Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
14/2 (2021), 127-135. DOI: http://dx.doi.org/10.21609/jiki.v14i2.968

Facial Expression Recognition using Residual Convnet with Image Augmentations

Fadhil Yusuf Rahadika[1], Novanto Yudistira[2], Yuita Arum Sari[3]

[1,2,3]Informatics Engineering, Faculty of Computer Science, University of Brawijaya
Jl. Veteran No. 8, Kota Malang, 65145, Indonesia

Email: [1]fadhilyusuf27@gmail.com, [2]yudistira@ub.ac.id, [3]yuita@ub.ac.id

Abstract

During the COVID-19 pandemic, many offline activities are turned into online activities via video meetings to prevent the spread of the COVID-19 virus. In the online video meeting, some micro-interactions are missing when compared to direct social interactions. The use of machines to assist facial expression recognition in online video meetings is expected to increase understanding of the interactions among users. Many studies have shown that CNN-based neural networks are quite effective and accurate in image classification. In this study, some open facial expression datasets were used to train CNN-based neural networks with a total number of training data of 342,497 images. This study gets the best results using ResNet-50 architecture with Mish activation function and Accuracy Booster Plus block. This architecture is trained using the Ranger and Gradient Centralization optimization method for 60000 steps with a batch size of 256. The best results from the training result in accuracy of AffectNet validation data of 0.5972, FERPlus validation data of 0.8636, FERPlus test data of 0.8488, and RAF-DB test data of 0.8879. From this study, the proposed method outperformed plain ResNet in all test scenarios without transfer learning, and there is a potential for better performance with the pre-training model. The code is available at https://github.com/yusufrahadika/facial-expressions-essay.

Keywords: *facial expression recognition, CNN, ResNet, Mish, Accuracy Booster Plus*

1. Introduction

Human behavior recognition is one of the growing research topics in computer vision and pattern recognition. Human behavior recognition is usually applied in machine learning to monitoring human activities and getting insight from them [1]. The behavioral examination can help solve many problems in indoor as well as outdoor surveillance systems. The numbers of video surveillance systems have been increasing every day to monitor, track and analyze the behaviors in different areas [2]. There are several applications of human behavior detection. Some of them are motion detection and facial expression recognition. Facial expression analysis is one of the most prominent clues to determine the behavior of an individual. However, it is very challenging due to many variations in face poses, illuminations, and different facial tones [3]. Facial expression recognition itself can be applied using images or videos that are extracted into images. Emotional facial images are direct/indirectly associated with other human behavior such as kindness, decision-making, awareness, memory, and learning. This emotion can be read mainly through facial emotion in an efficient way [4].

Facial expression recognition is a technique to understand human emotions from expressions shown as a reaction to something that occurs from the environment. In this digital world, facial expression recognition can be widely applied. For example, it can be used to understand human expressions in online video meetings. In online video meetings, there are missing micro-interaction aspects when compared to direct social interactions. Facial expression recognition in online video meetings is expected to increase understanding of users' interactions. The use of online video meetings currently reaches 300

million meetings per day. This indicates that the video meeting has become commonplace in today's digital world, especially during the COVID-19 pandemic [5].

Video meetings are generally preferred to audio-only because it has several benefits that cannot be obtained through audio-only meetings. Users can better understand by seeing the speaker's lips movements, tongues, jaw, and facial expressions and can help understand the speaker's intention [6] [7]. The problem faced in video meetings is the limitation of humans who cannot focus on many things. For example, when a teacher is delivering learning material in a video meeting, at the same time, the teacher cannot observe the reactions given by all of his students. Even though the students' reactions themselves need to be understood by the teacher to get insight from the use of the learning methods. Facial expressions are one of the most important non-verbal means for humans to understand human emotions and can be analyzed into meaningful insight [8]. From the example above, by gathering insights from student reactions, teachers can immediately look for the best learning method according to their students in conducting online learning to be done more effectively.

A facial recognition competition ever held using the FER2013 dataset, which was released into scientific papers. The best result of that competition is constructed by combining three methods using sparse filtering for feature learning, random forests for feature selection, and support vector machine for classification with 70.22% accuracy [9]. Other research that has been conducted using the Convolutional Neural Network architecture for facial expression recognition objects using the same dataset has the best accuracy of 72.7% with the VGG architecture [10]. Meanwhile, research using other deep neural networks has also been held using VGG and ResNet with various training methods resulting in the best accuracy of 84.986% [11]. Some recent research also shows that image augmentation has become a training method to improve model accuracy. Recent research on image augmentation using a novel method called FMix can increase model accuracy by up to 2% on the ImageNet dataset [12].

An accurate model built with deep learning or deep artificial neural network can be applied to solve the existing problems stated before by recognizing and classify human facial expressions. As an example, it can be used to detect student facial expressions in online learning via video meetings. Using a good classification model can help teachers observe their students and get feedback or insights from the learning methods. However, it is possible to implement a facial expression classification model in other fields and cases like online learning and the online hiring process.

The model must be light enough but at the same time must have good accuracy because many facial expressions need to be classified simultaneously. Furthermore, the model must be applied easily in various environmental conditions. In this study, we used a deep learning model, a machine learning field inspired by the human neural network that is arranged in a chain and performs a specific function [13]. In addition, image augmentation can also be used in the training phase to improve model accuracy and train models to adapt better and generalize new data. Moreover, open facial datasets on the internet are generally imbalanced in each class, so class weighting is required on the loss function or sampling process during training. Thus, in this research, our contributions are three folds:

1) We introduce large scale training model for recognizing facial expression using FMix as image augmentation to reduce overfitting.
2) We introduce new model architecture extended from the residual network by adding Accuracy Booster Plus block and changing ReLU activation function to Mish.
3) We evaluate the proposed model on three datasets of AffectNet, FERPlus, and RAF-DB and achieve the higher accuracy than previous studies that using same residual network backbone.

2. Material and Methods

2.1. Datasets

The datasets used in this paper are collected from many popular facial expression datasets such as AffectNet [8], FERPlus [11], facial_expressions [14], and RAF-DB [15] [16]. This merged dataset is divided into eight classes that are neutral, happy, surprise, sad, anger, disgust, fear, and contempt. Image samples from each class are shown in Figure 1 consecutively.

2.1.1. AffectNet. AffectNet is the largest dataset of facial expression image datasets to date. This dataset consists of 1 million face images comprised of approximately 420 thousand images manually labeled by humans and 580 thousand images labeled automatically using models trained using images labeled by humans. This dataset is divided into two types: class expressions and dimensional in numerics representing facial expressions' value [8].

Fig. 1. Image samples from collected datasets

2.1.2. FERPlus. The FERPlus dataset is an improved dataset from FER2013, where the data is re-labeled with ten annotators, thus achieving a higher agreement percentage of up to 90% [11]. This dataset is available in the form of the number of votes by each annotator. In this study, a majority voting scheme is used to decide the final label.

2.1.3. facial_expressions. The facial_expressions dataset is an open dataset in a public GitHub repository. This dataset is not explicitly divided. In this study, all data will be used as training data [14].

2.1.4. RAF-DB. The Real-world Affective Faces Database (RAF-DB) is a dataset of facial expressions with around 30 thousand data retrieved from the internet. The data collected were independently labeled by 40 annotators [16].

2.2. Proposed Method

In this study, we used the residual network as it has shortcut connection that is intended to solve the vanishing gradient problem [17]. By utilizing the residual network as the base network, we also extend it with Accuracy Booster Plus Block, and change the original activation with the Mish function.

2.2.1. Accuracy Booster. Accuracy Booster is an additional block to be appended to the residual block in the ResNet architecture. This block is a development from SENet, where the fully connected layer is replaced with the CNN and batch normalization layer. We used this block to recalibrate the features extracted from each residual block as mentioned by the original author. Based on experiments on ImageNet dataset, Accuracy Booster has shown an increase in performance compared to SENet while keeping computation costs almost the same. The performance of Accuracy Booster has outperformed SENet by about 0.3% in ImageNet classification with a class number of 1000 [18]. There are two variants of this block: Accuracy Booster (using depth-wise CNN, in Figure 2a) and Accuracy Booster Plus (using CNN, in Figure 2b).

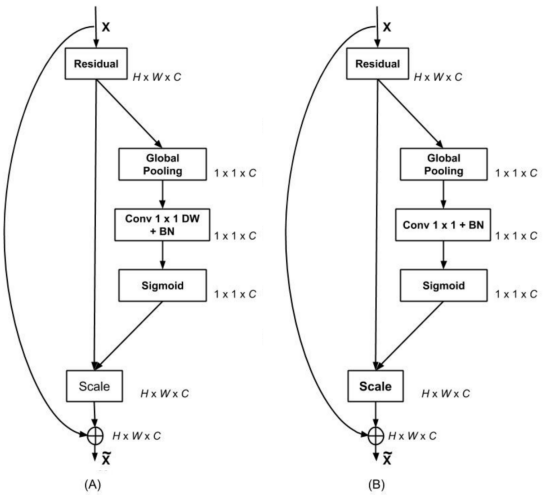

Fig. 2. Illustration of Accuracy Booster block [18]

2.2.2. Mish. Mish is a novel self-regularized non-monotonic activation function that can replace the ReLU activation function commonly found in many neural network architectures. Mish is related to the Swish function, where both have similar formula. We choose this activation as in some experiments especially on ImageNet dataset, Mish outperformed Swish and can generalize better [19]. The equation of Mish function is written in Equation 1.

$$F(x) = x * tanh(ln(1 + e^x)) \tag{1}$$

Where x is an input value.

2.2.3. Image Augmentation. In addition to the proposed method, we also used image augmentation to prevent neural network learning too quickly and to generate more varied training data. Image augmentation can be done by directly manipulating the pixel elements in the image such as flipping, rotating, cropping, color manipulation, and image perspective modification. We also used an advanced image augmentation method called FMix.

2.2.4. FMix. FMix is a form of image augmentation by combining two images into one image, also known as Mixed Sample Data Augmentation (MSDA). Merging is based on masks generated from binary random numbers forming a single form or continuous region. Next, the number 0 will be filled with the value from image 1, and the number 1 will be filled with the value from image 2 or

vice versa [12]. We choosed this method because it produces asymmetric merging patterns so it can help the artificial neural network to learn important features or details better. The illustrations of FMix augmentation on the face image dataset are shown in Figure 3.

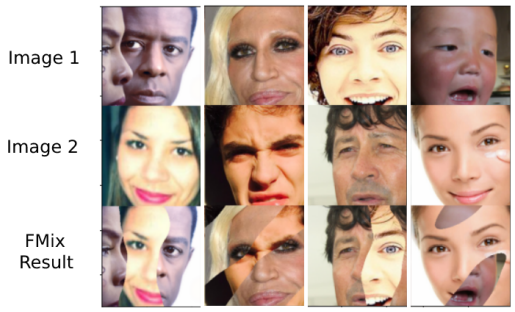

Fig. 3. Example mask and mixed images from AffectNet facial expression dataset for FMix

2.2.5. Loss Function. To evaluate how the proposed model perform during training process in this study, we used the most commonly used loss function in classification problem, namely log (cross-entropy) loss [20]. The equation of cross-entropy loss is written in Equation 2.

$$L = \sum_{j} y^{(j)} log(o^{(j)}) \tag{2}$$

where y is the classification target and o is the model output.

2.2.6. Class Weighting. We analyzed that the datasets we have used have imbalanced data in each class. To overcome this problem, we apply class weighting to the loss function. The weighting formula we have used is defined in two forms: class weighting with normalization is written in Equation 3 and class weighting without normalization is written in Equation 4.

$$W_i = 1 - (\frac{count(data_i)}{\sum_j count(data_j)}) \tag{3}$$

$$W_i = \frac{1}{count(data_i)} \tag{4}$$

Where W_{i} is weight calculation of class i.

2.3. Validation and Evaluation

Validation and evaluation are the final stage of this study. This stage will determine whether the proposed method has better performance or not. We will evaluate our proposed method result on validation and test set using accuracy metrics and confusion matrix.

During experiments, we calculated accuracy in two ways. First, in the mixed section, we mixed all validation and test sets from each dataset into one and then passed it through the network as a small batch. Second, in the AffectNet, FERPlus, and RAF-DB datasets, we calculated accuracy by separating them into their original partition of each dataset. For example, we used AffectNet and FERPlus validation set separately and then calculated accuracy for each set when we still develop the network. And then, in the final test, we used the AffectNet validation set, FERPlus test set, and RAF-DB test set separately and also calculate accuracy for each set.

2.4. Research Flow

This research will conduct experiments on the implementation of Residual ConvNet and image augmentation on facial expression classification. The research flow is shown in Figure 5. The hyperparameters that we tune in this research are learning rate (step size or how fast network weight updated), beta1 (exponential decay rates for the first-moment estimates), and beta2 (exponential decay rates for the second-moment estimates) [21]. Various hyperparameters and optimization methods were also used during training and testing to find the best result of that combination. The optimization methods used in this study are Stochastic Gradient Descent, Lookahead [22] + Rectified Adam [23] (this combination is also known as Ranger), and Lookahead [22] + Rectified Adam [23] + Gradient Centralization [24]. After training using certain combinations, the model is evaluated using accuracy metrics on validation data and test data.

As we have stated before, the proposed architecture block based on the residual network. And then, we have changed the default activation of the residual network from ReLU to Mish, and Accuracy Booster Plus block is also appended after the residual branch. We preserved the original form of the residual network that has two forms: basic block and bottleneck that showed in Figures 4a and 4b. In this research, we used only two types of the residual networks: ResNet-18 and ResNet-50. We also preserved the original architecture of the residual network, as shown in Table 1.

3. Result and Discussion

During training and testing hyperparameters and architecture, weighted loss with normalization is

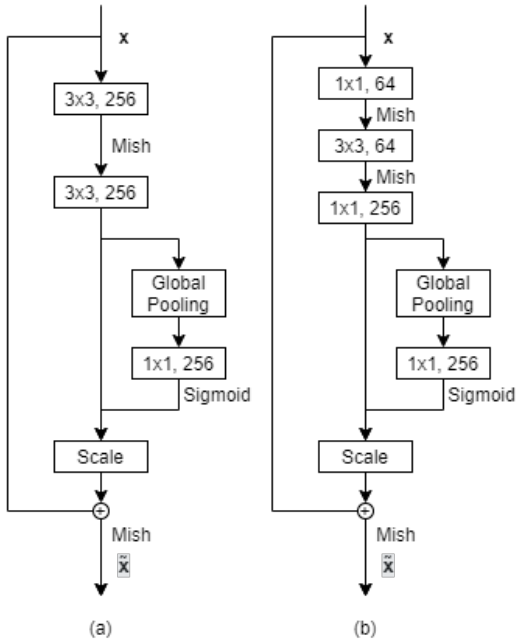

Fig. 4. Proposed residual network block architecture: (a) basic block and (b) bottleneck

TABLE 1
FULL ARCHITECTURE OF THE PROPOSED METHOD

layer name	output size	18-layer	50-layer
conv1	112x112	7x7, 64, stride 2	
conv2x	56x56	3x3 max pool, stride 2	
		[Basic Block 64d] x2	[Bottleneck 64d] x3
conv3x	28x28	[Basic Block 128d] x2	[Bottleneck 128d] x4
conv4x	14x14	[Basic Block 256d] x2	[Bottleneck 256d] x6
conv5x	7x7	[Basic Block 512d] x2	[Bottleneck 512d] x3
	1x1	average pool, 8-d fc, softmax	

used because the validation data and test data from each secondary data have different data characteristics. AffectNet dataset has balanced data for each class in the validation data. In contrast, FERPlus and RAF-DB have imbalanced data distribution for each class in the validation data and test data. The normalized weighted loss is chosen so that the model can produce good predictions on an imbalanced dataset and slightly better on a balanced dataset compared with no weighted loss.

During experiments, all training and testing use a step number of 60000, so all training using various models and configurations has the same treatment. Some image augmentations also applied in this study, such as brightness, contrast, hue, saturation, rotation, shear, and FMix. FMix is applied to all training processes as we found that this method can help to reduce overfitting in loss value and escalate accuracy in mixed validation set when the training step is quite large as shown in Figure 6. The FMix parameters that we used in all experiments are default parameters of official implementation, decay power (decay power for frequency decay prop $1/f^d$) = 3 and alpha (alpha value for beta distribution from which to sample mean of mask) = 1 [12]. Simultaneously, the augmentation methods used during training processes and the value of each image augmentation is shown in Table 2.

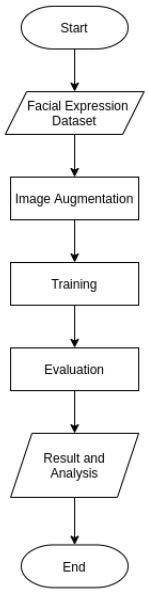

Fig. 5. Research flow

TABLE 2
AUGMENTATION METHODS

Augmentation Method	Value
Brightness	0.25
Contrast	0.25
Hue	0.05
Saturation	0.05
Rotation	15°
Shear	15°

3.1. Hyperparameter Testing and Evaluation

3.1.1. Learning Rate Testing and Evaluation. Learning rate testing is carried out on some optimization algorithms explained in Section 2.4. The

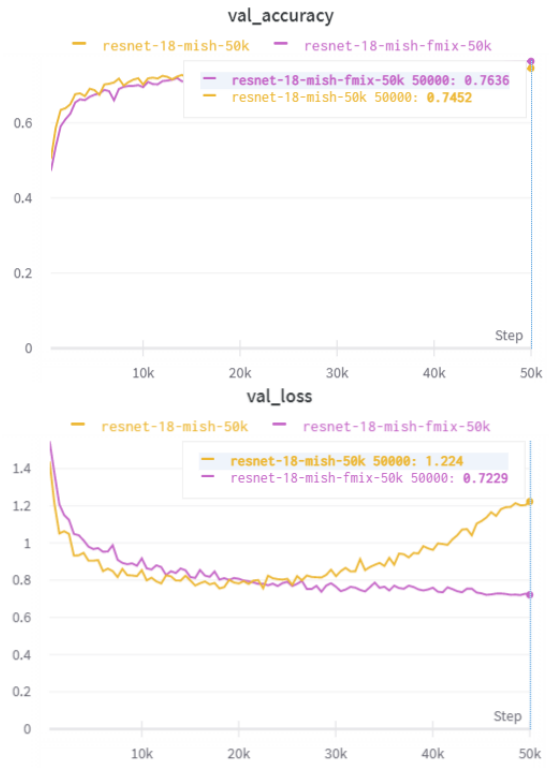

Fig. 6. FMix image augmentation usage comparison

beta1 and beta2 values used in this stage are 0.9 and 0.999. The evaluation results of various learning rates and optimizer methods are showed in Table 3.

TABLE 3
CLASSIFICATION RESULT WITH DIFFERENT LEARNING RATE AND OPTIMIZER METHOD

Learning Rate	Accuracy		
	AffectNet	FERPlus	Mixed
0.01 (SGD)	0.495	0.8377	0.7388
0.001 (SGD)	0.4708	0.8025	0.7009
0.001 (Lookahead + RAdam)	0.5135	**0.8577**	0.7606
0.001 (Lookahead + RAdam + GC)	**0.5195**	0.8534	**0.7641**

3.1.2. Beta1 Testing and Evaluation. Beta1 testing is carried out on the best optimization algorithm from the previous learning rate test results, namely Lookahead + Rectified Adam + Gradient Centralization with a learning rate of 0.001 and its accuracy evaluated on validation data or test data. The beta2 value used in this stage of testing is 0.999. The evaluation results of various beta1 values are shown in Table 4.

TABLE 4
CLASSIFICATION RESULT WITH DIFFERENT BETA1 VALUE

Beta1	Accuracy		
	AffectNet	FERPlus	Mixed
0.9	0.5195	0.8534	**0.7641**
0.95	**0.521**	**0.8553**	0.7589

3.1.3. Beta2 Testing and Evaluation. Beta2 testing is carried out on the best optimization algorithm from the previous learning rate test results, namely Lookahead + Rectified Adam + Gradient Centralization with a learning rate of 0.001 and its accuracy evaluated on validation data or test data. The beta1 value used in this stage of testing is 0.9. The evaluation results of various beta2 values are shown in Table 5.

TABLE 5
CLASSIFICATION RESULT WITH DIFFERENT BETA2 VALUE

Beta2	Accuracy		
	AffectNet	FERPlus	Mixed
0.98	0.5188	0.8504	0.7617
0.999	**0.5195**	**0.8534**	**0.7641**

3.2. Model Architecture Testing and Evaluation

The model architecture determines the capacity of the neural network for learning. Better architecture can learn patterns better from the data. Moreover, a model that has been previously trained using more extensive data (pre-trained) can help the model get better accuracy. The evaluation results of changing model architecture are shown in Table 6.

From the evaluation results in Table 6, pre-trained ResNet-18 produce a better result on two AffectNet and FERPlus validation sets than the model initialized from random value. Moreover, on the mixed dataset, this model is still has lower accuracy than the other models. Note that, pre-trained ResNet-18 has been trained using the ImageNet dataset with 1000 classes and a total of about 14 million images. This result proves that fine-tuning from a larger dataset to a smaller dataset helps the model to produce a better performance model since the feature extraction layer in the pre-trained model has better capabilities than the model initialized with random weights. However, although the pre-trained model has shown better results, this model still seems to

TABLE 6
CLASSIFICATION RESULT WITH DIFFERENT ARCHITECTURE OF THE PROPOSED METHOD

Architecture	Accuracy		
	AffectNet	FERPlus	Mixed
ResNet18	0.5195	0.8534	**0.7641**
CNN18	0.5270	0.8549	0.7609
ImageNet Pre-trained ResNet18	**0.5298**	**0.8630**	0.7623
ResNet18 + Mish (proposed method)	0.5280	0.8540	0.7640
ResNet18 + Accuracy Booster (proposed method)	0.5208	0.8539	0.7595
ResNet18 + Accuracy Booster + Mish (proposed method)	0.5210	0.8582	**0.7641**

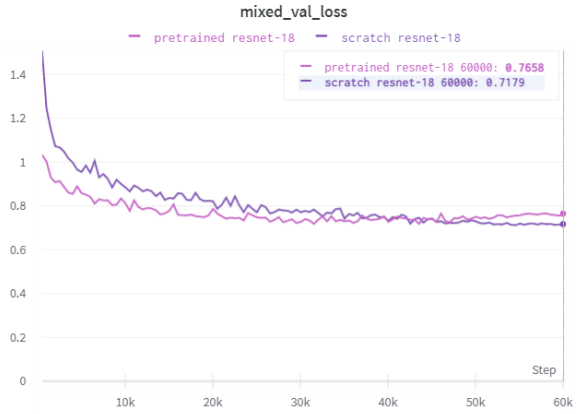

Fig. 7. Comparison of scratch ResNet-18 with pre-trained ResNet-18 validation loss graph

start to be overfitting, as shown by the increasing loss value in Figure 7 (magenta line). This overfitting phenomenon also indicates that the model can produce better results if regularizers are applied for training.

On the other hand, the proposed model showed positive results. Compared to the standard ResNet-18, we found that the proposed model outperforms the performance in terms of testing against training data and test data by a small margin. For example, in AffectNet and FERPlus datasets. However, we also found that sometimes the methods we proposed show no improvement in accuracy metrics. Figure 8 shows that the overfitting phenomenon also appears when compared to standard ResNet-18. Based on this result, we conclude that this model shows the potential to be developed in further studies for more significant improvements. For example, some regularizer methods can be added to the proposed model

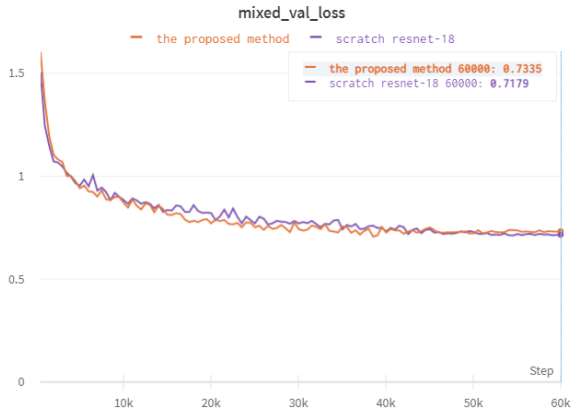

Fig. 8. The proposed method (ResNet-18 + Mish + Accuracy Booster Plus) validation loss graph

to reduce overfitting. Another suggestion is doing transfer learning from larger datasets, as the previous result has been mentioned. Thus, the models can reduce overfitting, find more optimum weights, and improve testing accuracy.

3.3. Specific Dataset Testing and Evaluation

Referring to the evaluation results in Table 6, the best model of training is obtained using ResNet-18 + Mish + Accuracy Booster Plus. Furthermore, to increase the model capacity, ResNet-50 + Mish + Accuracy Booster Plus was chosen to test and evaluate with a specific dataset. When training the AffectNet evaluation model, weighted loss without normalization is used. Besides, when training the FERPlus and RAF-DB evaluation model, the model is trained without weighted loss. The comparisons of our network to the previous study can be seen in Table 7. The evaluation results using confusion matrix from each dataset, namely AffectNet validation data, FERPlus test data, and RAF-DB test data, can be seen respectively in Table 8, Table 9, and Table 10. From the tables below, all testing scenarios still show the worst results in class with fewer data.

3.4. Discussion

Based on experiments on this study, many aspects can be improved. First, adding the pre-training method before the modified model is trained to classify the original dataset to make the model have better weights at the beginning of the training. Second, adding the augmentation method exploration scheme so the best combination of augmentation

TABLE 7
COMPARISONS OF THE PROPOSED METHOD RESULTS TO THE PREVIOUS STUDIES

Architecture	Accuracy		
	AffectNet	**FERPlus**	**RAF-DB**
AlexNet + Weighted Loss [8]	0.58	-	-
VGG-13 + Majority Voting [11]	-	0.8385	-
RAN (ResNet-18+) [25]	0.595	-	-
RAN (VGG-16) [25]	-	**0.8916**	-
RAN (ResNet-18) [25]	-	-	0.869
ResNet-18 (ours)	-	0.8372	0.883
ResNet-50 + Mish + Accuracy Booster Plus (proposed method)	-	0.8488	**0.8879**
ResNet-18 + Weighted Loss (ours)	0.596	-	-
ResNet-50 + Mish + Accuracy Booster Plus + Weighted Loss (proposed method)	**0.5972**	-	-

TABLE 8
CONFUSION MATRIX OF THE PROPOSED METHOD RESULT ON AFFECTNET VALIDATION SET EVALUATION

	NE	**HA**	**SU**	**SA**	**AN**	**DI**	**FE**	**CO**
NE	57.8	4.4	7.4	8.2	7.0	3.2	2.0	10.0
HA	2.8	81.2	2.6	1.2	0.6	1.4	0.2	10.0
SU	11.0	7.4	56.8	3.6	4.0	3.0	11.6	2.6
SA	13.2	1.8	2.8	62.6	9.4	3.8	3.6	2.8
AN	12.4	1.2	4.4	5.8	61.8	7.4	3.8	3.2
DI	6.4	5.4	3.8	7.8	21.6	47.8	4.6	2.6
FE	3.2	2.8	15.2	8.2	4.8	3.4	62.2	0.2
CO	16.4	19.6	2.2	3.4	6.0	4.4	0.6	47.4

methods can be obtained for facial image classification. Then, adding the regularization method in the training method or model architecture to overcome the overfitting problem encountered in almost all tests. Furthermore, adding tests using other weighted loss calculation functions or testing the use of sampler weighting to address the data imbalance problem.

4. Conclusion

During experiments in this study, the best results obtained using normalized weighted-loss with an accuracy of 0.7641 are obtained using Lookahead + RAdam + Gradient Centralization, the learning rate of 0.001, beta1 of 0.9, and beta2 of 0.999. We also observe that transfer learning using the ImageNet dataset brings accuracy improvement over the model generated from random values. The evaluation results show that the model produces fairly good accuracy when the data is imbalanced, where some facial expressions that rarely appear in the dataset also rarely appear in the real world. Meanwhile, the addition of the Mish activation function and the Accuracy Booster Plus block shows an improvement from the original model on the ResNet-18 architecture in all validation data and test data used in the study. The best evaluation results of the ResNet-50 model with the Mish activation function and the Accuracy Booster Plus block to the AffectNet validation data of 0.5972, the FERPlus validation data of 0.8636, the FERPlus test data of 0.8488, and the RAF-DB of 0.8879.

TABLE 9
CONFUSION MATRIX OF THE PROPOSED METHOD RESULT ON FERPLUS TEST SET EVALUATION

	NE	**HA**	**SU**	**SA**	**AN**	**DI**	**FE**	**CO**
NE	89.19	2.07	0.79	6.20	1.19	0.07	0.24	0.24
HA	2.59	93.32	1.62	1.19	1.08	0.11	0.11	0.00
SU	6.26	2.91	84.56	0.45	2.68	0.00	3.13	0.00
SA	21.52	1.79	0.22	70.40	3.81	0.90	1.34	0.00
AN	9.66	3.43	0.93	4.05	79.75	1.56	0.62	0.00
DI	10.00	10.00	5.00	5.00	20.00	50.00	0.00	0.00
FE	7.22	0.00	27.84	8.25	5.15	0.00	51.55	0.00
CO	25.00	0.00	0.00	7.14	10.71	7.14	3.57	46.42

TABLE 10
CONFUSION MATRIX OF THE PROPOSED METHOD RESULT ON RAF-DB TEST SET EVALUATION

	NE	**HA**	**SU**	**SA**	**AN**	**DI**	**FE**
NE	86.32	2.06	1.03	8.24	0.29	1.91	0.15
HA	2.70	95.19	0.25	0.59	0.25	0.84	0.17
SU	3.95	1.52	88.75	1.22	1.22	0.91	2.43
SA	5.23	1.26	0.21	91.42	0.42	0.84	0.63
AN	4.32	4.32	0.62	3.09	77.16	4.94	5.56
DI	10.00	6.87	1.87	10.62	5.00	65.00	0.62
FE	4.05	4.05	6.76	10.81	1.35	4.05	68.92

References

[1] N. Yudistira and T. Kurita, "Gated spatio and temporal convolutional neural network for activity recognition: towards gated multimodal deep learning," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, Dec. 2017. [Online]. Available: https://doi.org/10.1186/s13640-017-0235-9

[2] C. B. Thacker and R. M. Makwana, "Human behavior analysis through facial expression recognition in images using deep learning," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 2, pp. 391–397, Dec. 2019. [Online]. Available: https://doi.org/10.35940/ijitee.b6379.129219

[3] M. Sajjad, S. Zahir, A. Ullah, Z. Akhtar, and K. Muhammad, "Human behavior understanding in big multimedia data using CNN based facial expression

recognition," *Mobile Networks and Applications*, vol. 25, no. 4, pp. 1611–1621, Sep. 2019. [Online]. Available: https://doi.org/10.1007/s11036-019-01366-9

[4] S. Shakya, S. Sharma, and A. Basnet, "Human behavior prediction using facial expression analysis," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, Apr. 2016. [Online]. Available: https://doi.org/10.1109/ccaa.2016.7813754

[5] A. H. Mustaqim, "Persaingan ketat layanan vicon," May 2020. [Online]. Available: https://tekno.sindonews.com/read/29710/207/persaingan-ketat-layanan-vicon-1589497511?showpage=all

[6] V. Bruce, "The role of the face in communication: Implications for videophone design," *Interacting with Computers*, vol. 8, no. 2, pp. 166–176, 1996.

[7] C. O'Malley, S. Langton, A. Anderson, G. Doherty-Sneddon, and V. Bruce, "Comparison of face-to-face and video-mediated interaction," *Interacting with Computers*, vol. 8, no. 2, pp. 177–192, 1996.

[8] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, p. 18–31, Jan 2019. [Online]. Available: http://dx.doi.org/10.1109/TAFFC.2017.2740923

[9] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," 2013.

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

[11] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," 2016.

[12] E. Harris, A. Marcu, M. Painter, M. Niranjan, A. Prügel-Bennett, and J. Hare, "Fmix: Enhancing mixed sample data augmentation," 2020.

[13] J. Laserson, "From neural networks to deep learning," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 18, no. 1, pp. 29–34, Sep. 2011. [Online]. Available: https://doi.org/10.1145/2000775.2000787

[14] B. L. Y. Rowe, "facial_expressions," 2016. [Online]. Available: https://github.com/muxspace/facial_expressions

[15] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2584–2593.

[16] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[18] P. Singh, P. Mazumder, and V. P. Namboodiri, "Accuracy booster: Performance boosting using feature map recalibration," 2020.

[19] D. Misra, "Mish: A self regularized non-monotonic activation function," 2020.

[20] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," 2017.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[22] M. R. Zhang, J. Lucas, G. Hinton, and J. Ba, "Lookahead optimizer: k steps forward, 1 step back," 2019.

[23] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," 2020.

[24] H. Yong, J. Huang, X. Hua, and L. Zhang, "Gradient centralization: A new optimization technique for deep neural networks," 2020.

[25] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," 2019.