An Attention-based Method for Multi-label Facial Action Unit Detection

Duy Le Hoai
Chonnam National University
Gwangju, South Korea
hoaiduy1396@gmail.com

Eunchae Lim
Chonnam National University
Gwangju, South Korea
enechae78@gmail.com

Eunbin Choi
Chonnam National University
Gwangju, South Korea
iidmsqlss@gmail.com

Sieun Kim
Chonnam National University
Gwangju, South Korea
leaza34@gmail.com

Sudarshan Pant
Chonnam National University
Gwangju, South Korea
sudarshan@chonnam.ac.kr

Guee-Sang Lee
Chonnam National University
Gwangju, South Korea
gslee@jnu.ac.kr

Soo-Huyng Kim
Chonnam National University
Gwangju, South Korea
shkim@jnu.ac.kr

Hyung-Jeong Yang*
Chonnam National University
Gwangju, South Korea
hjyang@jnu.ac.kr

Abstract

Facial Action Coding System is an approach for modeling the complexity of human emotional expression. Automatic action unit (AU) detection is a crucial research area in human-computer interaction. This paper describes our submission to the third Affective Behavior Analysis in-the-wild (ABAW) competition 2022. We proposed a method for detecting facial action units in the video. In the first stage, a lightweight CNN-based feature extractor is employed to extract the feature map from each video frame. Then, an attention module is applied to refine the attention map. The attention encoded vector is derived using a weighted sum of the feature map and the attention scores later. Finally, the sigmoid function is used at the output layer to make the prediction suitable for multi-label AUs detection. We achieved a macro F1 score of 0.48 on the validation set and 0.4206 on the test set compared to 0.39 and 0.3650 from the ABAW challenge baseline model.

1. Introduction

Facial Expression Recognition (FER) is an important task in artificial emotional intelligence (AEI) research since it not only helps to identify human affective states but also allows to mimic various emotions during human-machine communication. Therefore, facial expression recognition has been receiving increasing attention in recent years. Facial expression is one of the common ways to express emotions in humans. In literature, there are three common approaches to encode facial expressions. As early as the twentieth century, a system comprising six basic human emotions including anger, disgust, fear, happiness, sadness, and surprise has been proposed by Ekman and Friesen [43]. Several years later, Ekman developed the Facial Action Coding System (FACS) [11] including 46 action units that match expressions and human emotions. Action Units (AU), the points related to specific facial muscle actions are related to the facial expressions [12]. Apart from the above categorical system, valence-arousal space, a continuous model is also widely used to represent human emotions.

Action Unit detection has become an important facial analysis task that has been applied to various areas such as the healthcare, robotics, and entertainment industry by extracting and recognizing features representing human emotions [3]. However, due to the scarcity of large datasets, facial expression recognition, especially facial action unit detection remains a challenge [12]. The third Affective Behavior Analysis in-the-wild (ABAW) 2022 Competition [12-20] provides a benchmark and a massive dataset based on the large scale in-the-wild Aff-Wild2 database for four challenges including Valence-Arousal Estimation, Expression Classification, Action Unit Detection, and Multi-Task Learning.

In this paper, we introduce our approach for Action Unit Detection challenge in the ABAW 2022 competition. First, we enforce a lightweight feature extractor to extract visual information from image frames. Second, we leverage the attention mechanism to capture the important intra-region in the face image to improve the AU detection performance. In addition, to counter the imbalance data

* Corresponding Author.

problem, we execute class reweights binary cross-entropy loss. In the next sessions, we will explain our method in detail.

2. Related Works

2.1. Facial Emotion Recognition

Existing Facial Emotion Recognition approaches can be classified into hand-crafted feature-based, temporal feature-based, and deep learning-based approaches according to the extracted features. Hand-crafted features include texture-based features such as Gabor filter [21], SIFT [22], and HOG [23], geometry-based features, and hybrid features. Temporal feature-based methods use temporal features from the face appearance related to AUs. Li *et al.* [24] used the Long Short-Term Memory (LSTM) network to detect AUs. Deep learning-based methods have been studied for large-scale datasets and better accuracy of AUs detection. Wang *et al*. [25] proposed Self-Cure Network (SCN), a deep neural network that avoids the overfitting phenomenon from uncertain facial images.

2.2. Action Unit Detection Analysis

From the ABAW 2020 competition so far, many teams proposed single-task and multi-task learning methods to recognize Action Units (AUs). At the first ABAW competition in 2020, for the AUs detection challenge, the FLAB2020 team [39] proposed a pairwise deep architecture to counter a problem of label inconsistency and wins the first prize. The NISL2020 team [40] gets the second prize in the same track by merging the Denver Intensity of Spontaneous Facial Action (DISFA) dataset with the Aff-wild2 dataset to tackle data imbalance and using a teacher-student algorithm for learning from missing labels. The TNT team [41] achieves the third prize with a fully convolutional network for learning from preprocessed visual and audio inputs. At the second ABAW 2021, numerous works have tried on AUs detection task. The Netease Fuxi Virtual Human team [42] is the winner of the AUs challenge with a multi-task approach. They employ the pre-trained Deviation Learning Network (DLN) to produce fine-grained expression embedding and then fed the embedding into a hierarchical streaming network for multi-task recognition. The CPIC-DIR2021 team [5] leverages information from visual and audio modality, along with temporal features extracted by Transformer encoder for multi-task learning including facial action unit and expression recognition and ranks second place in both tasks. The Maybe Next Time team [4] earns third prize in both action unit and expression challenges by using pre-trained ResNet50 as a feature extractor for cropped and aligned input images and a dynamic learning strategy to train the model from incomplete labels dataset.

Aside from the works in the competition, here we briefly introduce single task learning methods for recognizing AUs. Pahl *et al.* [26] propose a method based on a multi-label class balancing algorithm for unbalanced datasets and a ResNet with 18 layers. Ji *et al.* [27] use an end-to-end deep neural network to carry out the static and dynamic features extraction and the AU classification. Zhang *et al.* [28] present a method based on JAA-Net and Graph Convolutional Network. Saito *et al.* [29] introduce a method of inserting siamese-based networks, known as uncertainty models, into automatic AUs recognition methods using a pairwise deep architecture [30] to further reduce AUs recognition errors.

3. Proposed Method

The overall of our proposal is illustrated in Figure 1. Our approach comprises two main components: a feature extractor for extracting visual features from the input images and an attention module to induct the model focusing on the important local regions for predicting action units. At the output layer, we leverage the sigmoid function to adapt with the multi-label classification task of the action unit detection challenge.

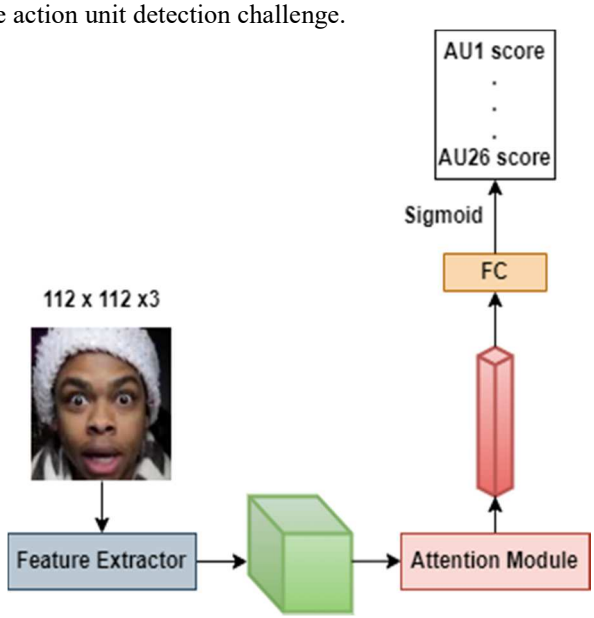

Figure 1. An overview of our proposed model.

3.1 Model Architecture

Feature Extractor. Previous approaches [4,5,6] make use of the pre-trained state-of-the-art models on the large-scale datasets as a feature extractor. In this work, we used a custom lightweight feature extractor. We used different pre-trained models on the ImageNet dataset and a custom CNN-based network as the feature extractor for training the model. From the experiment results, we explore that a

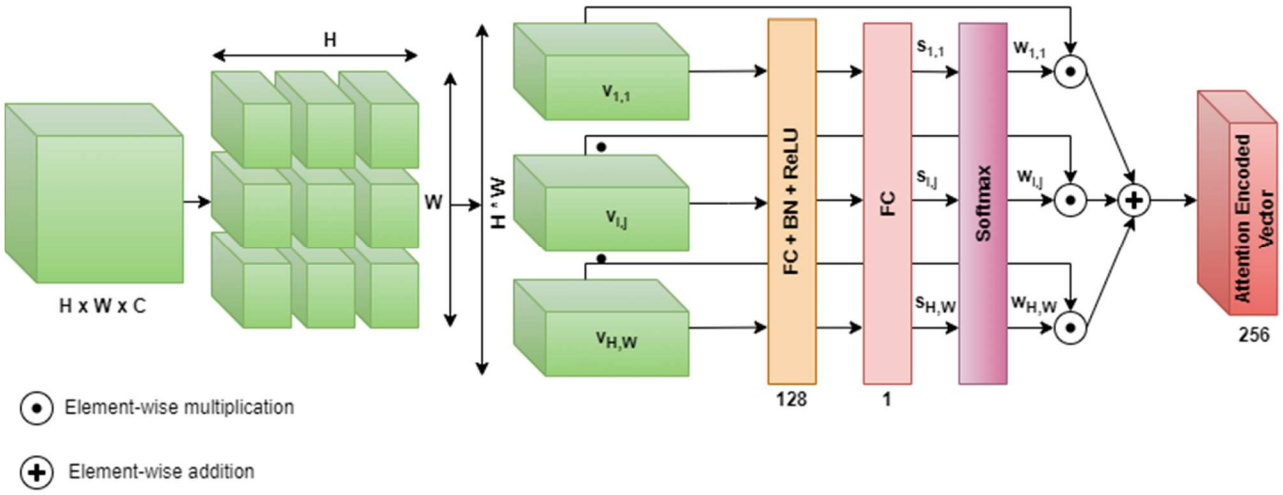

Figure 2. The details of the attention module that is based on the work [10].

custom CNN-based is the most effective feature extractor among them. In detail, we simply construct a convolutional block that contains a plain convolutional (Conv) layer followed by a Batch Normalization layer with ReLU as activation function and a max-pooling layer as the last layer of the block. We utilize a 3 x 3 kernel with a stride is 1 for all convolutional layers and strides of 2 x 2 for all max-pooling layers. The face image of size 112 x 112 x 3 is passed through a stack of 6 convolutional blocks with the number of filters for each block being 32, 64, 128, 128, 256, 256 sequentially. The extracted visual feature is a 256-dimensional feature map.

Attention Module. It is well known that the attention mechanism has been established a significant impact in many research fields such as natural language processing [7], computer vision [8], speech recognition [9], etc. In this work, we adopt the same Global-Local attention mechanism as described in [10] with a slight modification, Figure 2, to guide the network selectively focus on local salient parts to capture meaningful visual features for the action unit detection task. First, the feature map is split into a set of W*H sub-vectors, each vector $v_{i,j}$ has C elements corresponding to each location in the feature map. Next, a fully connected (FC) layer followed by a Batch Normalization and ReLU layer is added. Then a 1-unit FC layer is used to compute the score value $s_{i,j}$ for each sub-vector. After that, the weight $w_{i,j}$ of each sub-vector is calculated by a Softmax function. Finally, the attention encoded vector is obtained as a weighted sum of these sub-vector:

$$attention\ vector = \sum_{i=0}^{H}\sum_{j=0}^{W} w_{i,j} v_{i,j} \quad (1)$$

In the end, a feed-forward layer with a sigmoid activation function is adopted to produce the final action unit prediction.

3.2 Loss Function

Action Unit Detection challenge is a multi-label classification problem. Binary Cross-Entropy (BCE) is usually used for multi-label classification tasks with a sigmoid function as an activation function in the output layer. To solve the imbalanced class problem in the dataset, we introduce the sample weights in the BCE as follows:

$$\mathcal{L}(y, \bar{y}) = \frac{1}{C}\sum_{i=1}^{C} \mathcal{L}_{BCE}(y_i, \bar{y}_i) \quad (2)$$

$$\mathcal{L}_{BCE}(y_i, \bar{y}_i) = -[w_i y_i . log\bar{y}_i + (1 - y_i).\log(1 - \bar{y}_i)] \quad (3)$$

$$w_i = \frac{\#\ total\ training\ samples}{2 * (\#\ positive\ sample\ in\ i - th\ AU)} \quad (4)$$

where, y and $\bar{y}$ denote the ground truth and prediction, respectively. Variable C denotes the number of action units. In this work, C is equal to 12.

4. Experiments

4.1 Dataset and Metric

The proposed method was trained and validated on 547 provided videos that contain annotations in terms of 12 AUs. We directly used the face cropped and aligned images provided by the ABAW 2022 organizer to train and validate our model. These images have the size of 112x112x3 in RGB color space. We also removed all the labels annotated by the "-1". The performance metric in the competition is the macro F1 score across all 12 AUs:

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

Table 1. Ablation study of Action Unit detection results on Aff-Wild2 validation set.

Method	F1 score (%)
Baseline [12]	0.39
Ours with pre-trained ResNet50 [36] as feature extractor	0.3157
Ours with pre-trained InceptionV3 [37] as feature extractor	0.2383
Ours with pre-trained EfficientNet [38] as feature extractor	0.3184
Ours w/o attention module	0.4535
Ours w/o re-weighted BCE loss	0.4323
Ours (Custom CNN network feature extractor + Attention module + re-weighted BCE loss)	**0.4803**

Table 2. Our final AU detection result compared to other competitive works on the official testing set. Our result is indicated in bold.

Method	F1 score (%)
Baseline [12]	36.50
Netease Fuxi Virtual Human [31]	49.89
SituTech [32]	49.82
PRL [33]	49.04
STAR-2022 [34]	48.83
ISIR_DL [35]	44.32
Ours	**42.06**

4.2 Experiment Setting

Our model is implemented using TensorFlow [1] with an NVIDIA RTX 3080 graphics card. We used Adam [2] optimizer with the initial learning rate of 0.001 which is set to 0.0001 after 5 epochs, with the hyper-parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The mini-batch size is set to 256.

4.3 Results

Our experiment result on the validation set of the action unit detection task is shown in Table 1. We use the cropped and aligned images when doing validation. We conduct experiments with various feature extractors including pre-trained ResNet50, InceptionV3, EfficientNet, and a custom CNN-based network to select the most appropriate backbone for our model. In Table 1, it can be observed that our model with a custom CNN feature extractor followed by the attention module trained by re-weighted binary cross-entropy loss provides the highest performance. After 3 epochs, we achieve an F1 score of 0.48 compared to 0.39 of the baseline result on the validation set. To analyze the effects of the attention module and re-weighted BCE loss we also conducted ablation studies. The usage of re-weighted BCE loss leads to significantly improved model performance. The attention module helps to increase the model accuracy but has less impact than the re-weighted BCE loss.

Table 2 presents the comparison results between our method and the other teams on the official testing set. We first summarize the methods from all teams that scored higher than baseline for further comparison. The Netease Fuxi Virtual Human team [31] who is the winner of the AU detection challenge proposed a Transformer-based network to fuse the features from three different modalities including visual, textual, and acoustic. The SituTech [32] team (rank #2) leveraged external datasets in the training models and multi-models ensembling approach to enhance the performance. The PRL [33] team (rank #3) introduced temporal information from the video's frame sequence into the network using GRU combined with Local Attention. The STAR-2022 team [34] designed a network based on Transformer and Convolution for learning from visual and audio information. The ISIR_DL team [35] used Transformer for self-attention on the input images and cross-attention on label tokens. As illustrated in Table 2, we achieve an F1 score of 0.4206 on the test set. It indicates that our approach scored significantly better than the baseline. Moreover, our proposal is simple but competitively effective compared to other works in this challenge. In detail, we only used unimodal (visual) versus multi-modal (visual, textual, acoustic) from the winner team. Besides, we solely trained our model using provided dataset from the competition's organizer rather than using the external dataset like the second prize team. Furthermore, we extracted spatial-only information while spatial-temporal information was extracted in the top-3 team.

5. Conclusion

In this work, we presented a deep learning-based approach using attention mechanism and class re-weight binary cross-entropy loss for the Action Unit Detection challenge of the ABAW2022 competition. The experiment result shows that the attention mechanism is an effective technique for AU detection task and class re-weight can tackle the imbalance dataset problem in the multi-label classification task. Our proposed method greatly increases the performance compared to the baseline model on both validation and testing sets.

References

[1] Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado et al.

"Tensorflow: Large-scale machine learning on heterogeneous distributed systems." arXiv preprint arXiv:1603.04467 (2016).

[2] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[3] Shao, Zhiwen, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. "Facial action unit detection using attention and relation learning." IEEE transactions on affective computing (2019).

[4] Thinh, Phan Tran Dac, Hoang Manh Hung, Hyung-Jeong Yang, Soo-Hyung Kim, and Guee-Sang Lee. "Emotion Recognition With Sequential Multi-task Learning Technique." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3593-3596. 2021.

[5] Jin, Yue, Tianqing Zheng, Chao Gao, and Guoqiang Xu. "MTMSN: Multi-Task and Multi-Modal Sequence Network for Facial Action Unit and Expression Recognition." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3597-3602. 2021.

[6] Jacob, Geethu Miriam, and Bjorn Stenger. "Facial action unit detection with transformers." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[7] Galassi, A., Lippi, M., Torroni, P.: Attention in natural languageprocessing. IEEE Transactions on Neural Networks and LearningSystems (2020)

[8] Wang, F., Tax, D.M.J.: Survey on the attention based RNN model and its applications in computer vision. CoRRabs/1601.06823(2016). URL http://arxiv.org/abs/1601.0682358.

[9] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-based models for speech recognition.In: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama,R. Garnett (eds.) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pp. 577–585 (2015). URL http://papers.nips.cc/paper/5847-attention-based-models-for-speech-recognition

[10] Le, Nhat, Khanh Nguyen, Anh Nguyen, and Bac Le. "Global-local attention for emotion recognition." Neural Computing and Applications (2021): 1-15.

[11] Paul Ekman and Wallace V Friesen. Manual for the facial action coding system. *Consulting Psychologists Press*, 1978

[12] Kollias, Dimitrios. "ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Multi-Task Learning Challenges." arXiv preprint arXiv:2202.10659 (2022).

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

[14] D. Kollias, A. Schulc, E. Hajiyev and S. Zafeiriou, "Analysing Affective Behavior in the First ABAW 2020 Competition," 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 637-643, doi: 10.1109/FG47880.2020.00126.

[15] Dimitrios Kollias, Viktoriia Sharmanska, and StefanosZafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. arXiv preprint arXiv:2105.03790, 2021.

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

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

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

[19] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nico-laou, Athanasios Papaioannou, Guoying Zhao, Bj¨ornSchuller, 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, 127(6):907–929, 2019.

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

[21] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 11(4):467, 2002.

[22] P. C. Ng and S. Henikoff. Sift: predicting amino acid changes that affect protein function. Nucleic Acids Research, 31(13):3812–3814, 2003.

[23] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In IEEE Computer Society Conference on Computer Vision & Pattern Recognition, 2005.

[24] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6766–6775, 2017.

[25] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 6896–6905. IEEE, 2020.

[26] Jaspar Pahl, Ines Rieger, and Dominik Seuss. Multi-label class balancing algorithm for action unit detection. arXiv preprint arXiv:2002.03238, 2020.

[27] Xianpeng Ji, Yu Ding, Lincheng Li, Yu Chen, and Changjie Fan. Multi-label relation modeling in facial action units detection. arXiv preprint arXiv:2002.01105, 2020.

[28] Chenggong Zhang, Juan Song, Qingyang Zhang, Weilong Dong, Ruomeng Ding, and Zhilei Liu. Action unit detection with joint adaptive attention and graph relation. arXiv preprint arXiv:2107.04389, 2021.

[29] Junya Saito, Ryosuke Kawamura, Akiyoshi Uchida, Sachihiro Youoku, Yuushi Toyoda, Takahisa Yamamoto, Xiaoyu Mi, and Kentaro Murase. Action units recognition by pairwise deep architecture. arXiv preprint arXiv:2010.00288v2, 2020.

[30] Junya Saito, Xiaoyu Mi, Akiyoshi Uchida, Sachihiro Youoku, Takahisa Yamamoto, and Kentaro Murase. Action units recognition using improved pairwise deep architecture. arXiv preprint arXiv:2107.03143, 2021.
[31] Zhang, Wei, Zhimeng Zhang, Feng Qiu, Suzhen Wang, Bowen Ma, Hao Zeng, Rudong An, and Yu Ding. "Transformer-based Multimodal Information Fusion for Facial Expression Analysis." arXiv preprint arXiv:2203.12367 (2022).
[32] Jiang, Wenqiang, Yannan Wu, Fengsheng Qiao, Liyu Meng, Yuanyuan Deng, and Chuanhe Liu. "Facial Action Unit Recognition With Multi-models Ensembling." arXiv preprint arXiv:2203.13046 (2022).
[33] Nguyen, Hong-Hai, Van-Thong Huynh, and Soo-Hyung Kim. "An Ensemble Approach for Facial Expression Analysis in Video." arXiv preprint arXiv:2203.12891 (2022).
[34] Wang, Lingfeng, Shisen Wang, and Jin Qi. "Multi-modal Multi-label Facial Action Unit Detection with Transformer." arXiv preprint arXiv:2203.13301 (2022).
[35] Tallec, Gauthier, Edouard Yvinec, Arnaud Dapogny, and Kevin Bailly. "Multi-label Transformer for Action Unit Detection." arXiv preprint arXiv:2203.12531 (2022).
[36] Kaiming He et al. “Deep residual learning for image recognition”. In: CVPR. 2016.
[37] Christian Szegedy et al. “Rethinking the inception ar-chitecture for computer vision”. In:CVPR. 2016.
[38] Mingxing Tan and Quoc Le. “EfficientNet: Rethink-ing Model Scaling for Convolutional Neural Net-works”. In: ICML. 2019.
[39] Saito, Junya, Ryosuke Kawamura, Akiyoshi Uchida, Sachihiro Youoku, Yuushi Toyoda, Takahisa Yamamoto, Xiaoyu Mi, and Kentaro Murase. "Action units recognition by pairwise deep architecture." arXiv preprint arXiv:2010.00288 (2020).
[40] Deng, Didan, Zhaokang Chen, and Bertram E. Shi. "Multitask emotion recognition with incomplete labels." In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 592-599. IEEE, 2020.
[41] Kuhnke, Felix, Lars Rumberg, and Jörn Ostermann. "Two-stream aural-visual affect analysis in the wild." In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 600-605. IEEE, 2020.
[42] Zhang, Wei, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, Yu Ding, Runze Wu, Tangjie Lv, and Changjie Fan. "Prior Aided Streaming Network for Multi-task Affective Analysis." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3539-3549. 2021.
[43] P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion.”Journal of personality and social psychology, vol. 17, no. 2, pp. 124–129, 1971.