is still a component of a complex expression.

Furthermore, the intensity value of each emotion can be estimated and assigned by the annotators to each image. If this is achieved, the target FER model output can be transformed from the probability distribution of emotions to the intensity scores of emotions. In this way, the information of emotional intensity is learned by the model, which may help the model to differentiate the emotions with various intensities and thus to better interpret the expression. Table 6 provides an example to demonstrate how the intensity scores differ from the probability distribution. Note that the emotion intensity vector does not need to be normalized, since it does not represent probabilities anymore.

Table 6. The emotion distribution vectors and intensity vectors assigned to two images, both of which receive 10 votes on happiness in FERPlus. The intensity of emotion can be reflected in the intensity vector, but cannot be reflected in the probability distribution vector.

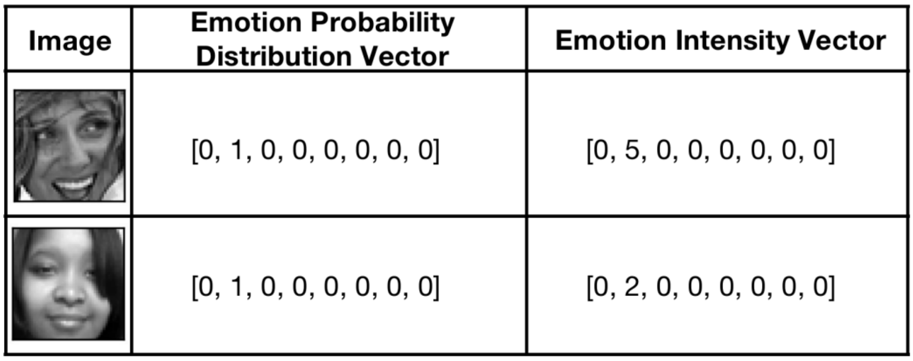


Image	Emotion Probability Distribution Vector	Emotion Intensity Vector
	[0, 1, 0, 0, 0, 0, 0, 0]	[0, 5, 0, 0, 0, 0, 0, 0]
	[0, 1, 0, 0, 0, 0, 0, 0]	[0, 2, 0, 0, 0, 0, 0, 0]

6. Conclusion

Facial expression recognition (FER) technology is a popular field of study in computer vision and deep learning, which aims to predict human emotion based on their facial images. Convolutional neural network (CNN) is a class of deep learning networks designed to extract image features and perform image analysis, thus can be utilized in FER-related tasks to predict the emotion of a face.

A CNN model needs to be trained by a facial image dataset in order to perform accurate emotion predictions. Most of the publicly available facial image databases such as AffectNet, SFEW, FER2013 adopt the single-emotion labeling scheme, i.e., only one emotion tag is given to each image in the dataset. This could have possibly become a limiting factor for researchers to develop models of higher prediction accuracy. A previous study has revealed that many human expressions are of great complexity, and can hardly be summarized by only a single sort of emotion. Thus the single-label scheme of these datasets suffers from information loss, which could weaken the performance of the models trained by their data.

This project focuses on the investigation of FER technology. Deep learning FER models are constructed and optimized by various approaches, and their test accuracy is measured. To explore the advantages of multi-label learning, the improvement of model performance trained by multi-label data compared to single-label data is evaluated.

The dataset selected for model training and testing in this project is FERPlus, which is a collection of approximately 35,000 grayscale facial images, each labeled by 10 taggers into one of 10 emotion categories: neutral, happiness, surprise, sadness, anger, disgust, fear, contempt, unknown, and not a face (N.F.). Data cleaning is carried out to remove samples of unknown and N.F. types, and test set (with ordinary or complex version) and training set (with N-hot label version) are derived.

Several classical CNN models (VGG, ResNet, DenseNet, MobileNet, etc.) and three advanced CNN-based FER models (RAN, DAN, MTEN) are implemented and compared to each other. VGG16 turns out to be the best model among the classical ones (with a test accuracy of 86.06%), and the three advanced models all showcased better performance than VGG16 (86.82%, 88.51% and 88.60% for RAN, DAN and MTEN respectively).

Some model optimization techniques are applied to evaluate their effectiveness. Data augmentation enhanced the performance of VGG16, and the best composition of the augmented training set is found to be Base + Flip + Rotate±20°. Although face alignment does increase the test accuracy of VGG16, its effect is not as significant as data augmentation. However, both these two techniques provide negligible improvement for the advanced models. MSE turns out to be the best loss function for model training, and ImageNet appears to be a better dataset for pre-training than AffectNet.

The N-hot label training illustrates that multi-label learning indeed outperforms the single-label learning, increasing the model accuracy by 2% to 4%. As the number of hot labels N of training data is increased from 1 to 3, the rise of test accuracy is dramatic for all the models. However, the accuracy increase is not as noticeable when N grows larger than 3.

When the models are tested by the complex test set, most of the test accuracy decreased by about 20%. VGG16 suffers the most from single-label training (test accuracy dropped to 55%), while the advanced models seem to have managed to learn sufficient information from single-label data, thus their performance on single-label training does not deteriorate as much as VGG16.

Some challenges of this project include the time-consuming training process and limited GPU memory. The former one restricts the attempt to implement more experiments with different setups, and the latter one makes it impossible to load large-scale datasets such as ImageNet and the entire AffectNet for model pre-training. Besides, since some source codes of past models are not available, some difficulties exist to optimize the model to the desired accuracy.

The project might have the following six aspects of limitations. First, due to the time constraint, experiments are not conducted multiple times or based on more general settings, and techniques like cross validation are not adopted either, likely causing the test results to be less reliable. Second, the data augmentation method applied currently might induce data redundancy and wrong features. Third, the face alignment technique is ineffective on some of the samples, due to reasons like image occlusions. Fourth, the data samples in FERPlus are highly imbalanced with regard to emotion classes and complexity levels, making the prediction of rare emotions and complex expressions extremely challenging. Fifth, the expression distribution derived from the voting result of 10 annotators may not be an accurate estimation of the ground truth. Sixth, the intensity of an emotion is not reflected in the expression probability distribution.

Future research can be carried out by considering the subsequent five perspectives. First, AffectNet can be further tested to determine whether it could serve as an appropriate dataset for model pre-training. Second, an adaptive data augmentation can be tried out to reduce the redundancy of the augmented dataset and avoid excessive rotation of images. Third, FER models based on not only the facial image but also other attributes such as age and ethnicity can be devised. Fourth, attempts can be made to construct a facial image dataset with relatively balanced samples with regard to emotion categories and complexities. Fifth, multi-label datasets can be created by allowing each annotator to assign multiple tags to an image, or with an intensity value marked for each emotion.

References

[1] B. Martinez and M. F. Valstar, “Advances, challenges, and opportunities in automatic facial expression recognition,” in Advances in Face Detection and Facial Image Analysis, Cham: Springer International Publishing, 2016, pp. 63–100.

[2] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “AffectNet: A database for facial expression, valence, and arousal computing in the wild,” IEEE Trans. Affect. Comput., vol. 10, no. 1, pp. 18–31, 2019.

[3] S. Li and W. Deng, “Deep Facial Expression Recognition: A Survey,” IEEE Trans. Affect. Comput., pp. 1–1, 2020.

[4] L. Wang, S. Xu, X. Wang, and Q. Zhu, “Eavesdrop the composition proportion of training labels in federated learning,” arXiv [cs.LG], 2019.

[5] A. V. Savchenko, “Facial expression and attributes recognition based on multi-task learning of lightweight neural networks,” in 2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY), 2021.

[6] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, “Region attention networks for pose and occlusion robust Facial Expression Recognition,” arXiv [cs.CV], 2019.

[7] I. J. Goodfellow et al., “Challenges in Representation Learning: A report on three machine learning contests,” arXiv [stat.ML], 2013.

[8] S. Du, Y. Tao, and A. M. Martinez, “Compound facial expressions of emotion,” PNAS, 15-Apr-2014. [Online]. Available: https://www.pnas.org/content/111/15/E1454. [Accessed: 17- Apr- 2022].

[9] S. Li and W. Deng, “Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning,” Int. J. Comput. Vis., vol. 127, no. 6–7, pp. 884–906, 2019.

[10] “TechDispatch #1/2021 - facial Emotion Recognition,” European Data Protection Supervisor. [Online]. Available: https://edps.europa.eu/data-protection/our-work/publications/techdispatch/techdispatch-12021-facial-emotion-recognition_en. [Accessed: 17-Apr-2022].

[11] S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” in 2017 International Conference on Engineering and Technology (ICET), 2017.

[12] "A Comprehensive Guide to Convolutional Neural Networks—the ELI5 way", Medium, 2021. [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53. [Accessed: 17-Apr-2022].

[13] Anonymous, “Real Emotion Seeker With Manually-Designed Teacher For Facial Expression Recognition,” unpublished.

[14] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv [cs.CV], 2014.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” arXiv [cs.CV], 2015.

[16] A. G. Howard et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” arXiv [cs.CV], 2017.

[17] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional Neural Networks,” arXiv [cs.LG], 2019.

[18] D. Han, S. Yun, B. Heo, and Y. Yoo, “Rethinking channel dimensions for efficient model design,” arXiv [cs.CV], 2020.

[19] Q. Huang, C. Huang, X. Wang, and F. Jiang, “Facial expression recognition with grid-wise attention and visual transformer,” Inf. Sci. (Ny), vol. 580, pp. 35–54, 2021.

[20] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv [cs.CV], 2020.

[21] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, “Training deep networks for facial expression recognition with crowd-sourced label distribution,” arXiv [cs.CV], 2016.

[22] S. Li and W. Deng, “Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition,” IEEE Trans. Image Process., vol. 28, no. 1, pp. 356–370, 2019.

[23] "GitHub - microsoft/FERPlus: This is the FER+ new label annotations for the Emotion FER dataset.", GitHub, 2021. [Online]. Available: https://github.com/microsoft/FERPlus. [Accessed: 17-Apr-2022].

[24] S.-H. Wang and Y.-D. Zhang, “DenseNet-201-based deep neural network with composite learning factor and precomputation for multiple sclerosis classification,” ACM trans. multimed. comput. commun. appl., vol. 16, no. 2s, pp. 1–19, 2020.

[25] Z. Wen, W. Lin, T. Wang, and G. Xu, “Distract your attention: Multi-head cross Attention Network for facial expression recognition,” arXiv [cs.CV], 2021.

[26] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[27] “dlib C++ Library,” Dlib.net. [Online]. Available: http://dlib.net. [Accessed: 17-Apr-2022].

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[29] H. Zhou et al., “Exploring emotion features and fusion strategies for audio-video emotion recognition,” arXiv [cs.CV], 2020.