Emotion Recognition Using Transformers with Random Masking

Seongjae Min, Junseok Yang, Sejoon Lim*
Kookmin University, Korea
{mugun19, 01079421063, lim}@kookmin.ac.kr

Abstract

In recent years, deep learning has achieved innovative advancements in various fields, including the analysis of human emotions and behaviors. Initiatives such as the Affective Behavior Analysis in-the-wild (ABAW) competition have been particularly instrumental in driving research in this area by providing diverse and challenging datasets that enable precise evaluation of complex emotional states. This study leverages the Vision Transformer (ViT) and Transformer models to focus on the estimation of Valence-Arousal (VA), which signifies the positivity and intensity of emotions, recognition of various facial expressions, and detection of Action Units (AU) representing fundamental muscle movements. This approach transcends traditional Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) based methods, proposing a new Transformer-based framework that maximizes the understanding of temporal and spatial features. The core contributions of this research include the introduction of a learning technique through random frame masking and the application of Focal loss adapted for imbalanced data, enhancing the accuracy and applicability of emotion and behavior analysis in real-world settings. This approach is expected to contribute to the advancement of emotional computing and deep learning methodologies.

1. Introduction

Recently, deep learning has undergone significant changes in various fields such as computer vision, natural language processing, and especially in analyzing human emotions and behaviors. One of the key developments in this field is the Affective Behavior Analysis in-the-wild (ABAW) competition held by Kollias et al. [6–17, 24] These competitions facilitate research by providing diverse and challenging datasets such as AffWild2, C-EXPR-DB, and Hume-Vidmimic2, encouraging the development of models capable of accurately assessing complex emotional states. These models provide keys through Valence-Arousal (VA) estimation, facial expression (EXPR) recognition, and Action Unit (AU) detection, which are essential components in understanding human emotions.

In the field of emotional analysis, VA estimation provides the foundation by quantifying the positivity (Valence) and intensity (Arousal) of emotions, while facial expression recognition focuses on classifying facial expressions into distinct emotions. Furthermore, Action Unit detection emphasizes identifying the basic muscle movements that constitute these expressions, offering finer details in interpreting emotional states.

Recent studies have embraced various deep learning approaches, including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), achieving notable success. Additionally, the emergence of transformer models [21] has introduced a new paradigm in understanding temporal and spatial features, expanding the limits of how machines can interpret human emotions and states.

This research builds on these advancements, proposing a new learning framework that utilizes temporally ordered pairs of masked features derived from facial expressions, Action Units, and valence-arousal indicators. By integrating advancements in feature extraction and sequence modeling, we aim to refine the accuracy and applicability of emotional and behavioral analysis in real-world environments and contribute to the evolving landscape of emotion computing and deep learning methodologies.

The main contributions of this study are as follows:

- Introduction of random frame masking learning technique: This study proposes a new learning method that improves the generalization ability of emotion recognition models by randomly masking selected frames.

- Application of Focal loss to imbalanced data: By using Focal loss, we have significantly improved the performance of the model in addressing the imbalance problem in facial expression recognition and Action Unit detection.

*Sejoon Lim is the corresponding author.

2. Related Works

The advancement of deep learning has brought significant changes to the study of human emotional behavior as well. The Affective Behavior Analysis in-the-wild (ABAW) competition has been a tremendous contribution to driving such needed changes and pushing the field forward. ABAW provides a wide variety of datasets, including Aff-Wild2 and C-EXPR-DB, for challenges and research opportunities. In this direction, apart from the Hume-Vidmimic2 dataset, it proposes a challenge with a number of tasks.

2.1. Valence-Arousal Estimation

Valence-Arousal Estimation is a type of emotional analysis that gives an emphasis on forecasting the Valence and Arousal of the persons. Valence is referred to as a characteristic of either positivity or negativity of the emotions. An increase in Valence will symbolize an increase in positive emotion, while a reduction in Valence will show negative emotions. Greater Arousal would indicate that the emotions were more actively energized, while lesser Arousal would mean that the emotions were cool and composed. Recent studies have been doing quite well with performance in [20] using CNNs and LSTMs. Some recent progress has been reported in the application of transformer models as well.

2.2. Expression Recognition

The task for Expression Recognition is a mutually exclusive class recognition problem. Each frame of the video should be classified to one of the defined categories: Neutral, Anger, Disgust, Fear, Happiness, Sadness, Surprise, Other. The research has been carried out using visual and audio information where there exists, to a greater extent, emotional content. References include [25–27]. Nguyen et al.[19] proposed to use only images, and for each of them, a feature vector is extracted using a pretrained network and then supplied to a transformer encoder.

2.3. Action Unit Recognition

In Action Unit Recognition, the determination of specific Action Units (AU) based on the human face's features is done in every frame of the video. It requires facial motion analysis down to the last detail. Yu et al.[23] proposed a feature fusion module based on self-attention, which is responsible for integrating overall facial characteristic and relationship feature between AUs. Zhang et al.[26] and Wang et al. [22] initialized a Masked Autoencoder This enabled the extraction of various general features associated with the face.

3. Approach

In this paper, we propose a network that can learn Valence-Arousal (VA), human expressions, Action Units (AU) and temporally masked features for each frame. So, the first step is the feature extraction for each of the input images. Section 3.1 details the feature extraction step. After the features have been extracted, they are randomly masked, put together into temporal pairs, and then inputted into the transformer encoder. This process is followed by an fully connected (FC) layer to produce the final output. We describe the functioning principle of the transformer classifier module in 3.2. Section 3.3 describes the loss function used for learning. In Figure 1, we show the schema of our whole network.

3.1. Feature Extractor

We utilize a pretrained Vision Transformer (ViT) [4] network in order to extract useful features. Instead of using the 'cls' token from ViT's final output in the conventional manner, we apply average pooling to the output of the last layer based on the method put forth in [1]. This approach saves computational power required during training time by pre-extracting features for the Aff-Wild2 dataset. Additionally, employing a large-scale pretrained network facilitates the extraction of generalized representations that are better adapted to the diverse contexts of the images. This enhances the network's ability to process and analyze the input images' complex emotional expressions and related action units.

3.2. Transformer Classifier

Masked inputs in the Transformer model have been validated in different parts of the Transformer Classifier: GPT [2], BERT [3], MAE [5] Motivated by the works discussed above, we propose designing a Transformer Classifier with features processed in the order of time and an input mask. The proposed encoder is designed to realize the self-attention mechanism that can process efficient sequences of image data. This approach significantly enhances our understanding of changes in facial expressions within a temporal image sequence, which is crucial for accurately recognizing emotions and AU. During training, the temporal feature pairs are input with a certain probability p, having been partially masked beforehand. This method ensures that overfitting is completely avoided, thereby increasing the model's generalization performance.

3.3. Loss function

For AU and Expression, Focal Loss [18] is effective against the imbalanced distribution of data. Focal loss performs very well when learning models from severely class-imbalanced datasets. It is defined as follows:

$$L_{\text{focal}} = -\alpha(1-p_t)^{\gamma}\log p_t \tag{1}$$

Where, p_t denotes the predicted probability, and α and γ are tuning parameters. These hyperparameters assign more

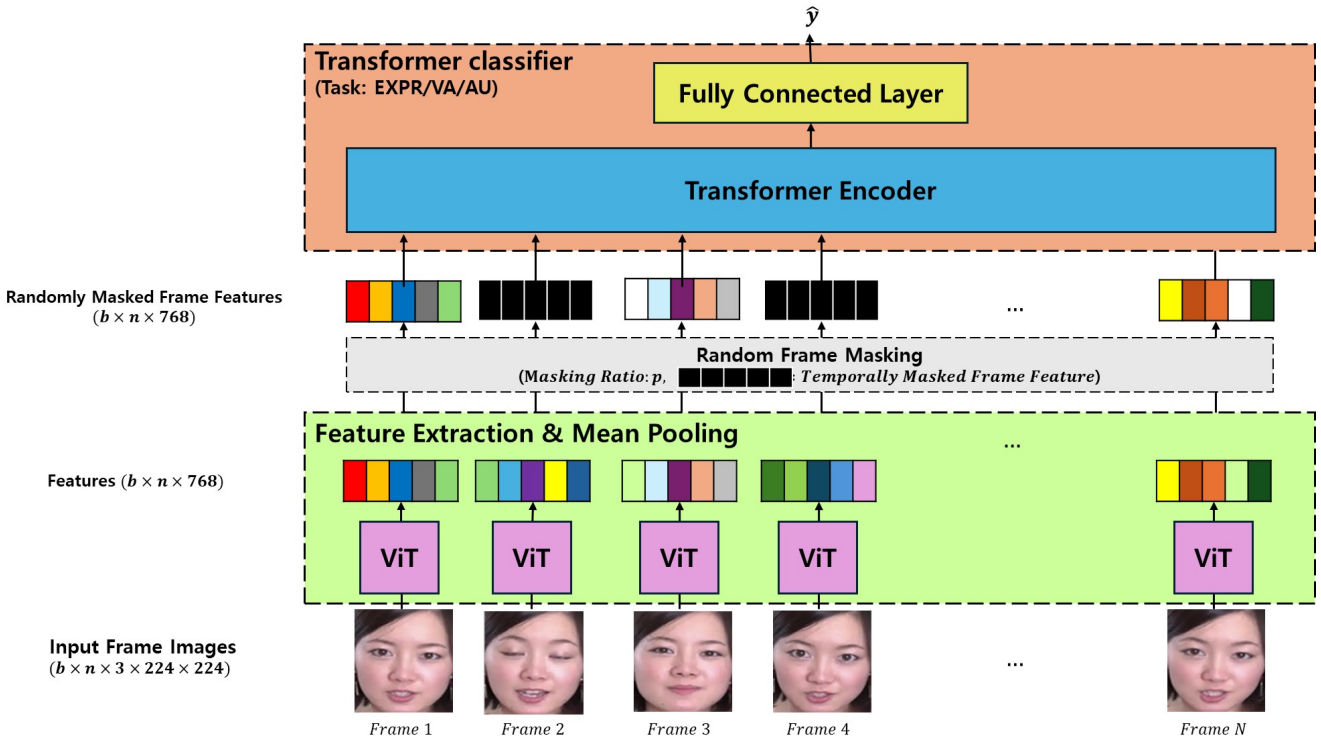

Figure 1. illustrates the comprehensive pipeline of the our model. Initially, a pretrained vision transformer individually extracts features from each input frame image (where b stands for batch size, and n represents sequential length), ensuring a detailed analysis of every frame. To avert the risk of overfitting, these extracted features from each frame are randomly masked. In the final step, a transformer classifier sequentially processes these randomly masked frame features to predict the outcome $\hat{y}$

importance to hard samples and reduce their importance for easier ones, so that the model gets to focus more on the part it struggles within the learning process. This approach significantly boosts the performance of focal loss on imbalanced datasets.

For VA measurement, Concordance Correlation Coefficient (CCC) loss was used. It is computed as follows:

$$L_{\text{ccc}} = 1 - \frac{2\rho\sigma_{xy}}{\sigma_x^2 + \sigma_y^2 + (\bar{x} - \bar{y})^2} \tag{2}$$

Here, ρ is the correlation coefficient between the two variables, and σ_{xy}, σ_x^2, and σ_y^2 represent their respective averages. The CCC loss function measures the concordance between the predicted and actual values, making it an appropriate loss function for predicting emotional states. The foregoing greater importance to the difficult samples and reducing the importance of easy samples allows the model to focus more on parts that should get more concentrating in the learning process. Consequently, Focal loss is the optimal function for substantially improving performance in highly imbalanced datasets. On the other hand, CCC loss function is particulary effective in predicting emotional states because it gives a way that makes it possible to quantify the agreement between the predicted and the target values.

4. Experiments

4.1. Experimental Setup

This study employs the ImageNet21k and Aff-Wild2 datasets to train our model. ImageNet21k, a comprehensive dataset comprising roughly 21,000 classes and 14 million images, is used for pre-training the feature extractor. The Aff-Wild2 dataset, specifically utilized for training the Transformer Encoder, is applied only to cropped facial images.

Our model architecture includes a ViT Base for the feature extraction and a Transformer Classifier with 8 heads and 6 layers, incorporating a dropout rate of 0.2. The model processes sequences with a temporal length of 100 and operates with a batch size of 512. Optimization is achieved through AdamW with a learning rate of 0.0001 and a consistent weight decay of 0.001. Focal Loss is used as the loss function, setting alpha at 0.25 and gamma at 2. A random masking probability of 0.2 is included to enhance training robustness and prevent overfitting. This configuration is designed to efficiently facilitate learning and address the com-

plexities of emotion recognition across diverse and extensive datasets.

4.2. Results on Validation set

Table 1 evaluates the effectiveness of our complete methodology on the Aff-Wild2 validation set across three key challenges: Valence-Arousal (VA) estimation, Expression (EXPR) recognition, and Action Unit (AU) detection, comparing these results with a baseline model to underscore the enhancements provided by our approach.

For VA estimation, we employ the Concordance Correlation Coefficient (CCC), which merges the scores of Valence and Arousal to encapsulate the emotional state. Our method significantly improves with a CCC of 0.39, compared to the baseline's 0.22, indicating a marked enhancement from our model's architecture and training regimen.

EXPR and AU are assessed using the F1 score. In EXPR recognition, our methodology achieves an F1 score of 0.29, slightly higher than the baseline score of 0.25, suggesting an increased capability in recognizing facial expressions. Similarly, AU detection performance improves marginally with an F1 score of 0.40, compared to the baseline's 0.39. These results demonstrate our model's enhanced ability to identify subtle nuances in facial expressions and behavioral units more effectively than the baseline configuration.

Challenge	Metric	Method	Result
VA	CCC	Ours	0.39 (V:0.33, A:0.44)
		Baseline	0.22 (V:0.24, A:0.20)
EXPR	F1 Score	Ours	0.29
		Baseline	0.25
AU	F1 Score	Ours	0.40
		Baseline	0.39

Table 1. Results on Validation set of Aff-Wild2

4.3. Results on Test set

In the VA estimation, EXPR recognition, and AU detection challenges conducted in this study, specific results were derived as described in Tables 2, 3 and 4 respectively. In the VA estimation category, 60 teams participated, and only 10 of them obtained scores that exceeded the baseline and submitted valid results. In the EXPR recognition category, 70 teams participated, and only 10 teams achieved performance that exceeded the baseline and submitted valid results. In the AU detection category, 40 teams participated, and only 7 of them scored higher than the baseline and submitted valid results. In particular, our team ranked 10th in the VA and EXPR categories, respectively, and scored higher than the baseline and was named on the Leaderboard. This shows that unlike most teams using a multi-modal approach that combines image data and audio data, important results could be achieved without additional data by using only image data and applying a random masking technique.

Team	CCC-V	CCC-A	Total Score
Netease Fuxi AI Lab	0.6873	0.6569	0.6721
DeepAVER	0.5418	0.6196	0.5807
CtyunAI	0.5223	0.6057	0.5640
SUN CE	0.5355	0.5861	0.5608
USTC-IAT-United	0.5208	0.5748	0.5478
HSEmotion	0.4925	0.5461	0.5193
KBS-DGU	0.4836	0.5318	0.5077
ETS-LIVIA	0.4198	0.4669	0.4434
CAS-MAIS	0.4245	0.3414	0.3830
Ours	0.2912	0.2456	0.2684
baseline	0.2110	0.1910	0.2010

Table 2. VA Estimation results of the 6th ABAW Competition

Team	F1 Score
Netease Fuxi AI Lab	0.5005
CtyunAI	0.3625
USTC-IAT-United	0.3534
HSEmotion	0.3414
M2-Lab-Purdue	0.3228
KBS-DGU	0.3005
SUN CE	0.2877
AIOBT	0.2797
CAS-MAIS	0.2650
Ours	0.2296
baseline	0.2250

Table 3. EXPR Recognition results of the 6th ABAW Competition

Team	F1 Score
Netease Fuxi AI Lab	0.5601
CtyunAI	0.4941
HSEmotion	0.4878
USTC-IAT-United	0.4840
KBS-DGU	0.4652
M2-Lab-Purdue	0.3832
baseline	0.3650
Ours	0.35

Table 4. AU Detection results of the 6th ABAW Competition

4.4. Ablation Study

The ablation studies specifically focus on the role of the masking component in our model, as detailed in Table 5. When the masking procedure is removed, VA estimation

scores decrease from 0.39 to 0.35. Although these scores still surpass the baseline, this reduction highlights the crucial role of masking in achieving higher accuracy. Similarly, for EXPR and AU detection, the removal of masking leads to reduced scores of 0.27 and 0.38, respectively. These scores represent a decrease from the full model's performance but remain above the baseline, emphasizing the importance of masking in enhancing model performance and generalization.

Additionally, Table 6 presents the empirical performance metrics of the model across individual frames for each task. It is evident that the performance for VA estimation and AU detection improves incrementally with an increase in the number of frames. In contrast, the EXPR task shows a negligible correlation with the frame count. Based on these observations, a temporal length of 100 frames has been determined to be optimal and was therefore employed in the final evaluation submission.

Challenge	Metric	Method	Result
VA	CCC	Ours	0.39 (V:0.33, A:0.44)
		w/o Masking	0.35 (V:0.28, A:0.42)
EXPR	F1 Score	Ours	0.29
		w/o Masking	0.27
AU	F1 Score	Ours	0.40
		w/o Masking	0.38

Table 5. ablation study on the impact of masking for VA, EXPR, and AU Challenges

Challenge	Metric	Frame	Result
VA	CCC	1	0.2868
		10	0.2965
		100	0.3509
		200	0.3906
EXPR	F1 Score	1	0.2683
		10	0.2622
		100	0.2699
		200	0.2599
AU	F1 Score	1	0.3006
		10	0.3522
		100	0.3756
		200	0.3652

Table 6. Ablation study on the impact of frame count for VA, EXPR, and AU Challenges

5. Conclusions

This study explores the use of ViT and Transformer models for VA estimation, EXPR recognition, and AU detection in the 6th ABAW competition. In line with the growing trend towards Transformer-based models, which are increasingly favored over traditional CNN and LSTM approaches, this research introduces a novel methodology utilizing these advanced models. The primary contribution of this work is the introduction of a random frame masking technique during the training process, which significantly improves the models' generalization performance. Notably, this technique has helped achieve top leaderboard rankings in the VA estimation and EXPR recognition tasks. Additionally, the validity of the proposed approach is demonstrated through ablation studies that assess frame-wise performance and the impact of the masking technique.

However, this study is limited to an image-based approach, in contrast to the majority of participating teams that employed multimodal approaches. Furthermore, the research is constrained by the inability to implement image augmentation techniques, highlighting a potential area for improvement. Future research will aim to address these limitations and explore methods to enhance the robustness and generalization capabilities of the models.

6. Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.2022R1F1A107262613) and the Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0020536, HRD Program for Industrial Innovation)

References

[1] Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Better plain vit baselines for imagenet-1k. *arXiv preprint arXiv:2205.01580*, 2022. 2

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[6] 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

[7] Dimitrios Kollias. Abaw: learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2023.

[8] Dimitrios Kollias. Multi-label compound expression recognition: C-expr database & network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2023.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5888–5897, 2023.

[17] 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

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[19] Dang-Khanh Nguyen, Ngoc-Huynh Ho, Sudarshan Pant, and Hyung-Jeong Yang. A transformer-based approach to video frame-level prediction in affective behaviour analysis in-the-wild. *arXiv preprint arXiv:2303.09293*, 2023. 2

[20] Geesung Oh, Euiseok Jeong, and Sejoon Lim. Causal affect prediction model using a past facial image sequence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3550–3556, 2021. 2

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[22] Zihan Wang, Siyang Song, Cheng Luo, Yuzhi Zhou, Shiling Wu, Weicheng Xie, and Linlin Shen. Spatio-temporal au relational graph representation learning for facial action units detection. *arXiv preprint arXiv:2303.10644*, 2023. 2

[23] Jun Yu, Renda Li, Zhongpeng Cai, Gongpeng Zhao, Guochen Xie, Jichao Zhu, Wangyuan Zhu, Qiang Ling, Lei Wang, Cong Wang, et al. Local region perception and relationship learning combined with feature fusion for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5784–5791, 2023. 2

[24] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 1

[25] Su Zhang, Ziyuan Zhao, and Cuntai Guan. Multimodal continuous emotion recognition: A technical report for abaw5. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5763–5768, 2023. 2

[26] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Multimodal facial affective analysis based on masked autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2023. 2

[27] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Leveraging tcn and transformer for effective visual-audio fusion in continuous emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5755–5762, 2023. 2