Multi Model Ensemble for Compound Expression Recognition

Jun Yu[1], Jichao Zhu[1]*, Wangyuan Zhu[1], Zhongpeng Cai[1],Gongpeng Zhao[1],
Zhihong Wei[1], Guochen Xie[1], Zerui Zhang[1],Qingsong Liu[2], Jiaen Liang[2]
[1]University of Science and Technology of China
[2]Unisound AI Technology Co., Ltd.
{harryjun}@ustc.edu.cn
{jichaozhu,zhuwangyuan,zpcai,zgp0531,weizh588,xiegc,igodrr}@mail.ustc.edu.cn
{liuqingsong,liangjiaen}@unisound.com

Abstract

Compound Expression Recognition (CER) plays a crucial role in interpersonal interactions. Due to the complexity of human emotional expressions, which leads to the existence of compound expressions, it is necessary to consider both local and global facial expressions comprehensively for recognition. In this paper, to address this issue, we propose a solution for compound expression recognition based on ensemble learning methods. Specifically, our task is classification. We trained three expression classification models based on convolutional networks (ResNet50), Vision Transformers, and multi-scale local attention networks, respectively. Then, by using late fusion, integrated the outputs of three models to predict the final result, leveraging the strengths of different models. Our method achieves high accuracy on RAF-DB and in sixth Affective Behavior Analysis in-the-wild (ABAW) Challenge, achieves an F1 score of 0.224 on the test set of C-EXPR-DB.

1. Introduction

In recent years, with advancements in artificial intelligence and human-computer interaction technology, automatic facial expression analysis has become a crucial research tool in fields such as clinical psychology, psychiatry, and cognitive science. It has demonstrated promising results on specific test databases and holds significant commercial prospects in various domains, including human-computer interaction[2, 7], virtual reality [26, 29] , augmented reality[1], smart driving[20] , depression recognition[30]. Companies like Affectiva and Kairos provide real-time assessment and prediction services, such as intelligent advertising and safe driving, by analyzing facial expressions along with other human behaviors such as language, gaze, body movements, and responses in human-computer interaction. Facial expressions have significant research value, however, in daily human life, facial expressions are not always singular in nature. They are composed of various basic expressions. For example, surprise may include both happiness and astonishment. Due to the ambiguity and diversity of expressions, researchers have started paying more attention to the recognition of compound expressions. These compound expressions are more diverse and accurately reflect the complexity and subtlety of our daily emotional expressions.

To facilitate the advancement of compound expression recognition in real-world settings, previous [9–17] and the 6th Affective Behavior Analysis in-the-wild (ABAW) [18] is hosting a challenge CER, aimed at designing a model capable of predicting seven compound expressions. Although the competition organizers have not provided a training dataset or baseline, given that zero-shot capabilities typically lag behind supervised methods, this paper trains models using supervised methods on multiple facial datasets. At this juncture, it constitutes a classification task, with our objective being to develop a model capable of accurately predicting compound expressions.

The current classification task models can generally be categorized into two types: convolutional neural networks (CNNs) [21] and Transformer[27] structures. ResNet [5] is one of the commonly used backbones in CNNs, which calculates advanced features of images by sliding convolutional kernels over them, thus focusing on local features. Vision Transformer[3], on the other hand, is the first backbone in the Transformer family to be widely applied to image modalities and achieve good results. It segments images into patches and then flattens them into sequences. By incorporating position encoding, Vision Transformer embeds the positional information of each patch into the sequence. Through Transformer's Encoder module, Vision Transformer can model the positional relationships of all

*Corresponding author

positions in the image simultaneously, capturing contextual information from different parts of the face and obtaining global information. Both of these models have shown promising performance in facial expression recognition tasks in real-world scenarios.

In this paper, we employ three different models to address the issue of compound expression recognition. The convolutional network used is ResNet50, and the Transformer used is Vision Transformer, each focusing on local and global features, respectively. Additionally, we introduce the Multi-scale and Local Attention Network[28], which combines the two backbone concepts by incorporating multi-scale and attention mechanisms on the feature maps generated by the convolutional network.

Overall, our contributions can be summarized as:

- We integrated different models through Late Fusion by fine-tuning ViT, MANet, and ResNet50 models on the dataset annotated with the same compound expressions.
- In the 6th ABAW competition, our method achieve F1 score of 0.224 on the official test set, ranking third. This result fully confirms the effectiveness and competitiveness of our proposed method.

2. Related Work

Facial Expression Recognition (FER) has evolved from a single expression to a composite expression, [23] constructing a Real World Emotional Face (RAF-CE) database with composite expressions. Meta based multi task learning (MML) combined with AU recognition improves the performance of composite FER. Empirical research has validated the effectiveness of this method, enhancing understanding of subtle differences in complex emotions. [19] built C-EXPR-DB, a real-world dataset consisting of 400 videos annotated with 13 compound expressions. And C-EXPR-NET was proposed on this dataset, which is a multitask learning method based on facial information and AU information. By updating the model through cross entropy and KL divergence, it improves the performance of CER and AU detection (AU-D). The experimental results validated the effectiveness of C-EXPR-NET and demonstrated its generalization ability in new emotion recognition environments. In [25], they proposed a manifold based deep learning network called Deep Bi Manifold CNN (DBM-CNN) , which preserves the local correlation of deep features and the manifold structure of expression labels to learn the feature description of composite expressions.

3. Method

This section we will present our late-fusion ensemble model designed to tackle compound expression recognition. The schematic of the pipeline is depicted in Fig. 1, comprising three distinct models: Vision Transformer (ViT), MANet, and ResNet.

3.1. Data collection

Due to disparities between the ImageNet dataset and facial expression recognition datasets, and the limited availability of datasets with annotations for compound expressions like RAF-DB, we construct a **Unity** based on single-expression annotations from AffectNet [24] and RAF-DB [22], a total of 306,989 facial images.

3.2. Encoders

In this section, we will introduce the three encoders we employ and delve into the details of feature extraction.

3.2.1 Vision Transformer (ViT)

We employed a pre-trained ViT from [6], which underwent pre-training on the ImageNet-1K dataset using self-supervised learning with masked reconstruction, which has validated its efficacy across various image-based tasks, achieving comparable or superior generalization performance compared to supervised training, including emotion recognition for facial expressions. The patch size and stochastic depth rate are set to 16 and 0.1, respectively.The model processes extracted facial images and yields 768-dimensional embeddings for each image, the average of all the output features treat as the image's feature.

3.2.2 Multi-scale and local Attention Network (MANet)

The feature extractor of MANet is used for pre-extracting intermediate-level features. The multi-scale module is employed for integrating features from different receptive fields. The local attention module can guide the network to focus on locally salient features. For each frame image, both the multi-scale module and the local attention module generate final features, which are weighted according to parameters. The initialization parameters of MANet come from pre training weights on RAFDB. Both of the outputs of multi-scale module and local attention module are 512-dimensional. During training, the features from the two scale branches are directly concatenated along the dimension. However, during late fusion, they are weighted and fused with coefficients of 0.6 and 0.4.

3.2.3 Residual Neural Network (ResNet)

ResNet50 is a deep CNN and a member of the ResNet (Residual Network) series. ResNet50 consists of 50 layers of depth and adopts the idea of residual learning. It has achieved good results in tasks such as image classification and object detection, and has become one of the classic backbones. We use the weights trained on the FER2013[4]

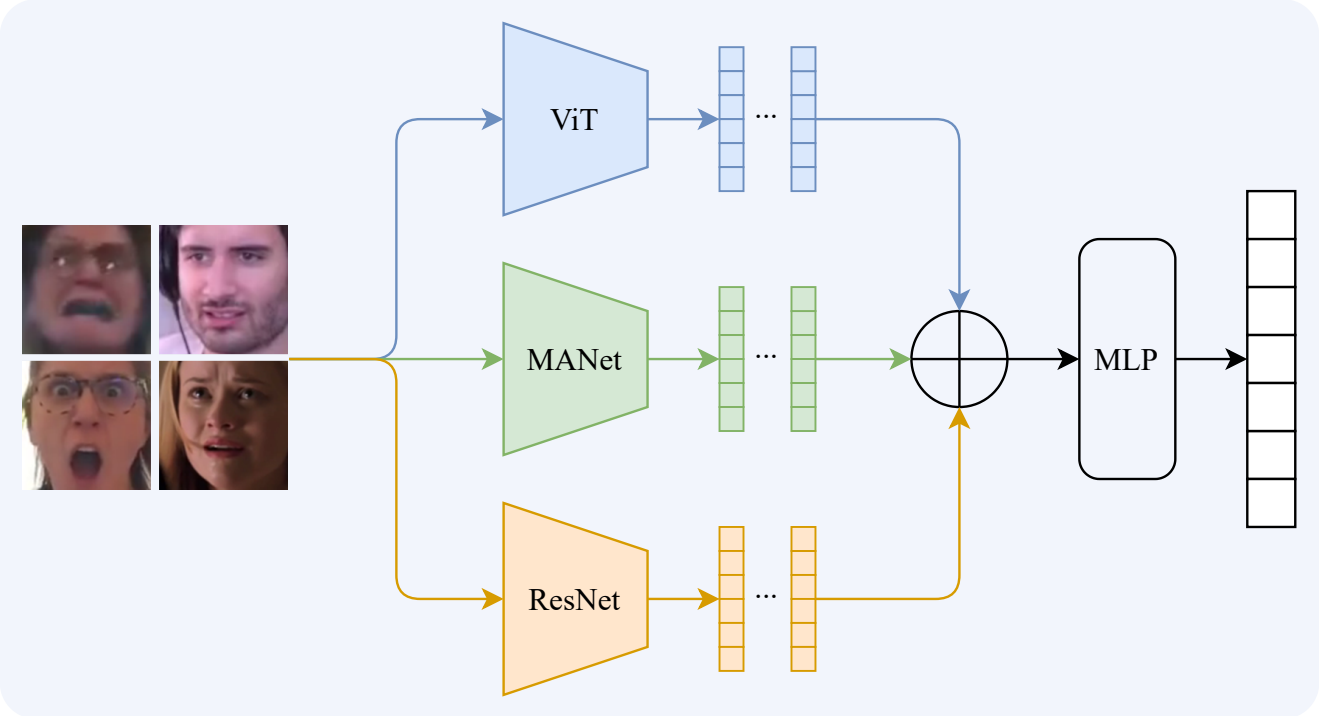

Figure 1. Illustration of the model. Our method consists of three fine-tuned feature extraction networks, ViT, MANet and ResNet50, as well as a multilayer perceptron (MLP) for classification into seven classes.

as initialization parameters, generating a 512 dimensional vector for each image during ensemble.

3.3. Ensemble

We take batched data images x as input to the model, where $x \in \mathbf{R}^{B\times 3\times H\times W}$, with B denoting the batch size, 3 representing the RGB channels, and H and W being the height and width of the images, respectively. Thus, the features after data augmentation can be represented as:

$$\begin{aligned} \text{feature}_1 &= \text{ViT}(x) \in \mathbf{R}^{B\times 768} \\ \text{feature}_2 &= \text{MANet}(x) \in \mathbf{R}^{B\times 1024} \\ \text{feature}_3 &= \text{ResNet}(x) \in \mathbf{R}^{B\times 512} \end{aligned} \tag{1}$$

To enhance model performance by integrating multiple feature maps for richer feature representation, we adopt a late fusion strategy. Specifically, we concatenate the three aforementioned features along a specified dimension, and then input the multiple feature maps into a multi-layer perceptron (MLP) and compute logit for seven compound expression using softmax, as follows:

$$\begin{aligned} \text{feature} &= [\text{feature}_1, \text{feature}_2, \text{feature}_3] \\ \text{logit} &= \text{softmax}(\text{MLP}(\text{feature})) \end{aligned} \tag{2}$$

where $[\cdot, \cdot, \cdot]$ denotes the concatenation operation.

4. Experiments

4.1. Dataset

C-EXPR-DB[19] stands as the largest and most diverse in-the-wild audiovisual database to date. It encompasses 400 videos, totaling around 200,000 frames, meticulously annotated for 12 compound expressions and various affective states. Additionally, C-EXPR-DB includes annotations for continuous valence-arousal dimensions [-1, 1], speech detection, facial landmarks and bounding boxes, 17 action units (facial muscle activation), and facial attributes. In the Compound Expression (CE) Recognition Challenge, a total of 56 unlabeled videos were selected, covering 7 types of compound expressions. The extracted video tags consist of seven compound expression: Fearfully Surprised, Happily Surprised, Sadly Surprised, Disgustedly Surprised, Angrily Surprised, Sadly Fearful, and Sadly Angry.

4.2. Implement Details

4.2.1 Evaluation metric

For Compound Expression Recognition Challenge, the evaluation metric is the F1 Score for seven compound expressions, which is used to measure the prediction accuracy of the model for each expression category, combining both precision and recall, providing a more comprehensive as-

sessment of the model's overall performance, then the metric can be defined as follows:

$$F_1 = \sum_{i=1}^{7} \frac{F_1^i}{7} \tag{3}$$

where F_1^i corresponds to the i-th expression.

4.2.2 Training setting

All experiments in this paper are conducted using the PyTorch and trained on RTX3090 GPU. The input image resolution is consistently set to 224×224. During training, the number of epochs is set to 100. Cross-entropy loss is utilized as the optimization objective, and the Adam[8] optimizer is employed for parameter updates. In pre-train ViT on Unity, in order to improve stability during training optimization, warm-up learning rate is employed to update the learning rate. In the RAF-DB compound expression experiment, the learning rate is set to $5e-5$, and the batch size is set to 128.

4.3. Results

4.3.1 Visual Models' performance on Unity

The Tab. 1 presents a comparison of the performance of ViT and ResNet on facial expression recognition tasks (anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise), along with reporting overall accuracy (acc) and F1 scores. ViT outperforms or is comparable to ResNet in recognizing expressions of happiness, neutrality, and sadness. However, ResNet achieves slightly higher accuracy in recognizing the surprise expression. For other expressions, the performance difference between the two models is not significant. Notably, ResNet outperforms ViT by 11.02% in accuracy on the contempt expression, indicating a clear advantage for ResNet. This may suggest that ResNet is more sensitive to capturing specific details or certain local features, which may be prominent in the contempt expression, such as eye movements or micro-expressions.

4.3.2 Visual Models' performance on RAF-DB (CE)

Fine-tune the top model on the Unity dataset with composite labels on RAF-DB. The Tab. 2 illustrates the performance comparison of ViT, MANet and ResNet in recognizing compound expressions on RAF-DB validation set. For each compound expression, the table lists the recognition accuracy of the three models, as well as their accuracy and F1 score. In recognizing the Happily Surprised expression, ViT leads with an accuracy of 92.59%. In the recognition of the Sadly Surprised expression, ResNet significantly outperforms ViT and MANet with an accuracy of 55.56%, while the latter two have accuracies of 38.89%.

Single expression	ViT(%)	ResNet(%)
Anger	66.62	66.31
Contempt	11.02	17.84
Disgust	45	46.52
Fear	49.48	54.53
Happiness	95.91	93.18
Neutral	85.85	76.53
Sadness	74.95	70.86
Surprise	62.97	65.98
acc	70.2	68.78
F1	64.48	62.37

Table 1. Visual Models' performance on Unity

Compound Expression	ViT	MANet	ResNet
Angrily Surprised	60.53	52.63	55.26
Disgustedly Surprised	51.43	37.14	60
Fearfully Surprised	80.17	81.03	75.86
Happily Surprised	92.59	89.63	85.93
Sadly Angry	84.85	75.75	84.85
Sadly Fearful	72.72	63.64	63.64
Sadly Surprised	38.89	38.89	55.56
acc	78.09	74.06	75.06
F1	70.25	63.57	68.19

Table 2. Results on RAF-DB (CE) validation set.

In terms of overall performance, ViT leads with an accuracy of 78.09%, followed by ResNet with 75.06%, and MANet with 74.06%. In terms of F1 score, ResNet and ViT perform similarly at 68.19% and 70.25%, respectively, while MANet has the lowest performance at 63.57%.

Considering both accuracy and F1 score, ViT demonstrates good balanced performance overall, although it may not be the best in some individual expressions. These results indicate that different network architectures have different strengths and weaknesses in recognizing complex emotional expressions, and no single model is optimal in all cases. This also suggests the need to combine multiple models to improve the accuracy and robustness of expression recognition in practical applications.

4.3.3 Ensemble models' performance on RAF-DB(CE)

The results of the ensemble models with late fusion on the RAF-DB dataset are shown in Tab. 3, and it can be seen that Compared with the single model, the integrated model is more accurate in predicting five compound expressions, namely, Angry Surprised, Disgustedly Surprised, Disgustedly Surprised, Sadly Fearful, and Sadly Surprised. In par-

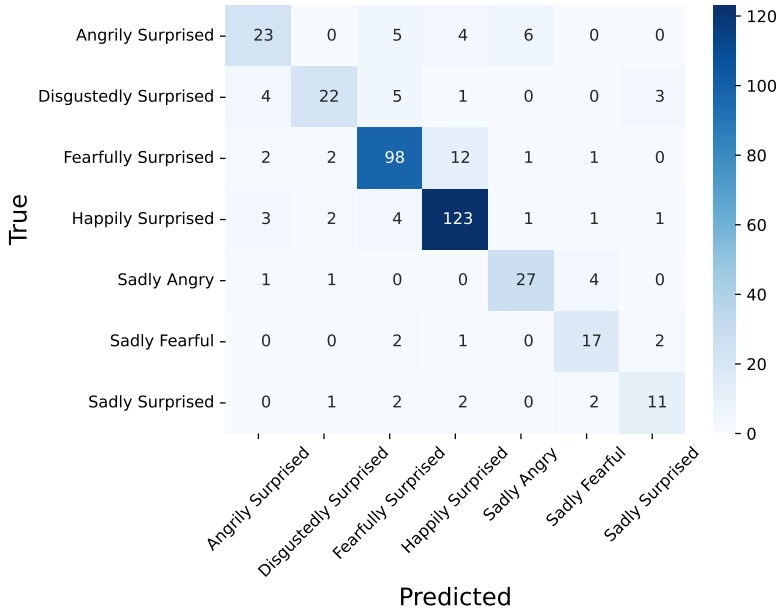

Figure 2. Confusion Matrix of Ensemble Models in RAF-DB Compound Expressions

Compound Expression	Ensemble
Angrily Surprised	60.53
Disgustedly Surprised	62.86
Fearfully Surprised	87.48
Happily Surprised	91.11
Sadly Angry	81.82
Sadly Fearful	77.27
Sadly Surprised	61.11
acc	80.86
F1	74.60

Table 3. Results of Ensemble Models in RAF-DB Compound Expressions

ticular, the expression of Sadly Surprised, which is difficult to identify, is 22.22% higher than that of ViT. In addition, the accuracy rate and F1 score are both better, which is consistent with our idea of using different models to bridge the gap between each other. Fig. 2 illustrates the associated confusion matrix, where the diagonal cells signify the count of accurately predicted samples for each expression.

4.3.4 Performance on Test set

In Table 4, the ranking and F1 scores on the test set are displayed for the final leaderboard. The top two positions are held by multi-modal models, whereas our method is based solely on single-modal images.

Teams	Rank	F1
Netease Fuxi AI Lab	1	0.5526
HSEmotion	2	0.2708
USTC-IAT-United	3	**0.2240**
SUN_CE	4	0.2201
USTC-AC	5	0.1845

Table 4. Performance on Testset

5. Conclusion

In this paper, we propose an ensemble learning-based approach to enhance the compound expression recognition challenge in zero-shot task. We employ three different feature extraction networks: ResNet50, ViT, and MANet. Using a late fusion technique to predict expression, our method achieves promising results on the test set (part) of C-EXPR-DB, demonstrating the effectiveness of ensemble methods with diverse structural models.

6. Acknowledgments

This work was supported by the Natural Science Foundation of China (62276242), National Aviation Science Foundation (2022Z071078001), CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2021-016B, CAAIXSJLJJ-2022-001A), Anhui Province Key Research and Development Program (202104a05020007), USTC-IAT Application Sci. & Tech. Achievement Cultivation Program (JL06521001Y), Sci. & Tech. Innovation Special

Zone (20-163-14-LZ-001-004-01).

References

[1] Vivek Bhardwaj, Anshal Joshi, G. L. Bajaj, Vikrant Sharma, Aayush Rushiya, and Sharanya Bharghavi. Emotion detection from facial expressions using augmented reality. *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1–5, 2023. 1

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 1

[4] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, 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*, pages 117–124. Springer, 2013. 2

[5] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 1

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021. 2

[7] Hiroshi Ishiguro, Tetsuo Ono, Michita Imai, Takeshi Maeda, Takayuki Kanda, and Ryohei Nakatsu. Robovie: an interactive humanoid robot. *Industrial Robot-an International Journal*, 28:498–503, 2001. 1

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[9] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022. 1

[10] Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. *arXiv preprint arXiv:2207.01138*, 2022.

[11] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.

[12] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021.

[13] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021.

[14] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019.

[15] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019.

[16] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800, 2020.

[17] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 1

[18] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (abaw) competition. *arXiv preprint arXiv:2402.19344*, 2024. 1

[19] Dimitrios D. Kollias. Multi-label compound expression recognition: C-expr database & network. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5589–5598, 2023. 2, 3

[20] Wanzeng Kong, Lingxiao Zhou, Yizhi Wang, Jianhai Zhang, Jianhui Liu, and Shenyong Gao. A system of driving fatigue detection based on machine vision and its application on smart device. *J. Sensors*, 2015:548602:1–548602:11, 2015. 1

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012. 1

[22] Shan Li, Weihong Deng, and Junping Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593, 2017. 2

[23] Ximan Li, Weihong Deng, Shan Li, and Yong Li. Compound expression recognition in-the-wild with au-assisted meta multi-task learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5735–5744, 2023. 2

[24] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10:18–31, 2017. 2

[25] Li Shang and Weihong Deng. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision*, 127:884 – 906, 2018. 2

[26] Luma Tabbaa, Ryan Searle, Saber Mirzaee Bafti, Md Moinul Hossain, Jittrapol Intarasisrisawat, Maxine Glancy, and

Chee Siang Ang. Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5 (4), 2022. 1

[27] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017. 1

[28] Zengqun Zhao, Qingshan Liu, and Shan Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021. 2

[29] Lim Jia Zheng, James Mountstephens, and Jason Teo. Eye fixation versus pupil diameter as eye-tracking features for virtual reality emotion classification. *2021 IEEE International Conference on Computing (ICOCO)*, pages 315–319, 2021. 1

[30] Wenbo Zheng, Lan Yan, and Fei-Yue Wang. Two birds with one stone: Knowledge-embedded temporal convolutional transformer for depression detection and emotion recognition. *IEEE Transactions on Affective Computing*, 14:2595–2613, 2023. 1