Lightweight and Accurate FER for Driver Emotion Analysis: Optimizing ResNet with Patch Extraction and Self-attention Techniques

Ibrokhim Muminov[1], Kamronbek Yusupov[2], Md Rezanur Islam[2], Mahdi Sahlabadi[3], and Kangbin Yim[3](✉)

[1] Department of Computer Software Engineering, Soonchunhyang University, Asan, Korea
theibrokhim@sch.ac.kr
[2] Department of Software Convergence, Soonchunhyang University, Asan, Korea
{arupreza,yuskamron}@sch.ac.kr
[3] Department of Information Security Engineering, Soonchunhyang University, Asan, Korea
sahlabadi@ieee.org, yim@sch.ac.kr

Abstract. Facial Expression Recognition (FER) is becoming one of the most widely utilized techniques for defining the emotional state of a vehicle operator to prevent traffic accidents. Deep CNN networks are heavily utilized for FER tasks, as they have achieved significant advancements and proved their efficiency during the last decade. However, deep CNN networks are computationally expensive due to a significant number of parameters and do not achieve high accuracy in facial expression classification. To address these issues, we optimized the widely-known ResNet-50 and ResNet-101 models by implementing a patch extraction block and a self-attention network. As a result, the optimized models achieved the test accuracies of 95% and 94%, respectively. Additionally, the number of model parameters reduced by 6.17 and 5.87 times, respectively, without any impact on accuracy.

1 Introduction

Emotions, whether constructive or destructive, can significantly interfere with driving performance [1], contributing to severe traffic accidents and endangering the lives of vehicle operators [2]. Consequently, progressive research in the field of emotion recognition is being conducted to prevent further possible road accidents caused by the emotional state of drivers. Voice recognition, electroencephalogram (EEG), and electrocardiogram (ECG) techniques have been broadly implemented for defining the emotional state of car operators in the last decade [3]. However, nowadays, facial expression recognition (FER) method is becoming the ultimate choice when it comes to the driver's emotion detection task on account of the latest advancements in the field of Artificial Intelligence (AI).

Pre-trained deep CNNs (Convolutional Neural Networks) are predominantly implemented in FER tasks due to their outstanding performance in ImageNet competitions

L. Barolli (Ed.): BWCCA 2024, LNDECT 231, pp. 171–180, 2025.
https://doi.org/10.1007/978-3-031-76452-3_17

[4]. In this research, we highlighted ResNet-family models [5], including ResNet-50 and ResNet-101. ResNet is highly effective in facial expression recognition due to its extremely deep neural network and superior emotion classification accuracy compared to other deep CNN models. Additionally, ResNet uses residual blocks to address the vanishing gradient problem [6], enabling it to extract the finest features from a facial image. Nevertheless, ResNet models are high in computational power as a result of tens of millions of parameters and low in accuracy during the FER tasks. Ultimately, these challenges make the implementation of the models unsuitable in resource-constrained devices like vehicles.

There are several techniques for optimizing deep CNN models, such as transfer learning [7], pruning [8], quantization [9], and parameter sharing [10]. However, while these optimization methods make the model lightweight, they can negatively impact its overall accuracy. To address this, we implemented the patch extraction block and self-attention network proposed by Ngwe et al. in [11] into the ResNet-50/101 models, effectively reducing the number of parameters while increasing accuracy, thereby proving the effectiveness of the proposed method.

For testing the efficiency of the proposed lightweight ResNet-50/101 models, FER+ dataset [12] was chosen. FER+ is based on the FER2013 dataset [13], using the same set of images but with improved labels and has better distribution of sentimental categories, which helps in building robust facial expression recognition systems. FER+ originally consists of 35,889 48x48 pixel images partitioned across 8 different emotion labels: anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise. However, in our research focused on a vehicle operator safety, we removed the disgust and contempt classes, as they are not as commonly studied or as clearly linked to immediate driving behaviors as other emotions, resulting in a dataset deduction of 877 images. To ensure the model receives new variations of the images at each training epoch, the dataset was augmented using the ImageDataGenerator method from the TensorFlow framework.

In the next stage of our research, the vanilla ResNet-50/101 models were tested on the augmented FERPlus dataset, where an accuracy of 85% and 83% was achieved respectively. Then, we applied newly defined patch extraction block and self-attention network to the models, which will be thoroughly discussed in the next sections. As a result, the number of parameters in ResNet-50 decreased by 6.17 times and in ResNet-101 by 5.87 times. Additionally, the accuracy of the two models increased by 10% and 11%, respectively. To the best of our knowledge, this research is the first one that implemented the patch extraction block and self-attention layers in pre-trained ResNet-50/101 models for FER tasks, drastically reducing the number of parameters while increasing the model accuracy significantly.

Remaining of this paper is organized as follows. Section 2 reviews related works with a focus on the optimization of ResNet-family models as well as the implementation of the patch extraction block and self-attention network in deep CNNs on the FER tasks. Section 3 provides a short, comprehensible overview of the steps taken during the research. Section 4 introduces the research results, followed by a thorough discussion in Sect. 5. Lastly, the conclusion and future work for this paper is included in Sect. 6.

2 Related Works

ResNet (Residual Network) was developed by He et al. [4] and achieved state-of-the-art results in ILSVRC and COCO 2015 competitions that assess algorithms for object detection and image classification at large scale. The main goal of the research was to design an extremely deep network that does not have a vanishing gradient issue compared to other deep CNN models. ResNet has various model variations with different numbers of layers, ranging from 20 to 1202. Among them, ResNet-50 is always a favorite choice when it comes to the FER tasks, considering the number of research studies conducted with ResNets in this field. [14].

The main idea of ResNet is the implementation of bypass pathway concept, which was utilized in Highway Nets [15] to address the issue of training a deeper network. ResNet proposed residual connections within layers to enable cross-layer connectivity and provided observational evidence showing that these residual networks are easier to optimize, minimalize the chance of overfitting, and can gain accuracy from substantially increased depth. However, the profound depth of the network results in tens of millions of parameters, and the accuracy in FER tasks often remains below 90%.

Research addressing the drawbacks of ResNet-50/101—namely, their high computational demands and relatively low accuracy—remains an active area of study. For instance, [16] incorporated an attention mechanism into the ResNet-50 architecture to boost its ability to recognize essential features correlated with particular emotional states. This mechanism helps the model focus on the most relevant parts of the input data, improving the accuracy of sentiment recognition. Moreover, the study utilized a customized Sigmoid cross-entropy loss function to better deal with emotion classification tasks. This function is designed to weigh different features according to their importance, ensuring the model learns to prioritize the most significant emotional indicators. The research suggests that the proposed model indicated substantial improvement in performance but still struggles with identifying uncommon emotional categories.

Jin et al [17] proposed several optimization methods to enhance the ResNet-34 model for emotion recognition based on EEG signals. First, the model developed in this research integrates a feature fusion mechanism that combines information from different feature sources. It is achieved through multiple parallel branches with varying convolution kernel sizes, allowing the network to capture and fuse diverse features from EEG data effectively. This fusion improves the network's ability to interpret complex data relationships, enhancing the accuracy of emotion recognition. Second, the model incorporates an attention mechanism after the residual network module that helps the model focus on the most relevant parts of the input data, thereby improving the effectiveness of feature extraction. Finally, the network parameters are optimized using both softmax loss and center loss functions, which enhances the model's classification performance and stability. Although the feature fusion technique improves model performance, it also increases the model's complexity. This added complexity could lead to higher computational costs, making the model less suitable for vehicles.

Ngo et al. [18] describes the optimization of the ResNet-50 model for FER on static images using transfer learning and fine-tuning techniques. The following approach leverages the knowledge the model has gained from a large-scale, diverse dataset and adapts it to the FER task with a smaller dataset to avoid overfitting issues. The model

proposed in this research is initially pre-trained on the ImageNet dataset for object detection tasks. Then, it is fine-tuned on the AffectNet dataset to recognize eight common facial expressions by specifically adjusting the model's parameters to the FER task, improving the model's performance in recognizing facial expressions from static images. While transfer learning and fine-tuning techniques improve performance metrics, the overall improvement is not as substantial as expected.

The most recent research which succeeded in enhancing performance metrics while reducing the number of parameters in deep CNN models is PAtt-lite [11]. Patt-lite proposed a combination of lightweight patch and self-attention network based on a MobileNetV1 structure to improve FER performance. In PAtt-lite, a truncated ImageNet-pre-trained MobileNetV1 is used as a baseline feature extractor in FER. In place of the truncated original layers, the model features a patch extraction block composed of three different convolutional layers: two depthwise separable convolutional layers and one pointwise convolutional layer. These convolutional layers are responsible for splitting the feature maps into four patches, learning higher-level features from its input. The patch extraction block is specifically designed and added to better adapt to the FER datasets than simply fine-tuning the top layers of the model. An attention classifier is also added to the network to enhance the learning of these feature maps from substantially lightweight feature extractor. As a result of these modifications, MobileNetV1 model achieved state-of-the-art results in CK +, RAF-DB, FER2013, and FERPlus datasets, with an average accuracy of 95.77% and with only 1.10M model parameters.

ResNet-50/101 models' performance could be enhanced through the implementation of the patch extraction and self-attention network too. As a reason, deeper layers of the model can be skipped and be replaced by a newly defined patch extraction block to capture more abstract and diverse features from the data. Depthwise separable convolutional layers, the main component of the patch extraction block, have a substantially smaller number of parameters while still preserving the same spatial dimensions of the input volume as a traditional convolutional layer does. Self-attention network could further enhance the model's ability to extract the finest features from the facial expression data.

3 Methodology

Most research on optimizing deep CNN models from the ResNet family has faced a trade-off: reducing computational power often comes at the cost of decreased accuracy. This research aims to achieve the opposite effect by optimizing popular ResNet models (ResNet-50/101) to reduce the number of parameters while simultaneously enhancing accuracy. To attain this goal, a patch extraction block and self-attention network were implemented. The efficiency of the method was tested based on the following steps. First, FERPlus was chosen as the optimal dataset for this research. Subsequently, vanilla ResNet-50/101 models were tested on the FERPlus dataset to establish an initial benchmark, which is to compare to the final benchmark of their lightweight versions. Having established the initial benchmark, we walked through all individual layers of the ResNet-50 model and 4th convolutional block layers of ResNet-101 model using grid search to find the most optimal location for attaching the patch extraction block. As a result, we discovered the ideal layer after which the patch extraction block was attached, enhancing the ability of ResNet-50/101 models to reveal the finest features from the FER data.

4 Experimental Evaluation

The patch adaptation process via grid search proved its effectiveness in defining the best variation of ResNet-50/101 models, which displayed the highest accuracy and low computational power, for FER tasks. In terms of ResNet-50, the patch extraction block was attached after each layer of the model. All layers consequent to the patch extraction block were truncated, and the top layers were replaced by a Global Average Pooling (GAP) layer, followed by a dot product self-attention layer situated.

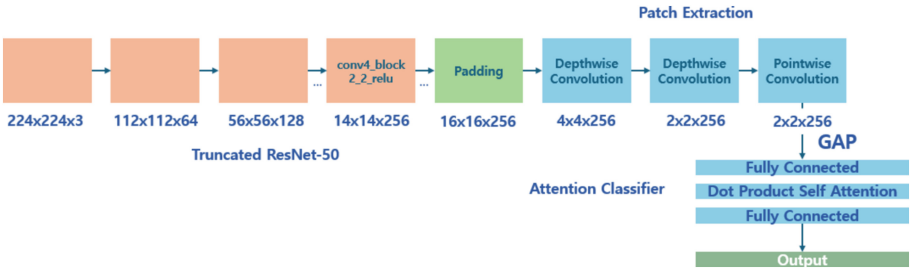


Fig. 1. Proposed ResNet-50 Architecture

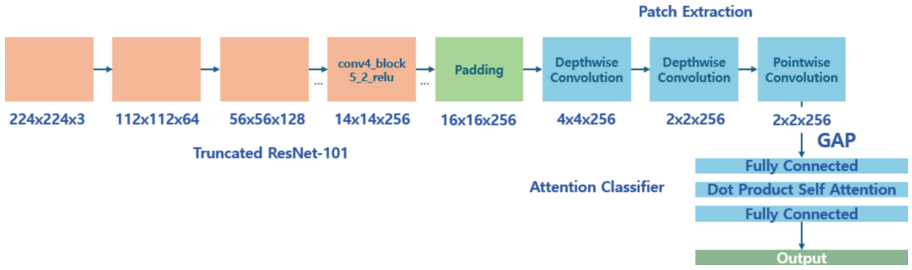
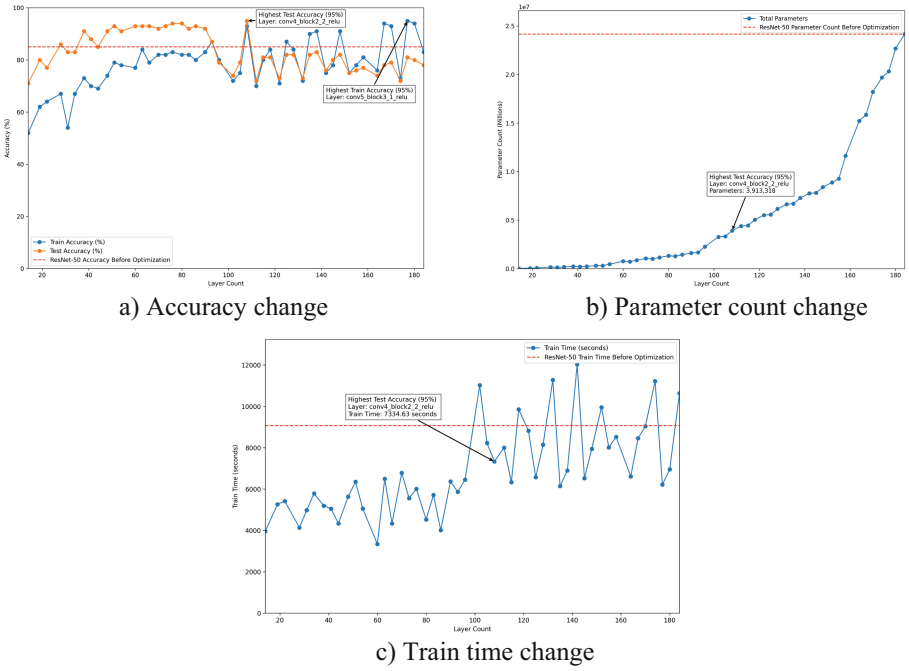


Fig. 2. Proposed ResNet-101 Architecture

between two fully connected layers. A total of 50 model variations were tested based on the same configuration and hardware mentioned in Table 1. During the patch adaptation process, the training accuracy of the test models fluctuated between 52% and 95%, while the test accuracy ranged from 71% to 95%, as shown in Fig. 3(a) . The parameter number of the models displayed a constant upward trend as the number of layers increased over time, as shown in Fig. 3(b). The lowest train time on the GPU specified in Table 1 was 3963.81 s, whereas the highest train time achieved 11279.43 s, as displayed in Fig. 3(c). Attaching the patch extraction layers after the block 'conv4_block2_2_relu' of the ResNet-50 model displayed the best performance in emotion classification. As a result of the optimization, the accuracy of the ResNet-50 model increased by a whopping 10%, followed by a drastic 6.17-fold decrease in parameter number. The optimized ResNet-50 model structure can be seen in Fig. 1.



a) Accuracy change

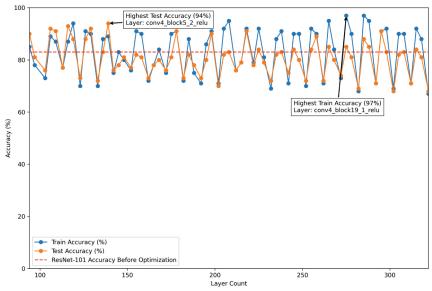
b) Parameter count change

c) Train time change

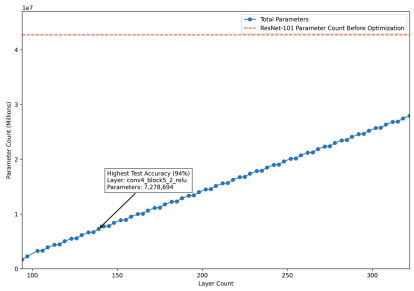
Fig. 3. ResNet-50 Patch Adaptation Result

In the case of ResNet-101, the patch extraction block was attached to each layer between the 'conv4_block1_1_relu' and 'conv4_block23_out' blocks. A total of 69 model variations were tested using the same configuration as in ResNet-50. The train and test accuracies, as well as the training time on the GPU, fluctuated significantly, as shown in Fig. 4(a, c). The patch extraction block after 'conv4_block5_2_relu' achieved the highest test accuracy of 94% compared to other model variations during the research. As a result of the optimization, the ResNet-101 model experienced an 11% increase in the test accuracy, along with an approximate 5.87-fold decrease in parameter count. The optimized ResNet-101 model structure can be seen in Fig. 2.

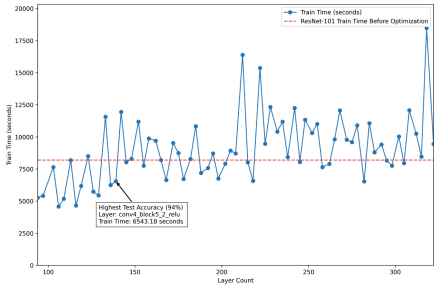
Training deep CNN models for ordinary FER tasks and especially optimizing them by running hundreds of model variations to find the one with the highest accuracy and low computational requirements presents significant challenges in terms of hardware and time. In the starting point of our research, CPU was utilized for the model training process, as the previously owned GPU was not sufficient to handle the FERPlus dataset. For running one epoch on FERPlus dataset on CPU, approximately 3.5 h were required. Consequently, achieving the results from the first model took roughly 10 days. As a result, the latest GPU, mentioned in Table 1, was purchased to speed up the research process. Thanks to the latest technology specifically designed for handling complex AI tasks, it took approximately 11 days instead of months to run 119 ResNet-50/101 models for the patch adaptation procedure and achieve state-of-the-art results.



a) Accuracy change



b) Parameter count change



c) Train time change during

Fig. 4. ResNet-101 Patch Adaptation Result

Table 1. Model Hyperparameters and Hardware Specifications

Hyperparameter	Value / Details
Epochs	200
Learning_rate	0.001
Batch_size	8
Steps_per_epoch	len(train_generator)
Validation_steps	len(val_generator)
Callbacks	[reducelr, earlystop, lambdacb, tensorboard, checkpoint]
Activation	ReLu
Optimizer	Adam
Loss	categorical_crossentropy
Hardware Specifications	Details
CPU	Intel® Core™ i9-10900K CPU @ 3.70GH.z
GPU	NVIDIA RTX A6000
GPU Memory	128 GB
Dedicated GPU Memory	47.5 GB
Shared GPU Memory	64.0 GB

5 Discussion

The patch adaptation process mentioned in the previous sections discovered the optimal place for attaching the patch extraction block in both ResNet-50/101, increasing the model accuracy and decreasing the number of parameters. The patch extraction block was attached in each layer of the ResNet-50 model, as the aim of this research was to define the global maximum of test accuracy, keeping the number of parameters low.

As shown in Fig. 3(a), the test accuracy of the models displayed fluctuation over time, achieving its apex at 95% during the 27th test. Although the graph may not reveal any patterns at first glance, a thorough investigation shows that the model variations in the first half of the research achieved significant improvements in test accuracy compared to ResNet-50's pre-optimization test accuracy. On the other hand, all model variations in the second half of the research performed worse than the vanilla ResNet-50 model. It means that the patch extraction block and the self-attention network boost the model's performance in FER tasks when it is attached to the middle layers. It can be explained by the fact that the middle layers of the pre-trained deep CNN model learns specific features of training samples, providing detailed information to the next layers [20]. In terms of the train time, 91.04% of the models were significantly faster compared to pre-optimized ResNet-50 as a result of the patch extraction block.

In ResNet-101, the patch extraction block was attached to specific middle layers based on the promising results observed in the post-optimized ResNet-50. With 69 tests conducted overall, the model achieved its global maximum accuracy of 94% after the patch extraction block was applied to the 'conv4_block5_2_relu' layer. 60.86% of the optimized models did not perform well compared to vanilla ResNet-101 model. Also, 53.62% of the models were significantly slower compared to the pre-optimized ResNet-101 model as a result of the patch extraction block. Based on these results, it can be argued that adapting the patch extraction block in ResNet-101 is not as effective as in ResNet-50.

A potential drawback of the patch extraction block used in this research is that it is dataset-specific. For example, in the FERPlus dataset, the ResNet-50 model achieved the highest accuracy when the patch extraction block was attached after the 'convolution4_block2_2_relu' layer. To achieve optimal results in other FER datasets, the patch extraction block might need to be attached to a different layer to effectively extract the most relevant features from that specific dataset. As a result, the patch adaptation process would need to be repeated to identify the best layer for attaching the patch extraction block, which is time-consuming and requires robust hardware.

6 Conclusion and Future Work

A total of 119 models were tested to provide solid evidence on the effectiveness of applying the patch extraction block and self-attention layer in ResNet-50/101 models, making them suitable for implementation in vehicles with limited computational power for pursuing FER tasks. As a result of the research, both models experienced a substantial boost in test accuracy and significant decrease in parameter count. ResNet-50's test accuracy increased from 85% to 95%, meanwhile the parameter number decreased from

24,813,574 to 3,913,318, by 84.22%. Similarly, ResNet-101's test accuracy increased from 83% to 94%, showing an 82.96% plummet in parameter count, from 42,724,070 to 7,278,694. In future work, we aim to optimize other deep CNN models, such as VGG16, DenseNet-121, and Xception. We also aim to enhance the patch extraction block so that it consistently delivers state-of-the-art results when attached to the same layer across all publicly available FER datasets.

Acknowledgments. This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Convergence security core talent training business support program(IITP-2024–2710008611) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and This work was supported by the Korea Institute for Advancement of Technology(KIAT) grant funded by the Ministry of Trade, Industry and Energy(MOTIE)(No.P0022671, Basis for Implementing Industrial Digital Transformation Conformity Assessment and Testbed Platform).

References

1. Mesken, J: The role of emotions and moods in traffic. SWOV Institute for Road Safety Research. Netherlands: Leidschendam (2003)
2. Zhang, Q., et al.: The effect of the emotional state on driving performance in a simulated car-following task. Transport. Res. Part F: Traffic Psychol. Behav **69**, 349–361 (2020)
3. Khare, S.K., et al.: Emotion recognition and artificial intelligence: a systematic review (2014–2023) and research recommendations. Information Fusion **102**, 102019 (2024). https://doi.org/10.1016/j.inffus.2023.102019
4. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
5. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
6. Hochreiter, S.: The vanishing gradient problem during learning recurrent neural nets and problem solutions. Internat. J. Uncertain. Fuzziness Knowl.-Based Syst. **6**(02), 107–116 (1998)
7. Zhuang, F., et al.: A comprehensive survey on transfer learning. Proc. IEEE **109**(1), 43–76 (2021). https://doi.org/10.1109/JPROC.2020.3004555
8. Cheng, H., Zhang, M., Shi, J.Q.: A survey on deep neural network pruning: taxonomy, comparison, analysis, and recommendations. IEEE Trans. Pattern Anal. Machine Intell. (2024). https://doi.org/10.1109/TPAMI.2024.3447085
9. Wu, J., et al.: Quantized convolutional neural networks for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and pattern Recognition (2016)
10. Ott, J., et al.: Learning in the machine: to share or not to share? Neural Netw. **126**, 235–249 (2020)
11. Ngwe, J.L., et al.: PAtt-Lite: lightweight patch and attention MobileNet for challenging facial expression recognition. IEEE Access **12**, 79327–79341 (2024)
12. Barsoum, E., et al.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction (2016)
13. Goodfellow, I.J., et al.: Challenges in representation learning: a report on three machine learning contests. In: Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, korea, november 3–7, 2013. Proceedings, Part III 20. Springer berlin Heidelberg (2013)

14. Iman, M., Arabnia, H.R., Rasheed, K.: A review of deep transfer learning and recent advancements. Technologies **11**(2), 40 (2023)
15. Srivastava, R.K., Klaus, G., Jürgen, S.: Highway networks. arXiv preprint arXiv:1505.00387 (2015)
16. Lin, J.: Application of resnet-50 based user sentiment analysis in digital media advertising design. IOS Press, International Symposium on World Ecological Design (2024)
17. Jin, E., et al.: An optimized residual network for emotion recognition based on a multi-features fusion technology and electroencephalography signals. J. Mech. Med. Biol. **24**(02), 24400086 (2024)
18. Ngo, T.Q., Yoon, S.: Facial expression recognition on static images. In: Future Data and Security Engineering: 6th International Conference, FDSE 2019, Nha Trang City, Vietnam, 27–29 Nov 2019, Proceedings 6. Springer International Publishing (2019)