*International Journal of Advanced Smart Convergence Vol.12 No.2 25-33 (2023)*
*http://dx.doi.org/10.7236/IJASC.2023.12.2.25*

**IJASC 23-2-3**

# A study on the effectiveness of intermediate features in deep learning on facial expression recognition

KyeongTeak Oh[1], Sun K. Yoo[2†]

[1] *Doctor, Department of Biomedical Engineering, Yonsei University College of Medicine, Korea*
[2] *Professor, Department of Biomedical Engineering, Yonsei University College of Medicine, Korea*

*OKT2704@yuhs.ac, SUNKYOO@yuhs.ac*

***Abstract***

*The purpose of this study is to evaluate the impact of intermediate features on FER performance. To achieve this objective, intermediate features were extracted from the input images at specific layers (FM1~FM4) of the pre-trained network (Resnet-18). These extracted intermediate features and original images were used as inputs to the vision transformer (ViT), and the FER performance was compared. As a result, when using a single image as input, using intermediate features extracted from FM2 yielded the best performance (training accuracy: 94.35%, testing accuracy: 75.51%). When using the original image as input, the training accuracy was 91.32% and the testing accuracy was 74.68%. However, when combining the original image with intermediate features as input, the best FER performance was achieved by combining the original image with FM2, FM3, and FM4 (training accuracy: 97.88%, testing accuracy: 79.21%). These results imply that incorporating intermediate features alongside the original image can lead to superior performance. The findings can be referenced and utilized when designing the preprocessing stages of a deep learning model in FER. By considering the effectiveness of using intermediate features, practitioners can make informed decisions to enhance the performance of FER systems.*

***Keywords:*** *Intermediate Feature, Artificial Intelligence, Facial Expression Recognition*

## 1. Introduction

Facial expression is an observable external expression that conveys affective hints regarding changes in our inner states. The technique of facial expression recognition (FER) is employed in various visual tasks, including driver safety surveillance and video conferencing. Numerous studies have been conducted to improve the classification accuracy of FER. Prior to the advent of deep learning, traditional FER studies primarily relied on handcrafted features such as histograms of oriented gradients (HOGs) [1], local binary patterns (LBPs) [2], and sparse presentation [3].

Deep learning features have proven to be efficient in extracting important patterns from images and have outperformed traditional methods by a significant margin [4-8]. While deep learning-based approaches

Manuscript Received: March. 10, 2023 / Revised: March. 13, 2023 / Accepted: March. 17, 2023
Corresponding Author: SUNKYOO@yuhs.ac
Tel: +82-2-2228-1919, Fax: +82-2-2227-6586
Professor, Department of Biomedical Engineering, Yonsei University College of Medicine, Korea

achieve promising performance on frontal faces with posed facial expressions, they exhibit poor accuracy in FER in the wild, which refers to unconstrained environments. Most of these methods were evaluated on lab-controlled datasets, such as MMI [9] and CK+ [10]. However, when tested on real-world FER datasets like FERPlus [11], RAF-DB [12], and AffectNet [13], which contain facial features affected by various factors, their performance significantly degrades.

In the unconstrained environment, some studies have demonstrated efficiency with in-the-wild databases, utilizing CNN structures. Georgescu et al. employed a convolutional neural network (CNN) architecture with handcrafted features for facial expression recognition [14]. Ruan et al. proposed a feature decomposition and reconstruction learning method to capture expression-specific variations and reconstruct expression features [15]. For basic feature extraction, they utilized ResNet-18 as the backbone network. Liu et al. introduced a clip-aware emotion-rich feature learning network for dynamic facial expression classification [16]. They employed a Deep CNN architecture for feature extraction and incorporated self-attention learning.

In addition to CNN structures, Transformer has demonstrated superior performance in natural language processing [17]. Drawing inspiration from its success, researchers have attempted to apply Transformer to computer vision tasks, leading to the development of Vision Transformer (ViT) by applying a vanilla Transformer to images with a few modifications [18]. ViT has shown outstanding performance compared to CNN-based methods when fully trained on large-scale datasets. Encouraged by this success, ViT has also been applied to the FER task. F. Ma et al. utilized ViT with attentional selective function [19]. They employed ResNet-18 to extract both RGB images and LBP features, using intermediate features from ResNet-18 for the attentional selective function. C. Liu et al. proposed patch attention convolutional ViT for recognizing expressions with occlusion [20]. They utilized pre-trained ResNet-18 as the backbone for extracting intermediate facial features.

In the previously mentioned studies, intermediate features were extracted from pre-trained CNN structures and used as inputs for ViT. Furthermore, some studies using CNN structures also incorporated intermediate features from pre-trained CNN models. Leveraging intermediate features or combining multi-level features from pre-trained networks enhances the representative capacity of the overall network, effectively improving generalization ability and recognition performance [21-22]. However, there has been limited research on which level of features extracted from pre-trained networks can be applied to the FER task. Therefore, this paper aims to investigate the impact of different levels of intermediate features on FER accuracy. The ResNet18, which was commonly used in previous studies, served as the backbone network for feature extraction, and ViT was employed for classification to verify the accuracy of FER.

## 2. Methods

Figure 1 illustrated the main components of the proposed method. In the FERPlus database, raw images were fed into a pre-trained ResNet18. At a specific layer of ResNet18, intermediate feature maps were extracted. A total of 4 feature maps were extracted. The extracted feature maps were utilized as inputs for ViT. Subsequently, ViT was trained using these extracted feature maps. In Figure 1, FM(#) represents the index of the extracted intermediate feature map.

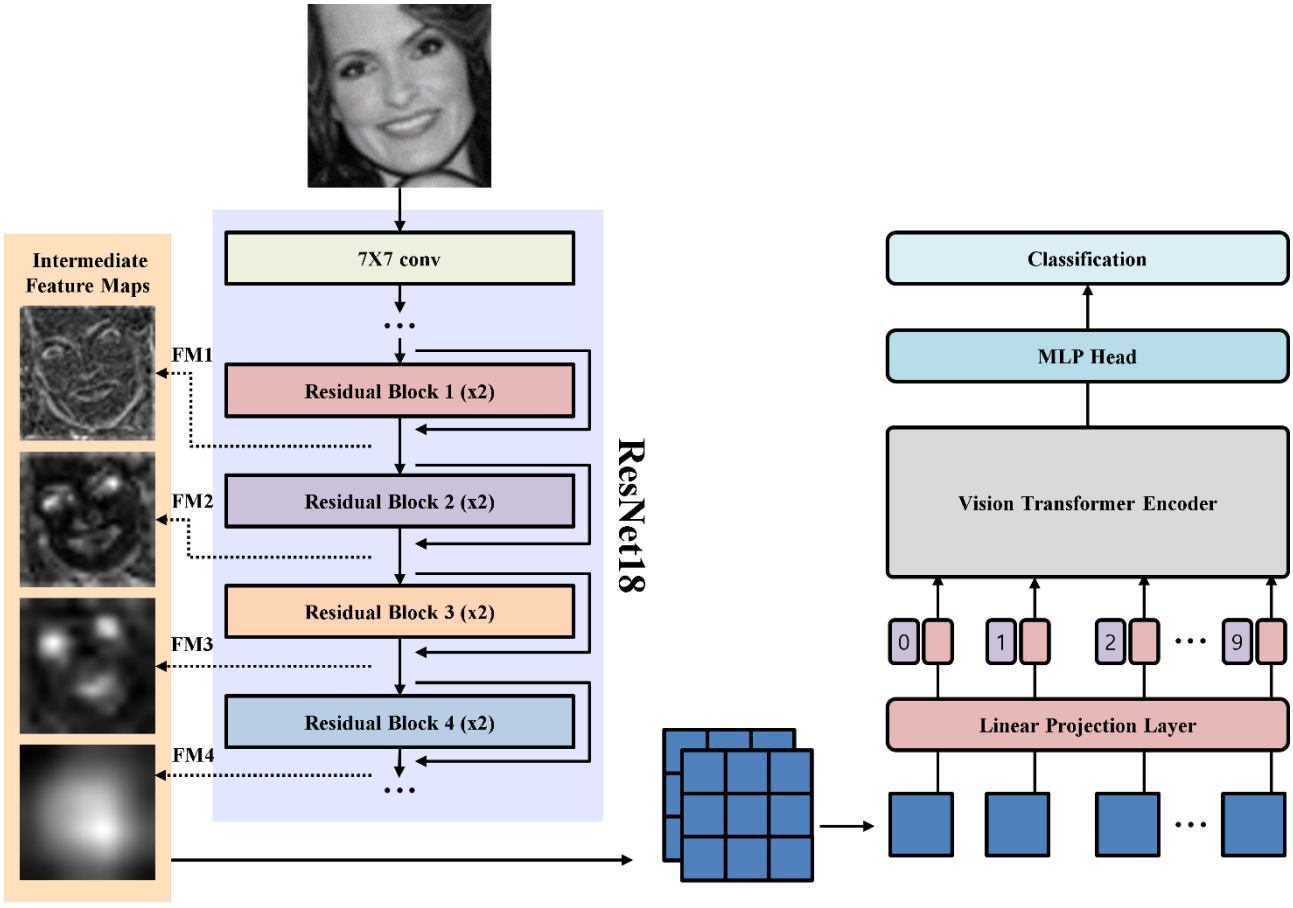


**Figure 1. Overview of the proposed method.**

### 2.1 Dataset

The proposed method was trained using public facial expression dataset, FERPlus database. FERPlus consists of 28,709 training images and 3,589 test images. FERPlus database is composed of labels for anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise. Due to the varying number of training and test images for each class, there is a class imbalance issue. To address this, we simply excluded the relatively low-sample classes of disgust and contempt and trained ViT using the remaining six classes.

### 2.2 Intermediate Feature Extraction

Figure 2 showed the extraction of intermediate feature maps after the first two residual blocks. The residual block was composed of two convolution layers, two batch normalization layers, and a ReLU activation function. After the residual block, there was an additional layer that combines the features before and after the block, followed by a ReLU activation function. The intermediate feature maps were extracted after the ReLU activation function following the two residual blocks.

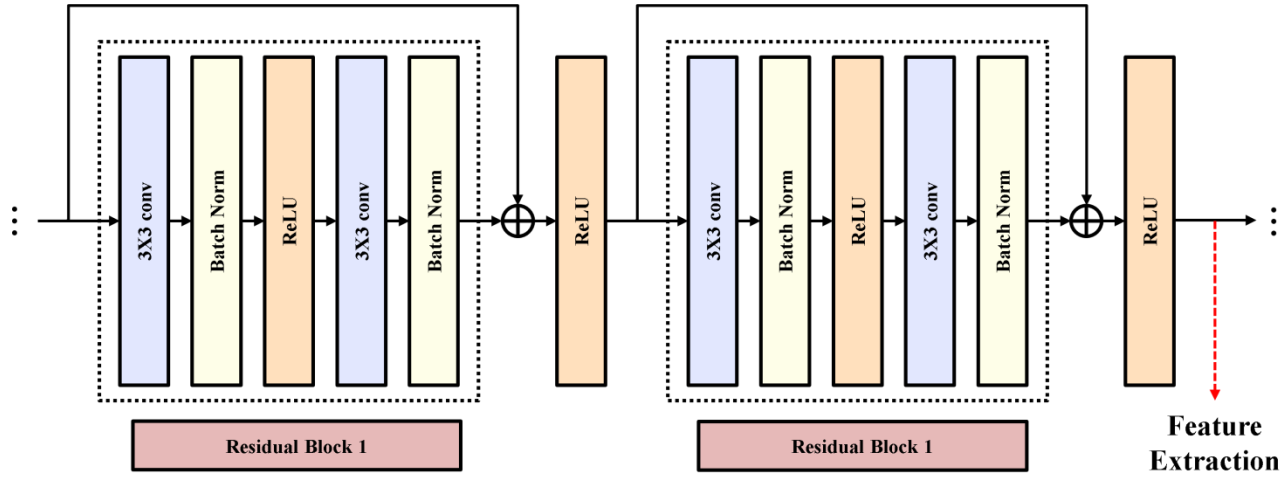


**Figure 2. Example of feature extraction in the first two residual blocks**

### 2.3 Implementation Details

The images in the FERPlus database were resized to a size of 224x224 to match the input size of ResNet18. For fair comparison, we used the ResNet18 with the same weights as the backbone network. To assess the impact of intermediate feature maps on the accuracy of FER, ViT was trained using raw images as well as each individual intermediate feature. Furthermore, we compared the accuracy by using different combinations of the extracted feature maps as input for ViT. We also examined the accuracy when both the original image and the feature maps were used together as input. For the training of ViT, the learning rate was set to 0.005. We used the Adam optimizer [24] for training. The batch size was set to 64, and the ViT was trained for 10,000 steps. The proposed method was implemented using PyTorch and trained the ViT model using an NVIDIA GTX 1080Ti graphics card.

## 3. Results

### 3.1 Comparison between the original images and the extracted intermediate feature maps

Figure 3 showed the original image and intermediate feature maps. When comparing the original image with the intermediate feature maps, it can be seen that the original image appears complex due to the presence of background, hair, or accessories, while the feature maps are simplified, highlighting the facial contours and clearly revealing the facial expressions. Furthermore, as the network gets deeper, the images become more simplified. Specifically, in FM3, it can be observed that the eyes and mouth are prominently activated, indicating their importance in capturing facial expressions. In the case of FM4, the simplification is too significant, making it difficult to discern the specific form. However, it can be observed that the facial region, excluding the background, is activated, indicating its relevance in capturing facial expressions.

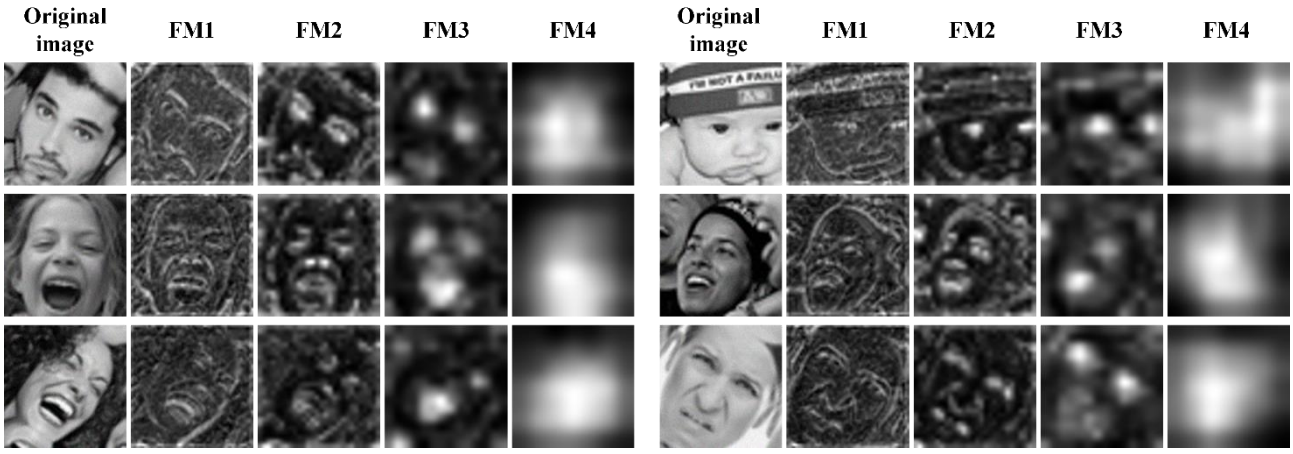


**Figure 3. Original image and the extracted intermediate feature maps**

### 3.2 Comparison of FER accuracy according to different input images

Table 1 compared the training and test accuracy of FER according to different input images. In terms of training accuracy, when the original images were used as input for ViT, the accuracy was 91.32%. However, when using the extracted feature maps as input, excluding FM4, the accuracy was higher. The accuracy for FM2, FM1, and FM3 was 94.35%, 94.13%, and 91.64% respectively. In terms of test accuracy, for all input types, the test accuracy was lower than the training accuracy. When the original images were used as input, the accuracy was 74.68%. The test accuracy was highest at 75.51% when using FM2.

**Table 1. FER accuracy according to different input images**

| Input Image | Training Accuracy (%) | Test Accuracy (%) |
|---|---|---|
| Original Image | 91.32 | 74.68 |
| FM 1 | 94.13 | 74.85 |
| FM 2 | **94.35** | **75.51** |
| FM 3 | 91.64 | 74.73 |
| FM 4 | 66.56 | 42.56 |

### 3.3 Comparison of FER accuracy for different combinations of feature maps

Table 2 compared the training and test accuracy of FER for different combinations of feature maps. The combination of FM1 and FM2 achieved the highest training and test accuracy of 95.32% and 76.58%, respectively. This achieved higher accuracy compared to when each of them was used as input individually. When FM3 and FM4 were input in the combination of feature maps, the training and test accuracies were the lowest, with values of 88.65% and 68.53%, respectively. However, they still showed higher accuracy compared to when they were input individually.

**Table 2. FER accuracy for different combinations of feature maps**

| Feature Map Combination | Training Accuracy (%) | Test Accuracy (%) |
|---|---|---|
| FM1, FM2 | **95.32** | **76.58** |
| FM1, FM3 | 94.78 | 75.35 |
| FM1, FM4 | 93.15 | 73.51 |
| FM2, FM3 | 93.68 | 74.96 |
| FM2, FM4 | 92.68 | 74.01 |
| FM3, FM4 | 88.65 | 68.53 |
| FM1, FM2, FM3 | 94.78 | 76.11 |
| FM1, FM2, FM4 | 94.11 | 75.85 |
| FM1, FM3, FM4 | 92.65 | 75.36 |
| FM2, FM3, FM4 | 91.65 | 74.12 |
| ALL | 93.37 | 74.52 |

### 3.4 Comparison of FER accuracy when using both feature maps and the original image

Table 3 compared the training and test accuracy of FER when using both feature maps and the original image. It was confirmed that utilizing both feature maps and the original image for training the ViT resulted in improved FER accuracy overall compared to training with each individually. When using the the original image and combination of FM2, FM3, and FM4 together, the training and test accuracy reached the highest values of 97.88% and 79.21%, respectively.

**Table 3. FER accuracy when using both feature maps and the original image**

| Feature Map Combination | Training Accuracy (%) | Test Accuracy (%) |
|---|---|---|
| FM1 | 94.52 | 74.91 |
| FM2 | 95.15 | 75.31 |
| FM3 | 95.32 | 75.44 |
| FM4 | 93.87 | 74.31 |
| FM1, FM2 | 95.32 | 76.98 |
| FM1, FM3 | 95.88 | 75.35 |
| FM1, FM4 | 94.15 | 74.51 |
| FM2, FM3 | 97.38 | 78.46 |
| FM2, FM4 | 94.68 | 74.31 |
| FM3, FM4 | 95.15 | 75.53 |
| FM1, FM2, FM3 | 96.78 | 76.41 |
| FM1, FM2, FM4 | 96.41 | 76.85 |
| FM1, FM3, FM4 | 97.05 | 77.17 |
| FM2, FM3, FM4 | **97.88** | **79.21** |
| ALL | 96.96 | 77.31 |

## 4. Conclusion

The aim this study was to evaluate the effectiveness of intermediate features in deep learning for facial expression recognition. Feature maps were extracted from a specific layer of a pre-trained ResNet18 model. To validate the usefulness of intermediate features, the original images and the extracted feature maps were

used as inputs to the ViT to compare the training and test accuracy of FER. The experimental results are as follows.

Using the extracted feature maps, excluding the feature map from the last layer, resulted in higher FER accuracy compared to using the original input images. This can be interpreted as the extracted feature maps simplifying the information related to the background, hair, or accessories, and focusing more on facial expressions, leading to higher accuracy compared to the original input images.

When using a combination of feature maps as input, the combination of FM1 and FM2 resulted in the highest FER accuracy, while incorporating other feature maps in the combination led to lower FER accuracy. This indicates that the combination of FM1, which highlights edge features, and FM2, which mainly captures facial expression-related areas, provides ViT with the most informative features for facial expression classification.

When the original image and the feature maps were used together as inputs, the combination of the original image with FM2, FM3, and FM4 achieved the highest FER accuracy. Contrary to the previous results, it was observed that including FM2 and FM3 instead of FM1 led to higher FER accuracy. This can be interpreted as an improvement in FER performance when the original image is used along with a combination of feature maps that exhibit activation in areas related to facial expressions, compared to the use of FM1 alone, which displays edge components in the original image. Furthermore, it was observed that adding FM4, which exhibits activation in the entire facial region, in addition to the use of FM2 and FM3, resulted in higher accuracy. Additionally, overall accuracy increased when the original image and feature maps were used together. This could be interpreted as the additional information from the feature maps providing attention during the training process, leading to improved accuracy.

A limitation of this study is that the training and test was performed using a single database when evaluating the effectiveness of intermediate features. This does not guarantee the accuracy when unseen data with different characteristics is used as input. Subsequent research should involve training on one database and validating on another database to demonstrate the effectiveness of the proposed method. This approach would help validate the usefulness of the method across different datasets. Another limitation is that to address the data imbalance issue, two classes were simply ignored, which may have an impact on the overall performance and fairness of the model. Using augmentation techniques like RandAugment [25] or employing methods that enhance the generalization capabilities are expected to partially address this issue and potentially improve the model's accuracy.

## Acknowledgement

This work was supported by the Industrial Technology Innovation Program (No. 20012603, Development of Emotional Cognitive and Sympathetic AI Service Technology for Remote (Non-face-to-face) Learning and Industrial Sites) funded By the Ministry of Trade, Industry and Energy (MOTIE, Korea).

## References

[1] Dalal, N., & Triggs, B., "Histograms of oriented gradients for human detection." In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, pp. 886-893, Jun 2005. DOI: https://doi.org/ 10.1109/CVPR.2005.177

[2] Shan, C., Gong, S., & McOwan, P. W., "Robust facial expression recognition using local binary patterns." In IEEE International Conference on Image Processing 2005, vol. 2, pp. II-370, Sep 2005. DOI: https://doi.org/ 10.1109/ICIP.2005.1530069

[3] Lee, S. H., Plataniotis, K. N., & Ro, Y. M., "Intra-class variation reduction using training expression images for sparse representation based facial expression recognition." IEEE Transactions on Affective Computing, Vol. 5, Issue 3, pp.340-351, Aug 2014.
DOI: https://doi.org/ 10.1109/TAFFC.2014.2346515

[4] Li, Y., Zeng, J., Shan, S., & Chen, X., "Occlusion aware facial expression recognition using CNN with attention mechanism." IEEE Transactions on Image Processing, Vol. 28, No. 5, pp.2439-2450, Dec 2018
DOI: https://doi.org/ 10.1109/TIP.2018.2886767

[5] Saeed, S., Shah, A. A., Ehsan, M. K., Amirzada, M. R., Mahmood, A., & Mezgebo, T., "Automated facial expression recognition framework using deep learning." Journal of Healthcare Engineering, Vol. 2022, Mar 2022.
DOI: https://doi.org/10.1155/2022/5707930

[6] Zhi, R., Zhou, C., Li, T., Liu, S., & Jin, Y., "Action unit analysis enhanced facial expression recognition by deep neural network evolution." Neurocomputing, Vol. 425, pp.135-148, Feb 2021.
DOI: https://doi.org/10.1016/j.neucom.2020.03.036

[7] Liang, D., Liang, H., Yu, Z., & Zhang, Y., "Deep convolutional BiLSTM fusion network for facial expression recognition." The Visual Computer, Vol. 36, pp.499-508, Feb 2020.
DOI: https://doi.org/10.1007/s00371-019-01636-3

[8] Sun, N., Li, Q., Huan, R., Liu, J., & Han, G., "Deep spatial-temporal feature fusion for facial expression recognition in static images." Pattern Recognition Letters, Vol. 119, pp.49-61, Mar 2019.
DOI: https://doi.org/10.1016/j.patrec.2017.10.022

[9] Valstar, M., & Pantic, M., "Induced disgust, happiness and surprise: an addition to the mmi facial expression database." In Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, p. 65, May 2010.

[10] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I., "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." In 2010 ieee computer society conference on computer vision and pattern recognition-workshops, pp. 94-101, Jun 2010.
DOI: https://doi.org/10.1109/CVPRW.2010.5543262

[11] Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z., "Training deep networks for facial expression recognition with crowd-sourced label distribution." In Proceedings of the 18th ACM international conference on multimodal interaction, pp. 279-283, Oct 2016.
DOI: https://doi.org/10.48550/arXiv.1608.01041

[12] Li, S., Deng, W., & Du, J., " Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition." IEEE Transactions on Image Processing, Vol. 28, Issue 1, pp. 356-370, Jan 2019.
DOI: https://doi.org/ 10.1109/TIP.2018.2868382

[13] Mollahosseini, A., Hasani, B., & Mahoor, M. H., "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, Aug 2017.
DOI: https://doi.org/ 10.1109/TAFFC.2017.2740923

[14] Georgescu, M. I., Ionescu, R. T., & Popescu, M., "Local learning with deep and handcrafted features for facial expression recognition.", IEEE Access, Vol. 7, pp.64827-64836, May 2019.
DOI: https://doi.org/ 10.1109/ACCESS.2019.2917266

[15] Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C., & Wang, H., "Feature decomposition and reconstruction learning for effective facial expression recognition." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.7660-7669, Apr 2021.
DOI: https://doi.org/10.48550/arXiv.2104.05160

[16] Liu, Y., Feng, C., Yuan, X., Zhou, L., Wang, W., Qin, J., & Luo, Z., "Clip-aware expressive feature learning for video-based facial expression recognition." Information Sciences, Vol. 598, pp.182-195, Jun 2022.
DOI: https://doi.org/10.1016/j.ins.2022.03.062

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., "Attention is all you need." Advances in neural information processing systems, Jun 2017.
DOI: https://doi.org/10.48550/arXiv.1706.03762

[18] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., "An image is worth 16x16 words: Transformers for image recognition at scale. ", arXiv, 2010.
DOI: https://doi.org/10.48550/arXiv.2010.11929

[19] Ma, F., Sun, B. and Li, S., "Facial expression recognition with visual transformers and attentional selective fusion." IEEE Transactions on Affective Computing, pp. 1-1, Oct 2021.
DOI: https://doi.org/10.1109/TAFFC.2021.3122146

[20] Liu, C., Hirota, K., & Dai, Y., "Patch attention convolutional vision transformer for facial expression recognition with occlusion.", Information Sciences, Vol. 619, pp. 781-794, Jan 2023.
DOI: https://doi.org/10.1016/j.ins.2022.11.068

[21] Pong, K. H., & Lam, K. M., "Multi-resolution feature fusion for face recognition." Pattern Recognition, Vol. 47, No. 2, pp.556-567, Feb 2014.
DOI: https://doi.org/10.1016/j.patcog.2013.08.023

[22] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S., "Feature pyramid networks for object detection." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117-2125, Jul 2017.
DOI: https://doi.org/ 10.1109/CVPR.2017.106

[23] He, K., Zhang, X., Ren, S., & Sun, J., "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.770-778. 2016.
DOI: https://doi.org/10.48550/arXiv.1512.03385

[24] Kingma, D. P., & Ba, J., "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980, Dec 2014.
DOI: https://doi.org/10.48550/arXiv.1412.6980

[25] Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V., "Randaugment: Practical automated data augmentation with a reduced search space." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp.702-703. 2020.
DOI: https://doi.org/10.48550/arXiv.1909.13719