International Journal of Electrical and Computer Engineering (IJECE)
Vol. 16, No. 1, February 2026, pp. 395~403
ISSN: 2088-8708, DOI: 10.11591/ijece.v16i1.pp395-403

Facial emotion recognition under face mask occlusion using vision transformers

Ashraf Yunis Maghari, Ameer M. Telbani
Faculty of Information Technology, Islamic University of Gaza, Gaza, Palestine

Article Info

Article history:

Received Jun 22, 2025
Revised Sep 24, 2025
Accepted Nov 23, 2025

Keywords:

Deep learning
Emotion recognition
Face mask
Occlusion handling
Vision transformers

ABSTRACT

Facial emotion recognition (FER) systems face significant challenges when individuals wear face masks, as critical facial regions are occluded. This paper addresses this limitation by employing vision transformers (ViT), which offer a promising alternative with reduced computational complexity compared to traditional deep learning methods. We propose a ViT-based FER framework that fine-tunes a pre-trained ViT architecture to enhance emotion recognition under mask-induced occlusion. The model is fine-tuned and evaluated on the AffectNet dataset, which originally represents eight emotion categories. These categories are restructured into five broader classes to mitigate the impact of occluded features. The model's performance is assessed using standard metrics, including accuracy, precision, recall, and F1 score. Experimental results demonstrate that the proposed framework achieves an accuracy of 81%, outperforming several state-of-the-art approaches. These findings highlight the potential of vision transformers in recognizing emotions under masked conditions and support the development of more robust FER systems for real-world applications in healthcare, surveillance, and human–computer interaction. This work introduces a scalable and effective approach that integrates self-attention, synthetic mask augmentation, and emotion class restructuring to improve emotion recognition under facial occlusion.

Corresponding Author:

Ashraf Yunis Maghari
Faculty of Information Technology, Islamic University of Gaza
P.O. Box 108, Gaza, Palestine
Email: amaghari@iugaza.edu.ps

1. INTRODUCTION

Facial expressions are fundamental to human communication and emotion understanding [1]. Recent advances in computer vision have integrated facial expressions into many applications like virtual reality (VR) and augmented reality (AR), security systems, and human-computer interaction (HCI) [2]. Facial expression recognition (FER) systems classify emotions such as happiness, sadness, anger, fear, surprise, and disgust from static images and video streams [3]. FER models are trained on numerous images conveying emotions like happiness, sadness, anger, fear, surprise, and disgust. However, emotion expression varies due to age, culture, and gender [4], raising accuracy challenges and ethical concerns.

This study aims to enhance FER under mask-induced occlusion, using a vision transformers (ViT) based model trained on synthetically masked data. Traditional FER relies on manual feature engineering with preprocessing, feature extraction, and classification [2], while deep learning approaches such as deep belief networks (DBNs), long short-term memory networks (LSTMs), generative adversarial networks (GANs), and convolutional neural networks (CNNs) have improved performance [5]. Transfer learning further increased accuracy using pretrained models such as ResNet50, MobileNet, and VGG19 [6], [7]. Face masks,

widely used during COVID-19, obscure key regions like the mouth, making emotion recognition harder [8]. Researchers have applied attention heatmaps [9], cropped eye regions [10], or retained full images [11], [12] to address occlusion. Since no emotion-labeled masked datasets exist, synthetic approaches are used. Through self-attention, ViT models capture global dependencies and context from image patches, which allows ViTs to extract contextual features from partially occluded faces. This work proposes a ViT-based FER framework using a synthetically masked AffectNet dataset and restructured emotion classes. Following Magherini *et al.* [11], our approach demonstrates improved recognition performance under occlusion.

The remainder of this paper is organized as follows: section 2 reviews related studies. Section 3 describes the ViT-based FER model. Section 4 presents results and discussion. Section 5 concludes the paper.

2. RELATED WORK

The vision transformer (ViT) [13] has shown promising results in various computer vision tasks, including FER [14]. It is built on transformer architecture which is initially designed for NLP. ViT employs multi-head self-attention and image patch processing. Huang *et al.* [15] utilized ViT with a StarGAN framework for data augmentation in FER. Squeeze ViT was proposed to combine global and local features with fewer dimensions [16]. Fatima *et al.* [17] demonstrated the value of self-attention in ViT for emotion recognition.

Studies have also addressed FER under partial occlusion. Techniques like Gabor wavelet texture analysis, DNMF decomposition, and landmark-based shape analysis have been applied to separate occluded areas and extract discriminant features [18]. Other studies considered clothing-based occlusion, such as hijab detection using transfer learning [19], which also demonstrates the impact of partial covering on recognition performance. In addition to occlusion handling, other works focused on improving data quality and model robustness. Feng and Shao [20] enhanced data quality using preprocessing (*e.g.*, histogram equalization, affine transforms) and used Inception-v3 with transfer learning to achieve high accuracy on CK+ and Jaffe datasets. Other methods expanded classic CNNs, such as LeNet-5 [21], by deepening convolution and pooling layers to improve performance under occlusion. Chen *et al.* [22] proposed efficient attention-based ERFNet enhancements using group convolutions and residual modules. Mask-aware FER approaches have emerged recently. One study used CNNs on synthetically masked AffectNet data, merging emotion classes to address occlusion and achieved 96% training accuracy and 70% validation accuracy [11]. ACNN [23] was introduced to assign adaptive weights to facial region of interests (ROIs), with variants like pACNN and gACNN integrating local and global features. Recent work combined face parsing with a ViT-based classifier using cross-attention to differentiate masked and visible regions, outperforming other methods on datasets like M-LFW-FER and M-FER-2013 [24]. Our paper utilizes the capabilities of ViT's self-attention to improve FER under mask occlusion, using synthetically masked AffectNet data and class recategorization following Magherini *et al.* [11].

3. ViT-BASED FER FRAMEWORK

Figure 1 illustrates the overall framework of the proposed ViT-based FER model for masked facial images. The framework includes data collection from the AffectNet dataset, preprocessing using the mask-the-face (MTF) tool, fine-tuning of the pre-trained ViT model, and final evaluation using standard metrics (precision, recall, accuracy, and F1 score).

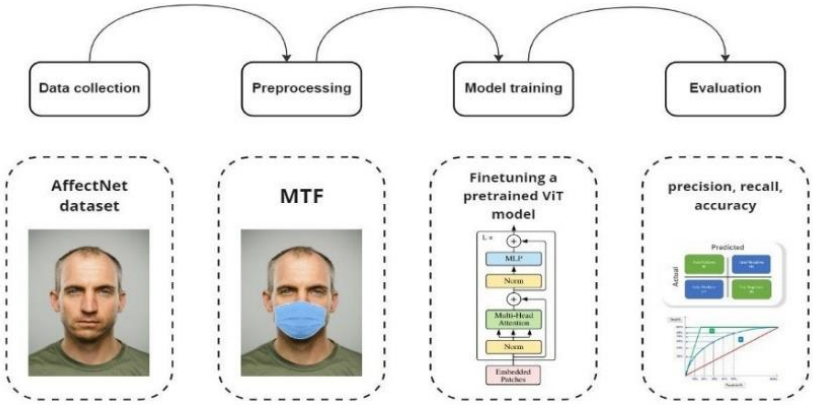

Figure 1. Workflow of the proposed FER framework

3.1. Data collection

AffectNet contains over one million facial images collected from the Internet by querying three search engines. About half of the retrieved images, around 450 thousands were manually annotated for eleven categories; neutral, happy, sad, surprise, fear, anger, disgust, contempt, none, uncertain, and non-face (he None (None of the eight emotions) [25]. These categories are shown in Figure 2.

AffectNet is widely used in facial expression recognition due to its scale and diversity, offering around 450K images [25] filtered from 120GB of data. With face masks occluding key features, emotion classification becomes difficult. Therefore, five classes (Anger-Disgust, Fear-Surprise, Happiness, Sadness, and Neutral) were created by merging similar expressions, as shown in Figure 3. This reclassification improves recognition under mask occlusion. The final distribution of these five classes is shown in Table 1, and stratified sampling ensured balance across training, validation, and test sets. It is worth noting that, in addition to AffectNet, other datasets like FER2013, JAFFE, and CK+ can also be considered for emotion recognition. These datasets may be used alone or combined with AffectNet to improve model performance. In this study, AffectNet was chosen due to its significantly larger number of facial images, more diverse and fine-grained emotion labels (seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral), and real-world image conditions. These advantages make it more suitable for building robust and scalable FER systems.

Figure 2. Sample images from the AffectNet dataset [25]

Figure 3. Reorganized emotion classes: Anger+Disgust and Fear+Surprise

Table 1. Samples per class after merging AffectNet categories

Expression Category	Number
Neutral	80276
Happy	146198
Sad	29487
Fear-Surprise	24479
Anger-Disgust	33394

3.2. Data preprocessing

The preprocessing step focuses on simulating real-world mask scenarios by masking the face images in the dataset. This step is crucial for adapting the model to emotion recognition under partial facial occlusion caused by mask usage. One commonly used tool is the "mask the face" (MTF) tool [26], depicted in Figure 4.

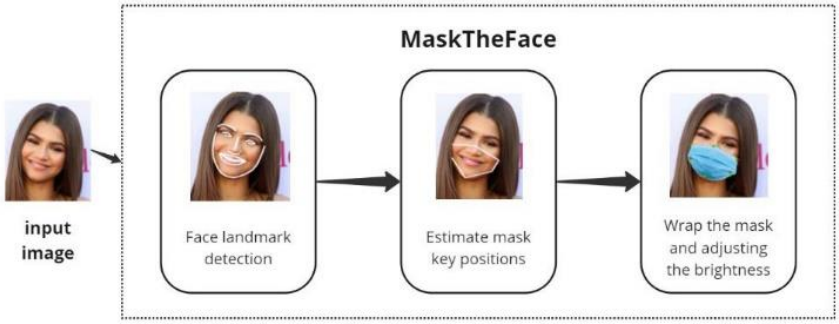

Figure 4. Workflow of the “Mask the FACE” (MTF) tool

3.3. ViT model finetuning

In this step, a pre-trained ViT model is fine-tuned using the augmented AffectNet dataset with synthetic masks. Fine-tuning involves further training the model on the masked images to refine its capability to the masked emotion recognition task.

3.3.1. Vision transformer architecture

The vision transformer (ViT) is a deep learning model that has shown state-of-the-art performance in classification tasks [13]. Originally developed for NLP, ViT later proved efficient in vision tasks [14], [27]. It consists of a patch embedding module that splits the image into patches and flattens them into tokens, followed by a transformer encoder composed of multi-head self-attention and feedforward layers. Each patch is linearly projected, and the attention mechanism computes a weighted sum across patches.

One key advantage of ViT is learning directly from data without manual feature engineering. It has demonstrated strong performance on datasets like ImageNet [28], making it a competitive alternative to traditional deep learning models. For FER, the self-attention mechanism is beneficial in capturing features from partially occluded facial images, such as those with face masks.

3.3.2. Fine-tuning steps

A pre-trained ViT model was fine-tuned using the masked AffectNet dataset to recognize emotions under occlusion. The steps include:

- Model selection: A pre-trained ViT model was chosen based on architecture and prior performance.
- Initialization: The model's parameters, learned from large datasets like ImageNet, were used as a starting point.
- Hyperparameter tuning: Learning rate, batch size, and regularization were adjusted experimentally.
- Fine-tuning: The model was trained on the masked dataset using backpropagation to optimize classification accuracy.

This process helps the ViT model learn the link between masked facial features and the recategorized emotions. An example of feature extraction by the fine-tuned ViT model is shown in Figure 5. This attention map was extracted from the final self-attention layer of the ViT model using visualization tools provided by the hugging face transformers library, and it highlights the facial regions (primarily the eye area) that most influence the model’s decision-making process.

3.4. Evaluation

The evaluation process typically comprises the following key component:

- Test dataset: Includes diverse facial expressions with masks, simulating real-world scenarios.
- Predictions: The model outputs probability distributions over emotion classes.
- Metrics: Precision, Recall, Accuracy, and F1 Score.
- Model comparison: The ViT-based FER model is compared with baseline and state-of-the-art methods to validate improvements.

Using these evaluation metrics can quantitatively measure the accuracy and effectiveness of the fine-tuned ViT model in recognizing emotions under face mask.

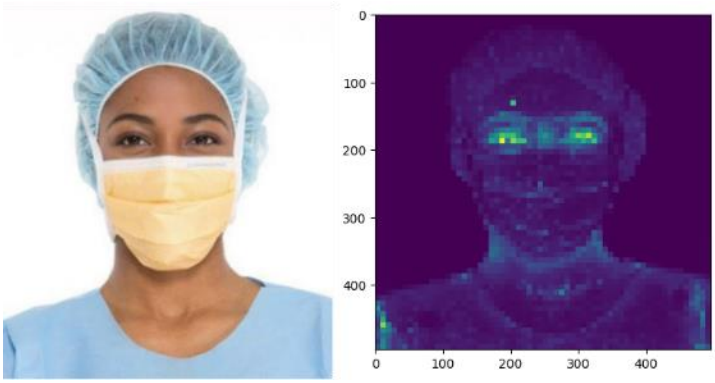

Figure 5. Attention map from the final layer of the ViT model, highlighting facial regions (mainly the eyes) that guided the prediction. Visualized using Hugging Face tools

3.5. Experimental environment setup

The proposed framework was implemented in the Google Colab Pro environment with graphics processing unit (GPU) acceleration (Tesla T4). An overview of the computational setup is illustrated in Figure 6, while the detailed description of the experimental environment is provided in section 4.1.

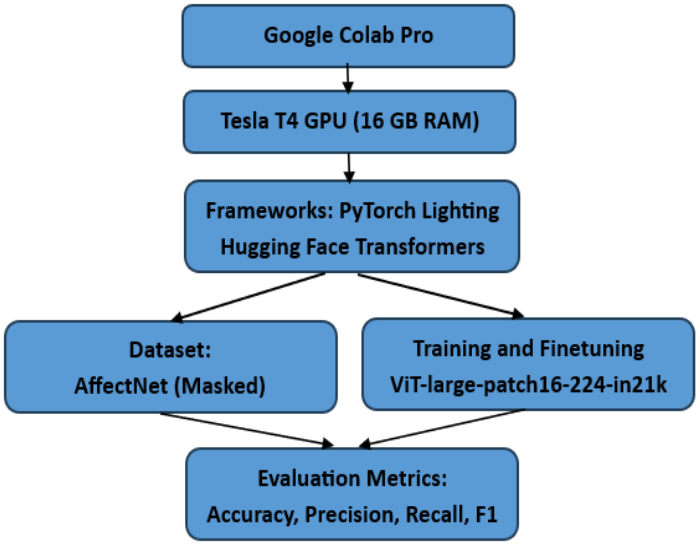

Figure 6. Experimental environment setup: the integration of Google Colab Pro, GPU resources, and the key libraries used to train and evaluate the ViT-based FER model

4. RESULTS AND DISCUSSION

4.1. Experiments environment

The experiments were implemented in the Google Colab Pro environment with GPU acceleration (Tesla T4). Colab Pro provides high-performance resources with pre-installed frameworks such as PyTorch and TensorFlow, along with supporting libraries for preprocessing and visualization. This setup ensured efficient training and reproducibility of the experimental results.

4.2. Experiment dataset

To address the feature loss from face masks, the original eight emotion classes in AffectNet were restructured into five (anger-disgust, fear-surprise, happiness, sadness, and neutral), as shown in Figure 7.

This recategorization ensured recognition of key emotions despite occlusion. The dataset (≈320K images) was split to 85% for training, and 15% for validation.

4.3. Implementation details

In this section, the implementation settings used during the fine-tuning process of the ViT model are detailed. The proposed model was implemented using PyTorch Lightning and fine-tuned in the Google Colab Pro environment with access to a Tesla T4 GPU. We used the pretrained ViT-large-patch16-224-in21k model from Hugging Face and trained it on the masked AffectNet dataset categorized into five emotion classes. Training was conducted for 3 epochs with a learning rate of 2e-5, using the Adam optimizer and mixed precision (16-bit). The dataset was split into 85% for training and 15% for validation, using stratified sampling. Evaluation metrics included accuracy, precision, recall, and F1 score, computed using macro-averaging to ensure fair comparison across all classes.

Figure 7. The original AffectNet classes were merged, based on visual similarity, into five categories: anger-disgust, fear-surprise, happiness, sadness, and neutral

4.4. Choosing the best ViT pretrained model

The first experiment evaluated several pre-trained ViT architectures for masked emotion recognition. Each model was fine-tuned on the dataset to adapt to occlusion effects. The google/ViT-large-patch16-224-in21k model achieved the highest accuracy of 80.8%.

4.5. Comparison with a state-of-the-art CNN model

This experiment compared the best ViT model identified in the previous experiment with ResNet-50. Both models were fine-tuned on the same masked facial dataset and evaluated using precision, recall, accuracy, and F1 score. As shown in Table 2, the ViT model outperforms ResNet-50 in all metrics. This result confirms the ViT superiority for accurately classifying emotions in the presence of face masks.

Table 2. Performance comparison of ViT and ResNet-50

Model / Metric	Accuracy	Precision	Recall	F1 score
ViT-large-patch16-224-in21k	**0.81**	**0.77**	**0.77**	**0.75**
ResNet-50	0.61	0.49	0.51	0.47

4.6. Comparison with other state of the art works

To further validate the effectiveness of the proposed ViT model for FER in the presence of face masks, we conducted a comparative analysis with other state-of-the-art approaches. The comparison includes CNNs and ViT-based architectures. As shown in Table 3, our ViT-based model achieved the highest accuracy (81%) among other compared methods. This underscores the effectiveness of our approach in addressing the challenges of facial expression recognition tasks under condition of facial occlusion.

Table 3. Comparison to other state-of-the-art works

Work	Model	Dataset	Year	Accuracy
Occlusion aware facial expression recognition using CNN with attention mechanism [23]	CNN	FED-RO	2018	66.50%
Face-mask-aware facial expression recognition based on face parsing and vision transformer [29]	ViT	M-FER 2013 and MCK+	2022	66.53%
Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people [30]	CNN	AffectNet	2023	69.3%
Emotion recognition in the times of COVID19: Coping with face masks [11]	ResNet	AffectNet	2022	70%
The proposed ViT-based model	ViT	AffectNet	2024	81%

4.7. Discussion

The experimental results revealed that the vision transformer (ViT) model outperformed state-of-the-art image classification methods, validating the claims made by Dosovitskiy *et al.* [13]. Their groundbreaking work demonstrated the effectiveness of ViT models in capturing spatial relationships and global context, leading to superior performance in tasks such as emotion recognition. Moreover, our findings showed that the ViT model achieved comparable results to the method proposed by Magherini *et al.* [11], despite of using only a smaller subset of AffectNet dataset. This finding highlights the ViT model efficiency in utilizing data resources to produce competitive performance compared to ResNet-50 which requires vast amount of data and extensive computing time.

Furthermore, based on the training environment used in this study (Google Colab Pro with Tesla T4 GPU), the estimated inference time per image for the fine-tuned ViT-large-patch16-224-in21k model is approximately 18–22 milliseconds. This performance indicates that the model is capable of near real-time emotion recognition, making it suitable for deployment in practical applications. The patch-based processing and self-attention mechanism employed by the vision transformer contribute not only to its accuracy, but also to its computational efficiency. These advantages make the model well-suited for environments with limited resources, such as mobile devices or embedded systems, where both speed and performance are critical.

One of the advantages of the ViT model that became evident during our experiments was its relatively faster execution time compared to traditional CNN architectures. Moreover, the self-attention mechanism allows it to capture long-range dependencies in the image, eliminating the need for computationally expensive convolutional operations. This advantage not only accelerates training and inference but also makes ViT models more scalable to larger datasets and computationally constrained environments. Additionally, the ViT model's ability to outperform existing state-of-the-art methods and achieve better performance with reduced database size demonstrates its potential as a powerful image classification tool. As we continue to explore and refine the ViT architecture, we can anticipate further improvements in accuracy, generalization, and efficiency, opening up new possibilities in various computer vision tasks.

5. CONCLUSION

In this paper, we employed ViT for facial emotion recognition under mask occlusion. Traditional facial emotion recognition has been primarily based on visibility of the face. To conduct our experiments, the AffectNet dataset, which contains a large collection of emotional facial images, has been used. A new approach is used to simulate real-world conditions of wearing face masks. We augmented the images in the AffectNet dataset by adding face masks using a custom script. This augmentation was essential to ensure that our ViT-based FER model would be exposed to the challenges posed by partially occluded faces, replicating the conditions we encounter in our daily lives.

Subsequently, we finetuned and evaluated our ViT-based model on this augmented dataset. The results of our experiments were quite promising, as our proposed method achieved an accuracy of 81%. This finding demonstrates the remarkable capability of ViT models to accurately recognize emotions even when the face is partially occluded by a mask. This is particularly significant in the context of our current times, where mask-wearing is prevalent and essential for public health.

In order to evaluate our proposed model, we employed various evaluation metrics, such as accuracy, F1-score, and recall. These metrics gave us more qualitative information on the model's ability to predict positive emotions, as well as on the distribution of the emotions in the dataset. Moreover, we compare the efficiency of our proposed ViT-based model with other state-of-the-art methods for masked facial emotion recognition. The results showed that the ViT-based model outperformed other techniques in the field of FER application. For future work, the FER system can be improved by optimizing the ViT model for masked faces, using larger and more diverse datasets to improve generalization, and exploring how the trained ViT model can be adapted to other tasks like facial expression analysis.

FUNDING INFORMATION

This paper is partially supported by the dean of higher studies and scientific research at Islamic University of Gaza.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ashraf Yunis Maghari		✓				✓		✓	✓	✓	✓	✓		
Ameer M. Telbani	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] Y. L. Tian, T. Kanade, and J. F. Conn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001, doi: 10.1109/34.908962.

[2] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: a survey," *Symmetry*, vol. 11, no. 10, p. 1189, Sep. 2019, doi: 10.3390/sym11101189.

[3] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 15, pp. E1454–E1462, 2014, doi: 10.1073/pnas.1322355111.

[4] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2022, doi: 10.1109/TAFFC.2020.2981446.

[5] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: review and insights," *Procedia Computer Science*, vol. 175, pp. 689–694, 2020, doi: 10.1016/j.procs.2020.07.101.

[6] H. P. Wei, Y. Y. Deng, F. Tang, X. J. Pan, and W. M. Dong, "A comparative study of CNN-and transformer-based visual style transfer," *Journal of Computer Science and Technology*, vol. 37, no. 3, pp. 601–614, 2022, doi: 10.1007/s11390-022-2140-7.

[7] K. Feng and T. Chaspari, "A review of generalizable transfer learning in automatic emotion recognition," *Frontiers in Computer Science*, vol. 2, no. 9, Feb. 2020, doi: 10.3389/fcomp.2020.00009.

[8] L. Zhang, B. Verma, D. Tjondronegoro, and V. Chandran, "Facial expression analysis under partial occlusion: A survey," *ACM Computing Surveys*, vol. 51, no. 2, pp. 1–49, 2019, doi: 10.1145/3158369.

[9] B. Yang, J. Wu, and G. Hattori, "Facial expression recognition with the advent of face masks," in *ACM International Conference Proceeding Series*, 2020, pp. 335–337, doi: 10.1145/3428361.3432075.

[10] G. Castellano, B. De Carolis, and N. Macchiarulo, "Automatic facial emotion recognition at the COVID-19 pandemic time," *Multimedia Tools and Applications*, vol. 82, no. 9, pp. 12751–12769, 2023, doi: 10.1007/s11042-022-14050-0.

[11] R. Magherini, E. Mussi, M. Servi, and Y. Volpe, "Emotion recognition in the times of COVID19: Coping with face masks," *Intelligent Systems with Applications*, vol. 15, p. 200094, 2022, doi: 10.1016/j.iswa.2022.200094.

[12] Q. Yang, Y. Zhang, W. Dai, and S. J. Pan, *Transfer learning*. Cambridge University Press, 2020.

[13] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[14] A. F. Alnabih and A. Y. Maghari, "Arabic sign language letters recognition using vision transformer," *Multimedia Tools and Applications*, vol. 83, no. 34, pp. 81725–81739, 2024, doi: 10.1007/s11042-024-18681-3.

[15] Z. Huang, Y. Yu, and C. Gou, "Driver facial expression recognition based on ViT and StarGAN," in *Proceedings 2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence, DTPI 2021*, 2021, pp. 254–257, doi: 10.1109/DTPI52967.2021.9540071.

[16] X. Fu, "Facial expression recognition based on squeeze vision transformer," in *Proceedings - 2022 International Symposium on Advances in Informatics, Electronics and Education, ISAIEE 2022*, 2022, vol. 22, no. 10, pp. 164–167, doi: 10.1109/ISAIEE57420.2022.00042.

[17] N. S. Fatima *et al.*, "Enhanced facial emotion recognition using vision transformer models," *Journal of Electrical Engineering and Technology*, vol. 20, no. 2, pp. 1143–1152, 2025, doi: 10.1007/s42835-024-02118-w.

[18] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," *Image and Vision Computing*, vol. 26, no. 7, pp. 1052–1067, 2008.

[19] H. Alabsi, A. M. Alashqar, and A. Maghari, "Woman hijab detection using transfer learning," *Journal of Information Systems and Digital Technologies*, vol. 7, no. 1, pp. 145–156, 2025.

[20] H. Feng and J. Shao, "Facial expression recognition based on local features of transfer learning," in *Proceedings of 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2020*, 2020, vol. 1, pp. 71–76, doi: 10.1109/ITNEC48623.2020.9084794.

[21] G. Wang and J. Gong, "Facial expression recognition based on improved LeNet-5 CNN," in *2019 Chinese Control And Decision Conference (CCDC)*, Jun. 2019, pp. 5655–5660, doi: 10.1109/CCDC.2019.8832535.

[22] M. Chen, J. Cheng, Z. Zhang, Y. Li, and Y. Zhang, "Facial expression recognition method combined with attention mechanism," *Mobile Information Systems*, vol. 2021, pp. 1–10, Sep. 2021, doi: 10.1155/2021/5608340.

[23] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, May 2019, doi: 10.1109/TIP.2018.2886767.

[24] B. Yang *et al.*, "Face-mask-aware facial expression recognition based on face parsing and vision transformer," *Pattern Recognition Letters*, vol. 164, pp. 173–182, 2022, doi: 10.1016/j.patrec.2022.11.004.
[25] 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, pp. 18–31, Jan. 2019, doi: 10.1109/TAFFC.2017.2740923.
[26] A. Anwar and A. Raychowdhury, "Masked face recognition for secure authentication," *arXiv preprint arXiv:2008.11104*, 2020.
[27] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, 2017.
[28] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
[29] B. Yang, W. Jianming, and G. Hattori, "Face mask aware robust facial expression recognition during the COVID-19 pandemic," in *Proceedings - International Conference on Image Processing, ICIP*, 2021, vol. 2021-Septe, pp. 240–244, doi: 10.1109/ICIP42928.2021.9506047.
[30] M. Mukhiddinov, O. Djuraev, F. Akhmedov, A. Mukhamadiyev, and J. Cho, "Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people," *Sensors*, vol. 23, no. 3, p. 1080, 2023, doi: 10.3390/s23031080.

BIOGRAPHIES OF AUTHORS

Ashraf Yunis Maghari is an associate professor of computer science at the Islamic University of Gaza (IUG). He holds a Ph.D. in computer vision and image processing from Universiti Sains Malaysia (USM). He has extensive research experience in computer science fields such as data mining, image processing, computer vision, and deep learning. He can be contacted at email: amaghari@iugaza.edu.ps.

Ameer M. Telbani was a lecturer in the Multimedia Department at the Islamic University of Gaza (IUG). He held a Master's degree in information technology from IUG. He had academic experience in multimedia technology disciplines such as image processing, 3D modeling, and animation. His research interests included computer vision, multimedia applications, and deep learning. This article is based on his master's thesis work. He tragically passed away during the war in Gaza before the completion of this research. He can be contacted at email: myashraf2@gmail.com.