Ensemble Spatial and Temporal Vision Transformer for Action Units Detection

Ngoc Tu Vu, Van Thong Huynh, Trong Nghia Nguyen, Soo-Hyung Kim*
Department of AI Convergence, Chonnam National University
{tu369,vthuynh,trongnghia7171,shkim}@jnu.ac.kr

Abstract

Facial Action Units detection (FAUs) represents a fine-grained classification problem that involves identifying different units on the human face, as defined by the Facial Action Coding System. In this paper, we present a simple yet efficient Vision Transformer-based approach for addressing the task of Action Units (AU) detection in the context of Affective Behavior Analysis in-the-wild (ABAW) competition. We employ the Video Vision Transformer(ViViT) Network to capture the temporal facial change in the video. Besides, to reduce massive size of the Vision Transformers model, we replace the ViViT feature extraction layers with the CNN backbone (Regnet). Our model outperform the baseline model of ABAW 2023 challenge [8], with a notable 14% difference in result. Our team has achieved a position within the top 5 teams in the ABAW 2023 competition, scoring slightly below the top three and four teams by a narrow margin of 0.27% and 0.43%, respectively.

1. Introduction

Affective computing is a foundation field in Artificial Intelligence that aims to enable machines to recognize, interpret, and respond to human emotions. Recent advances in deep learning and computer vision techniques have enabled significant breakthroughs in the field, but several challenges remain unsolved. In particular, Facial Affect Analysis in the Wild emerge as a notable challenge in recent years. This task plays a crucial role in applications such as Human-Machine Interaction and serves as an initial step for many systems. As such, the Affective Behavior Analysis in the Wild (ABAW) competition [7–16, 32] was organized to address these challenges. Since the first Workshop [32], ABAW has become an important platform for researchers to benchmark their approaches and collaborate on solving Affective Computing problems.

The competition comprises three tasks focusing on detecting and recognizing three commonly-used presentations in affect analysis: Expression, Facial Action Units, and Valence-Arousal. The competition involves three tasks that focus on different affect presentations of affect analysis. The Facial Action Units (AU) detection task utilizes the Unit defined by the Facial Action Coding System (FACS) [5] to capture and interpret facial muscle movements associated with different expressions. The Expression Recognition task, on the other hand, employs categorical and explicit definitions to represent human expressions. Finally, the Valence and Arousal estimation task uses continuous values to describe human emotional states, providing a more nuanced and comprehensive approach to affect the analysis. In this research, we particularly focus on the Action Unit detection task, which is the Multi-labels (12 labels) classification.

The Transformer architecture [27] has gained widespread popularity as a model of choice in the field of Deep Learning. Its successor, the Attention-based model, has emerged as the state-of-the-art approach not only for Natural Language Processing (NLP) tasks but also for achieving significant performance in various Computer Vision problems. In 2021, Dosovitsky et al. introduced the Vision Transformer [4], a model that surpassed other CNN-based methods on the Image Classification task in the ImageNet benchmark [3]. Since then, Transformer-based models have gained momentum in the field of Computer Vision due to their exceptional scalability and SOTA performance. In the ABAW competition, although there are many competitors who have already incorporated transformers as core components in their methods, Video-oriented Vision Transformer have yet to be extensively utilized.

For that reason, in this study, we present a Video Vision Transformer [1] based approach for the Action Units Detection task in the ABAW 2023 challenge. Furthermore, we integrate RegNet as the backbone of our model due to its proven decent performance, and light computing resource requirements [22]. This help us to shorten the training time when applying difference ideas on the model and also have a good balance between performance and efficiency.

Overall, the contribution of this paper can be summarized as following:

*Corresponding author

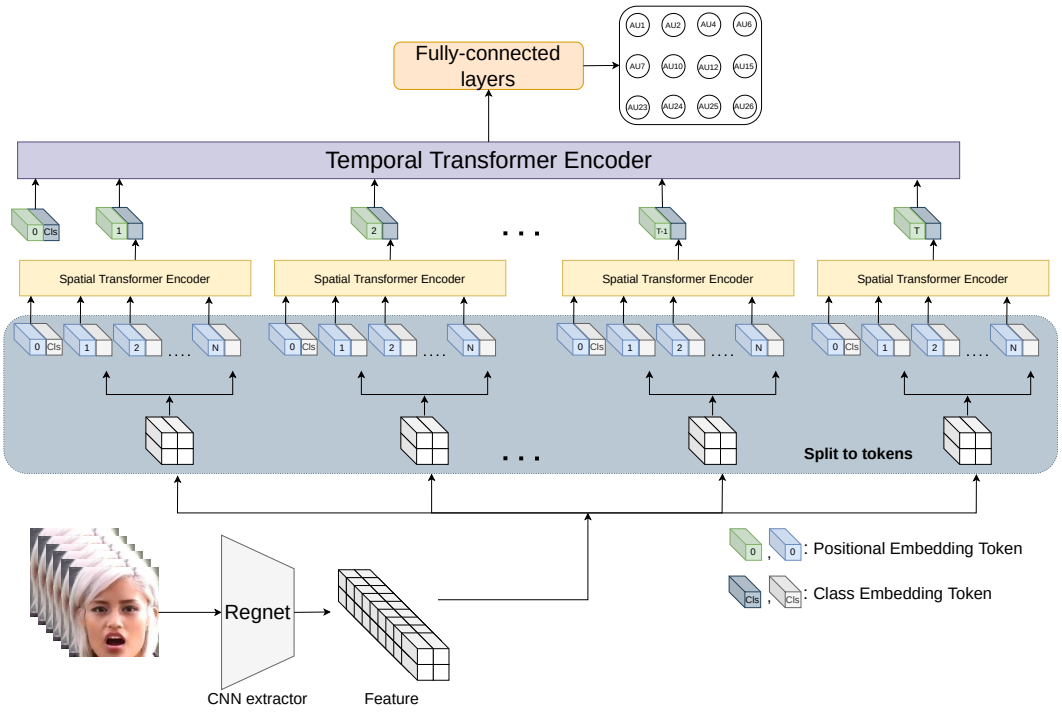

Figure 1. An overview of the action unit detection model.

- Instead of feeding the model raw video, we employ the CNN feature extraction model (Regnet) as the embedding module to get the presentation of video. This methods reduce the size of the model while still keeping the important information, hence help lightening the model.

- We utilize the ViVit model with Ensemble learning scheme for Facial Action Units model. The model outperform the baseline model and show competitive result comparing with the methods of ABAW 2023 competition.

2. Related Work

2.1. Facial Action Units Detection in the wild

Facial affective computing has been a significant challenge in the field of computer vision since its early stage. During this period, popular approaches primarily focus on single-modal affective analysis and statical 2D data processing [2]. In the deep learning era, this problem have been extended to different manner such as spontaneous facial expression [21, 35], multi-modalities, 3D facial expression [35] or automatic affect facial analysis [20]. DISFA [21] is one of the first publicly available databases of spontaneous facial expression with well-annotatated Action Units intensity. One year later, Zhang et al introduce the first spotaneous 3D facial expression database [35]. Until now, many Facial Affective Analysis researchs still uses these two dataset and attain remarkable result [6, 25].

Although having been research for a long time, most Affective Computing research are done in the controlled environment. Therefore, the applicability of these research in real world is quite limited despite of having achieved good results. To address this problem, Zafeiriou et al [32] introduce the first dataset for analysing human affective Behavior in real world scenario. The dataset can be access through the ABAW competition and is still updating each year.

2.2. Vision Transformers in Facial Actions Units Detection

Since the outstanding performance of ViT [4] and other NLP-influenced model in Image Classification and Video Classification, Transformer [27] has been adopted for various Computer Vision tasks. For Action Units Detection, "Facial Action Unit Detection With Transformers" [6] is one of the first study using the Transformer. Following previous Region of Interest attention based methods such as JAA-Net [25] or EAC-Net [17], the study use Transformer's Multi-head Attention Module as the ROI Attention module. While this methods achieve promising results on images, the size of model due to ROI Attention module may not be suitable for videos.

In the previous year's ABAW competition (CVPR 2022) [7], a number of competitors employed Transformer. Specifically, among the top five teams ranked highest, three

out of five teams utilized Transformer models as a core component of their model [22,28,34]. The winner team [34] use Transformer as an fusion module for their Multi-Modal architecture. On the other hand, the fourth places team [28] also use Multi-Modal scheme but Transformer is used for Feature Extraction purpose. While the third places team [22] treat each image feature from CNN extractor as a token and integrate Transformer to be the classification head. However, the Transformer model used in these methods are the original architecture which is not specifically suitable for video processing.

3. Methodology

Following the work of [22], we construct a ViViT-based model [1] for Facial Action Units Detection problems. Our architecture comprises of two core modules: Feature extraction module and Classification module. Overall architecture is showed in Figure 1.

3.1. Feature Extraction

For Feature Extraction module, due to good performance and small model size, we use RegNetY [23] as the backbone. Proposed by Radosavovic et al in 2020, RegNetY is a type of simple and regular convolutional network. The RegNetY design space is defined by three primary parameters: depth, initial width, and slope, and generates a unique block width for each block in the network. Notably, RegNet models are constrained by a linear parameterization of block widths, meaning that the design space only includes models with one specific linear structure. The RegNetY architecture is organized into multiple stages, each consisting of four blocks that collectively form the stem (start), body (main part), and head (end) of the network. Within the body section, multiple stages are defined, with each stage comprised of several blocks. It should be noted that both RegNetX and RegNetY employ a single type of block throughout the network, specifically, the residual bottleneck block with group convolution. However, the authors of RegNetY have introduced an additional Squeeze and Excitation layer to the standard residual block of RegNetX, in order to enhance the representational capacity of the model. The architecture of the residual block utilized in RegNetY is illustrated in Figure 2. This modification enables RegNetY to achieve improved performance over RegNetX, by allowing the model to more effectively capture complex feature representations.

We adopt the Transfer Learning approach, leveraging a pre-trained RegNetY model that has been trained on the ImageNet dataset [3]. To be more specific, we use this pre-trained model as a backbone for training on the ABAW dataset. However, instead of freezing the entire backbone, we just partially unfreeze the last three blocks of the backbone while keeping the first block frozen. This allows us to

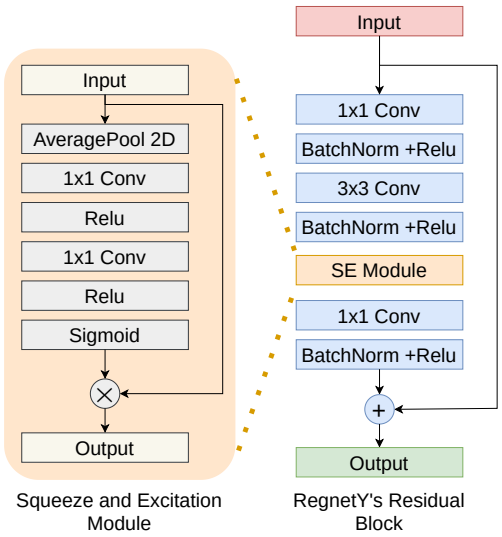

Figure 2. The Residual Block of the RegNetY [23] - feature extractor of our model.

fine-tune the pre-trained model for improved performance on our target task while still benefiting from the pre-existing knowledge gained during training on the large and diverse ImageNet dataset.

3.2. Frames-wise Classification

As mention earlier, we choose Video Vision Transformers [1] as the classification module. Inspired by ViT [4], this model process on the a sequence of spatio-temporal tokens that extracts from the video. Transformer layers employ a distinct approach wherein all pairwise interactions among spatio-temporal tokens are modeled within each layer. As a result, the transformer layer is able to capture long-range dependencies across the entire video sequence from the initial layer itself. However, as the Multi-Headed Self Attention has quadratic complexity with respect to the number the tokens, we reduce the computational complexity by removing the first 4. The selected variant employs a two-stage token encoding approach, consisting of spatial and temporal encoding. The Spatial Transformer Encoder treats each spatial patch of a frame's feature as an individual token, while the Temporal Transformer Encoder treats each frame embedding resulting from the spatial transformer as a separate token (Figure 1). This approach enables efficient and effective capture of spatio-temporal information within the input data, resulting in high-performance results while maintaining real-time processing capabilities.

From a extracted video embedding $V \in R^{B\times T\times E_l\times E_h\times E_w}$, with E_l, E_h, E_w is the length, height and width of each frame embedding, we use Tubelet Embedding of ViVit to convert into sequence of token. Then the token will be fed into Transformer layers com-

Table 1. Results of our methods on Validation Set and on K-fold Cross-validation.

Val Set	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU15	AU23	AU24	AU25	AU26	Avg.
Fold 1	42.45	33.13	41.23	66.20	73.02	72.20	76.60	42.44	33.47	26.67	84.35	32.20	52.00
Fold 2	55.63	44.37	47.78	68.39	76.60	75.24	77.56	31.66	26.43	22.22	86.63	41.56	54.50
Fold 3	52.26	34.73	41.00	55.89	70.03	71.70	68.08	26.81	26.20	18.15	85.80	36.07	48.89
Fold 4	43.62	33.42	45.34	54.84	70.06	64.80	65.64	25.41	15.82	19.60	79.39	22.44	45.03
Fold 5	43.22	30.95	44.32	65.19	74.00	72.83	71.62	21.29	28.27	19.48	85.11	44.34	50.07
Official	55.12	49.35	55.86	67.04	74.41	74.52	73.59	17.74	20.98	18.38	85.35	41.27	52.80

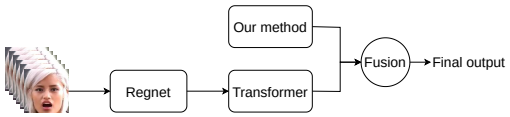

Figure 3. The fusion strategy ensemble our method with the model of Nguyen et al [22].

prises of Multi-Headed Self-Attention (MSA) [27], layer normalisation (LN) and MLP blocks. Each element is formalised in follow equation:

$$y^l = MSA(LN(z^l)) + z^l \quad (1)$$

$$z^{l+1} = MLP(LN(y^l)) + y^l \quad (2)$$

Given that the Video Vision Transformer was originally designed for multi-class classification problems, we make slight modifications to adapt the model for multi-label classification tasks. Specifically, we adjust the model's head to output 12 percentage values, each corresponding to a specific Action Unit. To convert these percentage values into binary values, we establish a threshold value of 0.5. Any percentage value above and equal the threshold was set to 1, while any value below the threshold was set to 0.

3.3. Ensemble models

Our approach involves incorporating the model from [22] into our framework for enhancing model performance (Fig. 3). Specifically, we employ the average fusion technique to combine the prediction scores of each model. Through this ensemble approach, we were able to improve the model's performance in test set by approximately 3% Tab. 1.

3.4. Prediction smoothing

Based on our observations, we identified a notable deviation in the trend of model score predictions. However, this situation do not make sense as the human face just slightly change between frames in the videos. Therefore, we suggest using the smoothing technique to reduce the unstable nature of the model's predictions. Specifically, we separately applied a smoothing technique to the data's ground-truth and the model's score predictions.

We implement the Robin Hood Label Smoothing (RHLS) [26] technique for label smoothing to decrease the model's confidence level. This technique is chosen for its ability to mitigate the imbalance characteristics of the dataset and prevent over-fitting on noisy examples. By reducing the model's confidence, the technique helps to enhance the model's generalization ability and improve its performance on previously unseen data. the Robin Hood Label Smoothing (RHLS) utilize a_i^- and a_i^+ to represent the probabilities of negative and positive values, respectively. With $\widetilde{y_i}$ is the smoothed label and y_i is the original label of the i-th Action Unit of the sample, the RHLS equation is:

$$\widetilde{y_i} = (1 - \frac{\alpha_i^+}{2})y_i + \frac{\alpha_i^-}{2}(1 - y_i) \quad (3)$$

Moreover, α_i^-, α_i^+ is calculated with the positive label frequencies in whole the training set so as to only the majorities class is smoothed:

$$\alpha_i^+ = \beta max(0, \frac{2f_i - 1}{f_i}), \alpha_i^- = \beta max(0, \frac{1 - 2f_i}{1 - f_i}) \quad (4)$$

For score prediction, we use 1D Gaussian smoothing to reduce the score deviation between consecutive frames. After using grid search to choose the best hyper-parameters, we choose the filter kernel size is 51 and the standard deviation values (σ) equal 3.0.

4. Experiments and results

4.1. Dataset

The AU task contains 541 videos that include annotations in terms of 12 AUs, namely AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25, and AU26 of around 2.7M frames and contain 438 subjects, 268 of which are male and 170 female. The dataset have been annotated in a semi-automatic procedure (that involves manual and automatic annotations).

4.2. Experiments setup

The networks were implemented with the Pytorch Lightning toolkit. We trained model by using SGD with learning rate of 0.9 and Cosine annealing warm restarts scheduler [19]. The networks is optimized with Sigmoid Focal

Table 2. Comparision of our method with other participants in ABAW 2023 competition.

Methods	Val Set	Test Set
Baseline	0.39	0.365
Netease Fuxi Virtual Human [33]	0.5667	0.5549
SituTech	-	0.5422
USTC-IAT-United [31]	0.5106	0.5144
SZFaceU [30]	0.543	0.5128
CtyunAI [36]	0.5174	0.4887
HSE-NN-SberAI [24]	0.544	0.4878
USTC-AC [29]	0.6983	0.4811
Nguyen et al [22]	0.5398	0.5002
Regnet-ViT	0.5280	0.4837
Regnet-ViT-Kfold	0.53	0.4914
Regnet-ViT-Smoothing	0.5351	0.4836
Regnet-ViT-Ensemble	**0.5493**	**0.5101**

loss function [18]. The number of frames in each sequence length is set as 256 frames. For the ViViT model, we set number of heads in each Attention module is 8 heads and the hidden dimension of the transformer is 1024. The inputs dimension of ViVit model is the last feature output of Regnet models which is $B \times 256 \times 440 \times 4 \times 4$ with 440 is the number of channels and 4 is the width and height of a token. Besides, we only keep the last 8 Transformer layers of ViVit and remove the rest. The entire training and testing process was conducted on a GTX 3090 GPU.

4.3. Metrics

According to challenge white paper [8], macro F1 Score is the official evaluation criterion for Action Units detection task. Therefore, the performance measure is calculated as the average F1 Score across all 12 AUs:

$$P_{AU} = \frac{(\sum_{au} F_1^a u)}{12} \tag{5}$$

4.4. Results

The results of our methods compared with the baseline and other top-8 methods of previous competition are presented in Tab. 2. According to the result, our methods significantly surpass the baseline methods with a 14% difference. On the other hand, our methods also achieve comparable performance with the highest rank competitors. Compared to other competitors, our model achieve top-5 in terms of performance on the Test Set. Although our model achieved results that were comparable to those of Team USTS-IAT-United and SZFaceU with F1 scores differing by only 0.27% and 0.42%, respectively, it still falls behind the top-performing teams who placed first and second in the competition with F1 scores that are 3% and 4% lower than theirs. The lower performance of our model can be attributed to the limitations of using a pretrained backbone that is not specifically designed for facial recognition tasks. While the top two teams trained their backbones on their own facial datasets, our model utilizes a backbone that was pretrained on ImageNet - a general-purpose dataset. This means that the features learned by our backbone may not be optimized for facial recognition tasks, which could limit our model's performance in the competition. Hence, this limitation may be a motivation for our future work.

Besides, we also report the results of the K-fold cross validation experiments with training set of ABAW challenge and evaluate on Validation set is showed in Tab. 1. As showing in the table, our model performance do not face overfitting problem in particular set and stabilize on the whole dataset. Various approaches of our methods are assessed and compared in Tab. 2. The Ensemble approach proved to be the most effective, yielding the highest scores on both the Validation and Test Sets, with 54.93% and 51.01%, respectively. The second most successful method on the Validation Set was the Smoothing approach, with a score of 53.51%. Nevertheless, this method demonstrated poor performance on the Test Set, resulting in a reduced performance compared to the baseline. This phenomenon can be attributed to the overfitting of the method caused by our hyper-parameter selection process.

5. Conclusion

In this paper, we present the Vision Transformer-based model for AU detection in the ABAW Competition. To reduce the computational burden and improve the effectiveness of our approach on small images, we propose using a CNN-based model for feature extraction instead of relying solely on the long Transformer backbone layers. Our method outperforms the baseline model and achieves competitive results compared to other methods in ABAW 2023 competition.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1A4A1019191) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF2021R1I1A3A04036408).

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF inter-*

national conference on computer vision, pages 6836–6846, 2021. 1, 3

[2] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 3

[4] 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. In *International Conference on Learning Representations*, 2021. 1, 2, 3

[5] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 1

[6] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 2

[7] 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, 2

[8] Dimitrios Kollias. Abaw: learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2023. 1, 5

[9] 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. 1

[10] 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. 1

[11] 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

[12] 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. *arXiv preprint arXiv:2303.01498*, 2023. 1

[13] 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. 1

[14] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1

[15] 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. 1

[16] 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. 1

[17] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 103–110. IEEE, 2017. 2

[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. 5

[19] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 4

[20] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, 10(3):325–347, 2017. 2

[21] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 2

[22] Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. An ensemble approach for facial behavior analysis in-the-wild video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2512–2517, 2022. 1, 3, 4, 5

[23] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 3

[24] Andrey V. Savchenko. Emotieffnet facial features in uni-task emotion recognition in video at abaw-5 competition, 2023. 5

[25] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaanet: joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129:321–340, 2021. 2

[26] Gauthier Tallec, Arnaud Dapogny, and Kevin Bailly. Fighting noise and imbalance in action unit detection problems. *arXiv preprint arXiv:2303.02994*, 2023. 4

[27] 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, 2, 4

[28] Lingfeng Wang, Shisen Wang, and Jin Qi. Multi-modal multi-label facial action unit detection with transformer. *arXiv preprint arXiv:2203.13301*, 2022. 3

[29] Shangfei Wang, Yanan Chang, Yi Wu, Xiangyu Miao, Jiaqiang Wu, Zhouan Zhu, Jiahe Wang, and Yufei Xiao. Facial affective behavior analysis method for 5th abaw competition, 2023. 5

[30] 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, 2023. 5

[31] Jun Yu, Renda Li, Zhongpeng Cai, Gongpeng Zhao, Guochen Xie, Jichao Zhu, and Wangyuan Zhu. Local region perception and relationship learning combined with feature fusion for facial action unit detection, 2023. 5

[32] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2017. 1, 2

[33] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Facial affective analysis based on mae and multi-modal information for 5th abaw competition, 2023. 5

[34] Wei Zhang, Feng Qiu, Suzhen Wang, Hao Zeng, Zhimeng Zhang, Rudong An, Bowen Ma, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2428–2437, 2022. 3

[35] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 2

[36] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Continuous emotion recognition based on tcn and transformer, 2023. 5