Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 4 (2025) 714 - 720 e-ISSN: 2580-0760

Real-time Emotion Recognition Using the MobileNetV2 Architecture

Triyani Hendrawati[1*], Anindya Apriliyanti Pravitasari[2], Nazamuddin[3], Riza Fazhriansyah Hermawan[4], Satrio Adilia Subekti[5], Muhammad Yasyfi[6]
[1,2,3,4,5,6] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Bandung, Indonesia

[1]triyani.hendrawati@unpad.ac.id, [2]anindya.apriliyanti@unpad.ac.id, [4]riza2002@mail.unpad.ac.id, [6]muhammad20102@mail.unpad.ac.id

Abstract

Facial recognition technology is now advancing quickly and is being used extensively in a number of industries, including banking, business, security systems, and human-computer interface. However, existing facial recognition models face significant challenges in real-time emotion classification, particularly in terms of computational efficiency and adaptability to varying environmental conditions such as lighting and occlusion. Addressing these challenges, this research proposes a lightweight, yet effective deep learning model based on MobileNetV2 to predict human facial emotions using a camera in real time. The model is trained on the FER-2013 dataset, which consists of seven emotion classes: anger, disgust, fear, joy, sadness, surprise, and neutral. The methodology includes deep learning-based feature extraction, convolutional neural networks (CNN), and optimization techniques to enhance real-time performance on resource-constrained devices. Experimental results demonstrate that the proposed model achieves a high accuracy of 94.23%, ensuring robust real-time emotion classification with a significantly reduced computational cost. Additionally, the model is validated using real-world camera data, confirming its effectiveness beyond static datasets and its applicability in practical real-time scenarios. The findings of this study contribute to advancing efficient emotion recognition systems, enabling their deployment in interactive AI applications, mental health monitoring, and smart environments. Real-world camera data is also used to evaluate the model, demonstrating its usefulness in real-time applications and its efficacy beyond static datasets. The results of this work advance effective emotion identification systems, making it possible to use them in smart settings, interactive AI applications, and mental health monitoring.

Keywords: facial recognition; deep learning, MobileNetV2, CNN, tensorflow

How to Cite: T. Hendrawati and A. Apriliyanti Pravitasari, "Real-time Emotion Recognition Using the MobileNetV2 Architecture", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 4, pp. 714 - 720, Jul. 2025.
*Permalink/*DOI: https://doi.org/10.29207/resti.v9i4.6158

Received: October 31, 2024
Accepted: July 2, 2025
Available Online: July 17, 2025

1. Introduction

In computer vision, face recognition has emerged as one of the most crucial technologies. It has found widespread applications in contactless payments, security, access control, healthcare, smart surveillance, and personalized services, enabling automated person identification or authentication based on facial images [1]. Over the past few decades, the rapid advancement of deep learning-based methods and the availability of large-scale datasets have significantly improved the accuracy and robustness of facial recognition systems [2].

A face recognition system generally consists of three main steps: face detection, feature extraction, and face matching. Over time, these steps have evolved from traditional handcrafted feature-based techniques to modern convolutional neural network (CNN)-based methods, which have revolutionized the field [1]. CNNs are particularly effective in capturing intricate facial features, allowing recognition models to perform reliably in challenging conditions, such as poor lighting, pose variations, occlusions, and facial expressions [3].

Despite these advancements, several challenges persist. Issues such as pose variations, lighting conditions, occlusions (partial face coverage), and privacy concerns

continue to pose significant hurdles for researchers. Additionally, ethical concerns and bias in facial recognition technology have drawn increasing attention, necessitating improvements in fairness and transparency [4]. To address these issues, researchers continue to refine facial recognition models by improving accuracy, computational efficiency, and real-world adaptability. Among various CNN architectures, MobileNet has been specifically designed to overcome the high computational costs associated with deep learning models. MobileNetV2, created by Google researchers, offers a compromise between performance and efficiency and is tailored for embedded and mobile applications [5].

Real-time facial expression categorization is a significant face recognition task that is essential to intelligent tutoring systems, affective computing, mental health monitoring, and human-computer interface (HCI). Real-time facial expression recognition (FER) has potential uses in a variety of industries, such as security, entertainment, customer service, and education. However, it is still very difficult to achieve great accuracy while preserving computing efficiency, especially in situations with limited resources like embedded systems and mobile devices [6]. Traditional CNN architectures, such as VGG16 and ResNet, provide excellent accuracy but are computationally intensive, making them unsuitable for real-time scenarios. To address these limitations, MobileNetV2 has been proposed as a highly efficient deep learning model that balances computational cost and classification performance [7]. MobileNetV2 is perfect for edge computing and real-time AI applications because it uses depthwise separable convolutions, which drastically reduce the number of parameters while preserving excellent accuracy. Several studies have validated that MobileNetV2 achieves performance comparable to larger architectures while consuming fewer computational resources, making it a promising candidate for real-time facial emotion recognition [8],[9],[10]. According to earlier research, MobileNetV2 offers a good balance between computational efficiency and accuracy [5-8], which makes it perfect for real-time deployment on low-resource devices. This empirical evidence supports its implementation in our work by illustrating the efficacy of MobileNetV2 for real-time face emotion identification, as evidenced by comprehensive assessments on the FER-2013 dataset.

The novelty of this work lies in integrating MobileNetV2 for real-time face emotion identification with immediate validation utilizing actual camera input. In contrast to prior research that concentrated exclusively on offline FER-2013 accuracy, our implementation evaluates the model's resilience in practical, real-time scenarios. This study validates the robustness of MobileNetV2 in recognizing human emotions accurately and efficiently, making it a feasible solution for affective computing, human-computer interaction, and real-time AI applications [11], [12].

2. Methods

2.1 Deep learning

One class of machine learning methods for extracting features from data is called deep learning. Artificial neural networks are used in deep learning to process big datasets with multiple layers of information processing stages in a hierarchical architecture that is used for feature learning without supervision. Higher-level features or components are dictated by lower-level ones in deep learning, which computes features or hierarchical representations from observational data. In deep learning, computer systems are typically trained to convert picture pixel values into internal representations using sizable datasets. This produces internal representations that allow classifiers to identify input patterns [12].

Multi-layered structures in artificial neural networks are represented by the learning technique known as deep learning. One machine learning technique called representation learning automatically finds and learns features from the unprocessed input data. The unprocessed data can be used for classification, or recognition as it is commonly known. The core architecture types in deep learning are recurrent neural networks, autoencoders, convolutional neural networks, and deep belief networks. Deep learning is still being researched since it can manage very big data sources, perform novel activation functions, find new methodologies, and significantly increase chip processing capabilities [11].

2.2 Convolutional Neural Network

A kind of neural network focused on deep learning, the convolutional neural network has undergone a great deal of testing and produced excellent outcomes in real-world applications. CNN is very useful in the field of large-scale video and image recognition. CNN can be said to be the most dominant approach for all recognition problems. There have been many applications of CNN in the industrial field, such as Amazon, Facebook, and Google, where CNN is typically used to extract house numbers from street view images [13], [14]. The CNN workflow is depicted in Figure 1.

From Figure 1, the input image is fed into the network, where its features are extracted through convolutional layers with ReLU activation functions, detecting patterns such as edges and textures. Next, the pooling layer reduces feature dimensions while preserving important information, improving computational efficiency. To extract more complicated features, this convolution and pooling process is done several times. After the output is flattened into a one-dimensional vector, fully connected layers process it, assigning the learnt features to categories like neutral, anger, disgust, fear, joy, sadness, and surprise. Finally, Softmax layer generates probabilities for each class, allowing the model to determine the detected object type.

Due to CNN's ability to represent each neuron in two dimensions, this approach is typically appropriate for processing input data in the form of images. The feature learning stage and the classification stage are CNN's two primary phases. Convolution layers, pooling layers, and ReLU (activation function) are all part of the feature learning step. The classification stage includes prediction, flattening, and fully connected layers. The feed-forward and backpropagation processes are the two primary functions of CNN [14], [15].

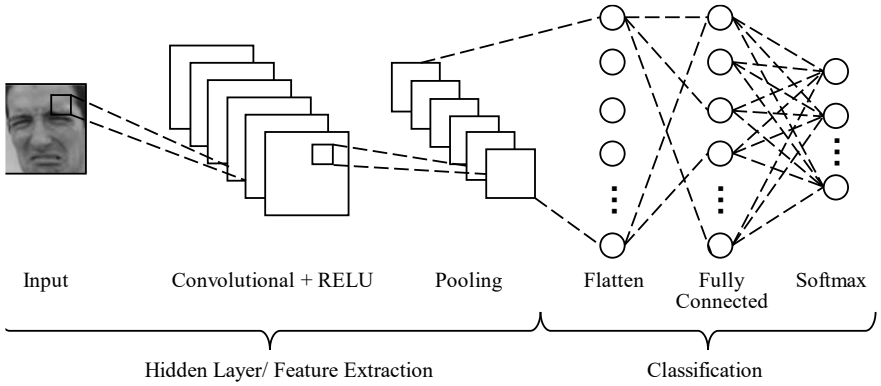

Figure 1. Workflow of CNN

2.3. MobileNetV2

The application of Deep Learning methods has spread across various computer vision tasks [16] and one general overview of its application is in FER (The Face Expression Recognition). With only a minor drop in performance, the modifications to the MobileNet design can drastically lower computational costs when compared to competing Deep Learning techniques for recognizing objects.

MobileNetV2 is an architecture of CNN that is usually used to address excessive computing resource needs. MobileNet differs from CNN architecture in that it uses convolutional layers with filter thicknesses that can match the thickness of the input image [17]. MobileNetV2 improves model performance on some tasks and benchmarks across the spectrum at different model sizes [4].

The linear bottleneck and the shortcut links between bottlenecks are the two new elements of MobileNetV2. The inner layers may contain the model's capacity to convert inputs from lower-level concepts to higher-level ones, whilst the bottleneck segment comprises inputs and outputs between the models. For instance, turning pixels into a picture. Faster data training and increased accuracy are made possible via shortcuts between bottlenecks [4]. In several previous studies, MobiExpressNet was able to achieve an accuracy of 67.96% on the Face Emotions Recognition (FER)-2013 dataset with a model size of approximately 75,000 parameters and a computation of 1x10^6 FLOP. Figure 2 shows the MobileNetV2 workflow.

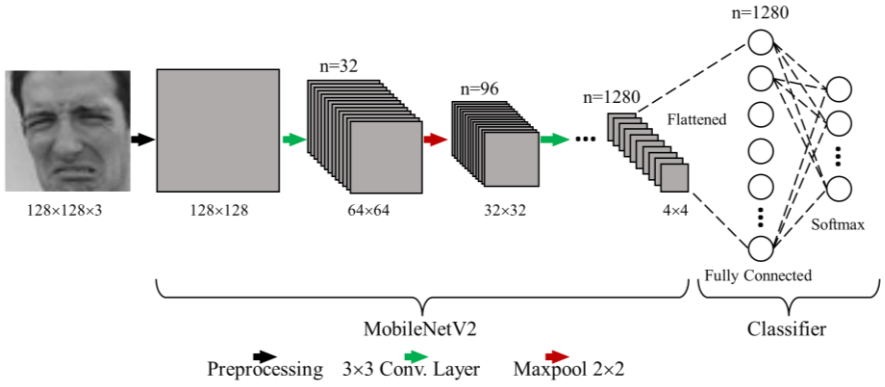

Figure 2. Workflow of MobileNetV2

Preprocessing is the first step in Figure 2, where a 128 × 128 × 3 input image is ready for feature extraction. The first stage involves 3×3 convolutional layers with ReLU activation, generating 32 feature maps of size 64×64. These are then processed through additional convolutional layers, increasing the depth to 96 feature maps of size 32×32, followed by max pooling (2×2) to reduce spatial dimensions. The next stage applies further convolutions, expanding the depth to 1280 feature maps with a resolution of 4×4, effectively extracting high-level features while maintaining computational efficiency. After being flattened, the collected characteristics are sent to a fully connected layer, which uses a Softmax classifier to map them to discrete categories (0 for anger, 1 for disgust, 2 for fear, 3 for joy, 4 for sadness, 5 for surprise, and 6 for neutral). MobileNetV2's lightweight architecture, using depthwise separable convolutions, ensures that the

model achieves high accuracy with minimal computational cost, making it suitable for mobile and embedded applications.

The activation function used in this investigation is the Rectified Linear Unit, or ReLU. ReLU in the MobileNetV2 model is one of the important components that support the performance of neural networks in recognising patterns from data, such as images. Equation 1 illustrates the straightforward formula that defines ReLU.

$$\text{ReLU}(x) = \max(0, x) \quad (1)$$

This means that if the input is positive, it will be passed as is; if it is negative, then the value will be changed to zero. All things considered, MobileNetV2's usage of ReLU boosts effectiveness and performance while maintaining the model's portability and speed for mobile applications without compromising accuracy or the model's capacity to execute tasks like object identification and picture categorization [18] .

The optimiser used in this research is the Adaptive Moment Estimation optimiser or Adam. Adam is an optimiser that maintains the learning rate of each weight and the average gradient exponentially. Adam is an efficient stochastic method that only requires first-order gradients with minimal memory requirements.

A loss function computes the difference between actual and expected outputs to determine the loss associated with every scenario that a model generates [19]. The model is less effective at capturing patterns the higher the loss function value. Using the loss function in this investigation A loss function used in multi-class classification models is called Sparse Categorical Cross Entropy [10], [20].

In this study, the model's performance is assessed using the confusion matrix (Table 1). The main components of the confusion matrix are False Positives (FP), False Negatives (FN), True Positives (TP), and True Negatives (TN). We calculate performance metrics using precision. How often a model correctly predicts the outcome given in Equation 2 is measured using a parameter known as accuracy [10], [20].

$$\text{Accuracy} = (TP+TN) / (TP+TN + FP+FN) \quad (2)$$

Table 1. The confusion matrix

Predicted	Actual Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

3. Results and Discussions

The data set used as the training set is the Face Emotions Recognition (FER)-2013 data available on the Kaggle website. This data consists of grayscale facial images sized 48x48 pixels. Each face in the data is divided into seven categories according to the emotions conveyed by the visible facial expressions. The training data used to build the model consists of 15,958 images. These 7 categories are represented by the following numbers: 0 denotes anger, 1 disgust, 2 fear, 3 joy, 4 sadness, 5 surprise, and 6 neutral.

3.1 Results

Before the model is built. Data preprocessing needs to be done first. The data preprocessing includes up sampling the dataset from 48x48 to 224x224 using bi-linear interpolation, followed by rescaling the image values by multiplying the image values by 1/255 so that the image values are in the range of 0 to 1.

Following the completion of data preparation, the model is constructed using MobileNetV2, which is a component of the Keras TensorFlow framework. Next, tuning is performed on the dense layer using a combination of 128-64-7 neurons. Thus, the model to be used contains a total of 2,430,663 parameters, with 2,396,551 parameters being trained and 34,112 parameters not being trained. Next, model training is conducted.

The model was developed using the Adam Optimizer, a total of 25 epochs, using Sparse Categorical Crossentropy as the loss function. The training model's evaluation results revealed an accuracy of 0.9423. It can be said that the model is quite good at facial emotion recognition with an accuracy of 0.9423.

The bounding box and face detection in real-time scenarios were implemented using OpenCV before passing the face region into the MobileNetV2 model for emotion classification, where predictions will be made in real-time with a camera, resulting in the following outcomes.

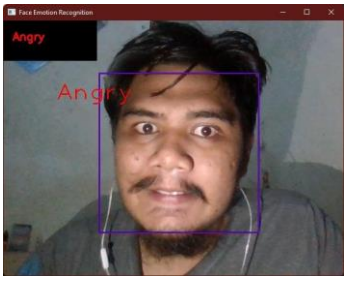

Figure 3. Angry facial expression

The results of real time face emotion recognition were tested with samples of people showing angry facial expressions. Figure 3's results demonstrate how well the model can identify the facial emotions of anger.

Figure 4. Happy facial expressions

Next, a real time face emotion recognition test was conducted on individuals with happy facial expressions.

Based on the results in Figure 4, it can be seen that the model can recognize happy facial emotions well.

Figure 5 shows the real-time face emotion recognition test on a person with a sad facial expression. The results indicate that the model can recognize the sad facial emotion well.

Figure 5. Sad facial expressions

Another trial provided by the facial expression of a surprised person. The results were obtained as shown in Figure 6. It is indicate that the model can recognize the emotion of surprised on faces well.

Figure 6. Expression of surprise

The following test was conducted with a sample of people with disgusted facial expressions. The results obtained (Figure 7), indicate that the model can recognize the emotion of disgust on faces well.

Figure 7. Disgusted facial expressions

In Figure 8, indicate that the model can recognize the emotion of fear on faces well.

Figure 8. Fear facial expressions

The last one was tested with a sample of people with normal facial expressions. The results obtained are as shown in Figure 9, indicating that the model can recognize normal facial emotions well.

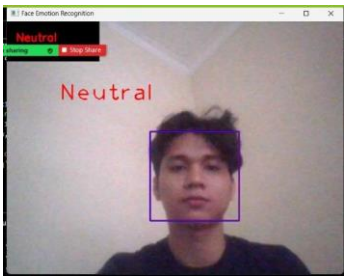

Figure 9. Fear facial expressions

3.2 Discussions

This study contributes by demonstrating that MobileNetV2 achieves high real-time emotion recognition accuracy (94.23%) on a resource-constrained environment, offering a practical solution for embedded AI applications in mental health, HCI, and smart surveillance systems. As it is derived from related performance, this could be more computation-efficient, important in real application scenarios for limited-resource devices, such as a smartphone or on-board system.

MobileNetV2 incorporates depth wise separable convolutions, which drastically cut down on the number of parameters while marginally lowering predictive accuracy when compared to conventional CNN models. This makes the model suitable for real-time emotion recognition without requiring extensive computational resources. When tested in realtime using a camera, this model can detect various facial expressions in realtime with high accuracy results as visualized in Figure 3-9. This is in line with the research conducted by Li & Deng [21] and Hou, et al [22].

However, a few limitations must be acknowledged. First is the FER-2013 dataset used in this work; grayscale images of small resolution (48x48 pixels) may not fully represent complex facial expressions from diverse environmental conditions. Second, variations in lighting and occlusion challenges are still there and would call for more enhancements in data augmentation or model architecture for further improvements.

To further evaluate the effectiveness and contribution of this study, we compare our results with several recent studies that employed deep learning methods for facial emotion recognition using the FER-2013 dataset. Table 2 summarizes the comparison of model architectures, datasets, accuracy, and whether real-time testing was conducted.

As illustrated in Table 2, the proposed MobileNetV2 model achieves higher accuracy than comparable methods, while also being validated in real-time through camera-based implementation. Unlike previous works that primarily focused on offline performance

using static datasets, our model's validation in live conditions highlights its practical relevance and adaptability in real-world environments. This constitutes a notable contribution to the domain of efficient and deployable facial emotion recognition systems.

Table 2 The summaries of comparison model architectures

Study	Model	Dataset	Accuracy	Real-time Imple-mentation
Li & Deng (2022) [21]	Deep CNN	FER-2013	91.4%	No
Huo et al. (2023) [22]	Improved DSCNN	FER-2013	92.1%	No
Santemiz et al. (2024) [23]	Side-view Hybrid CNN	FER-2013	89.5%	No
This study	MobileNetV2	FER-2013 + Real-time input	94.23%	Yes

4. Conclusions

This study shows how the MobileNetV2 architecture balances accuracy and computing efficiency, which makes it perfect for problems requiring real-time facial expression identification. With an accuracy of 94.23%, the suggested model proves to be resilient and flexible. Despite the strong results, challenges such as dataset limitations and real-world variations require further exploration in future studies. The results emphasize the significance of ethical considerations in facial recognition deployment while also promoting deep learning applications in emotional computing and human-computer interaction.

The future work will explore more extensive hyperparameter tuning (e.g., learning rate, dropout, and batch size) and testing under different lighting, occlusion, and expression intensities to improve robustness. Future research work can also investigate hybrid models that incorporate attention mechanisms or transformer-based architectures to improve robustness against these types of variations.

Acknowledgements

This work was funded under the Higher Education Research program [Grant No. 1493/UN6.3.1/PT.00/2024] by the Ministry of Research, Technology, and Higher Education Indonesia's Directorate for Research and Community Service (DRPM) in collaboration with Universitas Padjadjaran.

References

[1] B. Amirgaliyev, M. Mussabek, T. Rakhimzhanova, and A. Zhumadillayeva, "A Review of Machine Learning and Deep Learning Methods for Person Detection, Tracking and Identification, and Face Recognition with Applications," *Sensors*, vol. 25, no. 5, p. 1410, Feb. 2025, doi: 10.3390/s25051410.

[2] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou,"ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *IEEE Trans Pattern Anal Mach Intell*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022, doi: 10.1109/TPAMI.2021.3087709.

[3] C. Xie, C. Xia, M. Ma, Z. Zhao, X. Chen, and J. Li, "Pyramid Grafting Network for One-Stage High Resolution Saliency Detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2022, pp. 11707–11716. doi: 10.1109/CVPR52688.2022.01142.

[4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.

[6] I. D. Raji and J. Buolamwini, "Actionable Auditing," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA: ACM, Jan. 2019, pp. 429–435. doi: 10.1145/3306618.3314244.

[7] Y. Qu *et al.*, "Joint Hierarchical Category Structure Learning and Large-Scale Image Classification," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4331–4346, Sep. 2017, doi: 10.1109/TIP.2016.2615423.

[8] V. Maeda-Gutiérrez *et al.*, "Comparison of Convolutional Neural Network Architectures for Classification of Tomato Plant Diseases," *Applied Sciences*, vol. 10, no. 4, p. 1245, Feb. 2020, doi: 10.3390/app10041245.

[9] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, Mar. 2021, doi: 10.1016/j.neucom.2020.10.081.

[10] V. Maeda-Gutiérrez *et al.*, "Comparison of Convolutional Neural Network Architectures for Classification of Tomato Plant Diseases," *Applied Sciences*, vol. 10, no. 4, p. 1245, Feb. 2020, doi: 10.3390/app10041245.

[11] F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU," May 2023. Available: http://arxiv.org/abs/2305.17473

[12] M. M. Taye, "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions," *Computers*, vol. 12, no. 5, p. 91, Apr. 2023, doi: 10.3390/computers12050091.

[13] J. S. Chung, R. Arandjelović, G. Bergel, A. Franklin, and A. Zisserman, "Re-presentations of Art Collections," 2015, pp. 85–100. doi: 10.1007/978-3-319-16178-5_6.

[14] A. C. S, "Advancements in CNN Architectures for Computer Vision: A Comprehensive Review," in *2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS)*, IEEE, Nov. 2023, pp. 1–7. doi: 10.1109/AICERA/ICIS59538.2023.10420413.

[15] W. Wang, Y. Sun, W. Li, and Y. Yang, "TransHP: Image Classification with Hierarchical Prompting," Apr. 2023.

[16] Y. Liang and W. Liang, "ResWCAE: Biometric Pattern Image Denoising Using Residual Wavelet-Conditioned Autoencoder," Jul. 2023.

[17] R. Patel and A. Chaware, "MobileNet architecture and its application to computer vision," in *Computer Vision and Recognition Systems Using Machine and Deep Learning Approaches: Fundamentals, technologies and applications*, Institution of Engineering and Technology, 2021, pp. 253–275. doi: 10.1049/PBPC042E_ch11.

[18] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," Apr. 2021.

[19] Z. Leng *et al.*, "PolyLoss: A Polynomial Expansion Perspective of Classification Loss Functions," Apr. 2022.

[20] M. H. D. Farahani, M. K. Mohamadi, and M. Lotfizad, "Visual Object Tracking using Sparse Representation and

Interest Points in a Double Step Approach," *Int J Comput Appl*, vol. 175, no. 10, pp. 1–9, Aug. 2020, doi: 10.5120/ijca2020920563.

[21] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Trans Affect Comput*, vol. 13, no. 3, pp. 1195–1215, Jul. 2022, doi: 10.1109/TAFFC.2020.2981446.

[22] H. Huo, Y. Yu, and Z. Liu, "Facial expression recognition based on improved depthwise separable convolutional network," *Multimed Tools Appl*, vol. 82, no. 12, pp. 18635–18652, May 2023, doi: 10.1007/s11042-022-14066-6.

[23] P. Santemiz, L. J. Spreeuwers, and R. N. J. Veldhuis, "A Survey on Automatic Face Recognition Using Side-View Face Images," *IET Biom*, vol. 2024, no. 1, Jan. 2024, doi: 10.1049/2024/7886911.