# Classification of Facial Expression In-the-Wild based on Ensemble of Multi-head Cross Attention Networks

Jae-Yeop Jeong[1], Yeong-Gi Hong[1], Daun Kim, and Jin-Woo Jeong*
Department of Data Science, Seoul National University of Science and Technology
Gongreung-ro 232, Nowon-gu, Seoul, South Korea
{jaey.jeong, yghong, daun, jinw.jeong}@seoultech.ac.kr

Yuchul Jung and Sang-Ho Kim
Kumoh National Institute of Technology
Daehak-ro 61, Gumi, South Korea
{jyc, kimsh}@kumoh.ac.kr

## Abstract

*How to build a system for robust classification and recognition of facial expressions has been one of the most important research issues for successful interactive computing applications. However, previous datasets and studies mainly focused on facial expression recognition in a controlled/lab setting, therefore, could hardly be generalized in a more practical and real-life environment. The Affective Behavior Analysis in-the-wild (ABAW) 2022 competition released a dataset consisting of various video clips of facial expressions in-the-wild. In this paper, we propose a method based on the ensemble of multi-head cross attention networks to address the facial expression classification task introduced in the ABAW 2022 competition. We built a uni-task approach for this task, achieving the average F1-score of 34.60 on the validation set and 33.77 on the test set, ranking second place on the final leaderboard.*

## 1. Introduction

Recognition of facial expression is becoming more important for various interactive computing domains, such as human-computer/machine interaction, human-robot interaction, and human-AI interaction. Understanding the user's affective states is essential for intelligent agents to provide appropriate services to the users. However, previous studies on facial expression mainly utilized a set of human faces captured in a controlled setting, resulting in various limitations of the application in-the-wild. Recently, several works focusing on the affective behavior analysis in-the-wild have been introduced to realize the generation of trust, understanding, and closeness between humans and machines in real-life environments. [10].

The 3rd competition on Affective Behavior Analysis in-the-wild (ABAW), held in conjunction with the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2022, is a place where researchers can present their own contributions on the automatic analysis of human behavior and emotion recognition which is robust to video recording conditions, diversity of contexts, and timing of display [10]. The 3rd ABAW competition is based on the Aff-Wild2 database [11–17] which is an extension of the Aff-wild database [27] and consists of the following four tracks: 1) Valence-Arousal (VA) estimation, 2) Expression classification, 3) Action Unit (AU) detection, and 4) Multi-Task-Learning (MTL). For the VA estimation challenge, totally 564 videos of around 2.8M frames that contain annotations in terms of valence and arousal (values ranged continuously in [-1, +1]) are used. Here, arousal is the level of autonomic activation which ranges from calm to excited while Valence is the level of pleasantness defined along a continuum from negative to positive [4]. Similarly, for the expression classification challenge, totally 548 videos of around 2.7M frames that contain annotations in terms of the 6 basic expressions (i.e., Anger, Disgust, Fear, Happy, Sad, Surprised), plus the neutral state, plus a category 'other' that denotes expressions/affective states other than the 6 basic ones are used [10]. The AU detection task provides 547 videos of around 2.7M frames containing annotations in terms of 12 action units based on the facial action coding system (FACS) are used [2]. Finally, the multi-task learning task provides a set of videos annotated with both all of these expression labels [10].

[1]Both authors contributed equally to this research.
*Corresponding author

In this paper, we propose a method based on the multi-head cross attention networks proposed by [25] to solve the the 8-class facial expression classification challenge. To handle various challenging issues introduced in this task, such as a class imbalance problem and the lack of visually diverse images for certain emotional classes, we extended our training dataset with external databases and data augmentation techniques, applied the focal loss algorithm [18], and implemented an ensemble-based prediction approach.

## 2. Method

To generate a generalizable and robust deep learning model for facial expression classification in-the-wild, we first focused on analyzing characteristics and distributions of the training dataset provided by the competition organizers and then adopting more advanced CNN architectures and various learning techniques to improve the performance.

### 2.1. Data Pre-processing

To develop successful deep learning applications in terms of model performance, obtaining a large number of data with diversity is essential. However, as shown in Table 1, the Aff-wild2 dataset has a class imbalance problem, resulting in some emotional categories having far fewer images than others. For example, the number of images with "Neutral" class is 18x larger than that of images with "Fear" class. To address this issue, we used the following two strategies: 1) adding external databases and 2) applying various data augmentation techniques. First, we downloaded and processed external facial expression databases, such as AffectNet [20], ExpW [29], and Ai-Hub dataset [1]. The images included in these databases were generally captured and recorded under in-the-wild settings. The sample images from each dataset can be found from Figure 1. As shown in the figure, Aff-wild2 and AffectNet datasets share the same facial expression categories while ExpW and Ai-Hub datasets provide only part of expression categories included in the Aff-wild2 dataset. Note that Ai-hub dataset [1] is comprised of facial expression images taken by Korean actors in-the-wild. Among various images included in the Ai-hub dataset, we only used a set of images with "Neutral", "Anger", "Fear", and "Surprise" expressions. We believe that this extension could not only add more diversity in terms of visual appearances but also mitigate the imbalance problems between the classes except for the "Disgust" case. Second, we employed various data augmentation techniques (i.e., color jitter, random crop, horizontal flip, color jitter with random crop, random crop with flip) to prevent over-fitting. Finally, we cropped the face region of each image using DeepFace face detector algorithm [23,24] and then resized every patch into 224 x 224 scale. Table 1 shows the statistics of raw dataset we used when training the model and Figure 2 depicts the difference between the data distribution of the original training dataset and our extended dataset.

### 2.2. Model Architecture

The overall architecture of our method is illustrated in Figure 3. Our method is based on the approach called "DAN" [25] which consists of the following two modules: Feature Clustering Network (FCN) and attention phases: Multi-head cross Attention Network (MAN) and Attention Fusion Network (AFN). Specifically, the FCN module extracts the intermediate visual features from a set of input images in a class discriminative manner to maximize the inter-class margin and minimize the intra-class margin [25]. To secure class discrimination, they introduced a new loss function called affinity loss which makes it possible for the features from the same class to move closer to the median of the class while getting further away from other classes. The affinity loss is defined as:

$$L_{af} = \frac{\sum_{i=1}^{M} \left\| x_i^{'} - c_{yi} \right\|_2^2}{\sigma_c^2} \tag{1}$$

, where *i*-th input vector is $x_i \in X$ (input feature space), and target is $y_i \in Y$ (target space), $c$ denotes a class center ($c \in R^{m \times d}$, where $M$ is the dimension of $Y$ and $d$ is the dimension of class centers) which is randomly sampled from Gaussian distribution, and $\sigma_c$ indicates the standard deviation among class centers. For feature extraction in the FCN module, we utilized a ResNet-50 network [8] pretrained on VGGFace2 dataset [5] as a backbone.

The MAN module is composed of parallel cross-head attention units which are combinations of spatial and channel attention units. The MAN module takes the features from the FCN module as input and outputs a set of attention maps by implicitly combining the features extracted from the two above dimensions (i.e., spatial and channel-wise ones). Finally, the AFN module attempts to merge these attention maps (i.e. outputs from the MAN module) in an orchestrated fashion. The cross attention heads are supervised to center on different critical regions and avoid overlapping attentions using the partition loss, which is designed to maximize the variance among the attention maps, defined as:

$$L_{pt} = \frac{1}{NC} \sum_{i=1}^{N} \sum_{j=1}^{C} log(1 + \frac{k}{\sigma_{ij}^2}) \tag{2}$$

, where $k$ is the number of cross attention, $N$ is the number of samples, $C$ is the channel size of the attention maps, and $\sigma_{ij}^2$ depicts the variance of the *j*-th channel on the *i*-th sample. The merged attention map is then used for computing the class confidence with a linear layer.


Figure 1. Examples of training samples in each dataset

Table 1. Data statistics

| Database | Neutral | Anger | Disgust | Fear | Happy | Sad | Suprise | Other |
|---|---|---|---|---|---|---|---|---|
| Aff-Wild2 | 177,498 | 16,573 | 10,810 | 9,080 | 95,633 | 79,862 | 31,637 | 165,866 |
| AffectNet | 74,874 | 24,882 | 3,803 | 6,378 | 134,415 | 25,459 | 14,090 | 3,750 |
| ExpW | 34,883 | 3,671 | 3,395 | 1,088 | 30,537 | 10,559 | 7,060 | - |
| AI-Hub | 43,299 | 59,696 | - | 59,262 | - | - | 59,643 | - |
| Total | 330,554 | 104,822 | 18,008 | 75,808 | 260,585 | 115,880 | 112,430 | 169,616 |

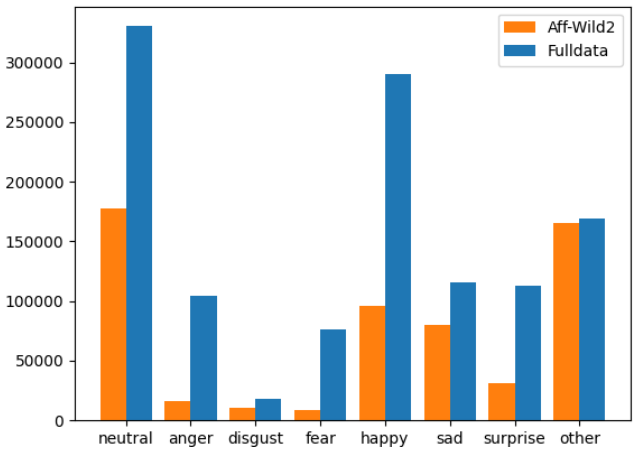


Figure 2. Training data statistics

The final loss function is therefore defined as:

$$L = L_{af} + L_{pt} + L_{fo} \tag{3}$$

, where $L_{fo}$ denotes the focal loss proposed by [18] that is known to produce a more robust performance in case a class imbalance problem occurs. For more details about DAN architecture, please refer [25].

### 2.3. Ensemble approach

Generally, it is widely known that an ensemble of multiple weak models even show better performances than a single strong model [6]. Therefore, an ensemble approach typically consists of a set of individual models that predict their own labels for a given sample. Among several ensemble approaches available, in this study, we employed a bagging approach [3] where each classifier is trained with a subset of training data, as illustrated in the bottom of Figure 3. By training with the sub-sampled training data, each individual model can observe different aspects of training samples, thereby learning different representations/features. Finally, we applied a soft voting (i.e., probability-based voting) method to integrate the predictions from a set of trained models, thereby resulting in the final label of a test sample.

## 3. Experiments and Results

In this section, we report the performance of our framework on the both official validation set and testing set prepared for the ABAW 2022 competition. The evaluation metric for the expression classification task is defined as :

$$P_{EXPR} = \frac{\sum F1_{class}}{8} \tag{4}$$

, where $F1_{class}$ denotes F1 score per class.

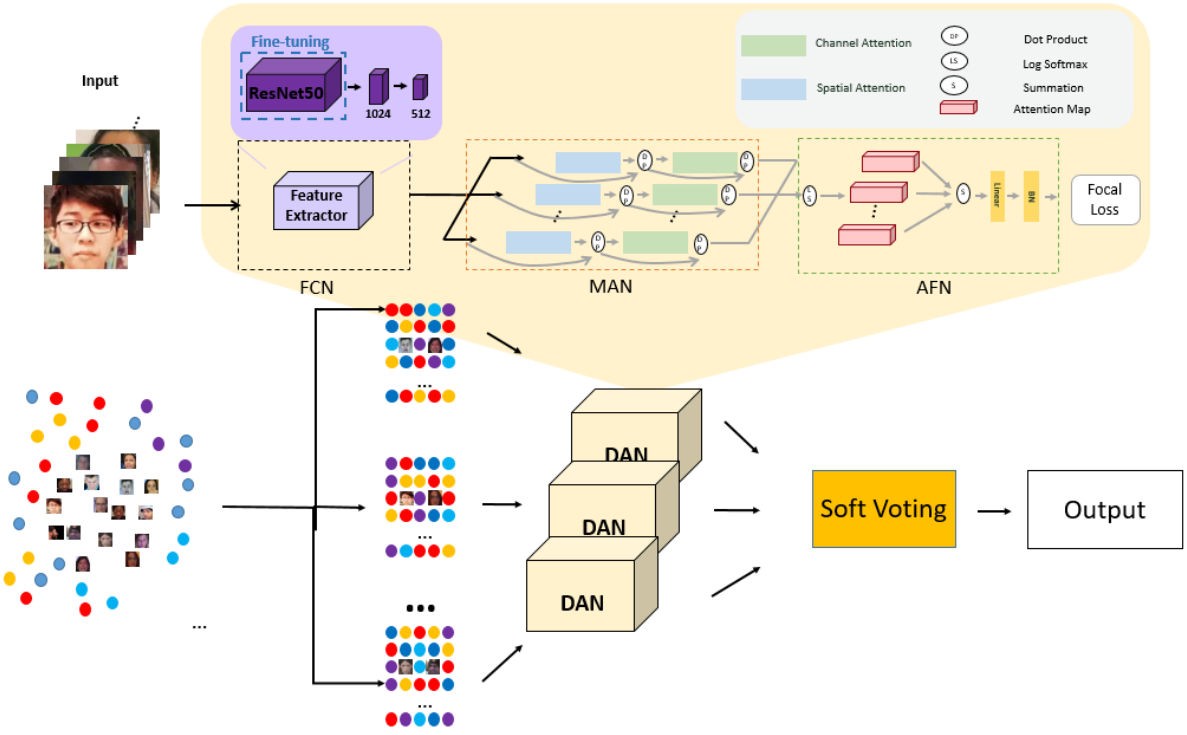


Figure 3. Overview of the proposed architecture

### 3.1. Training setup

Our framework was trained with a batch size of 1,024 and the ADAM optimizer with a learning rate of 0.0001. To avoid overfitting and achieve a stable performance, we adopted the following regularization and learning techniques: weight decay with a value of 0.0001, gradient clipping with a threshold value of 1.0, and the Exponential decay learning rate scheduler. Also, the number of cross attention heads was set to 4. The summary of the hyper-parameters used in this study is shown in Table 2.

As depicted in Table 1, the training dataset provided by the ABAW 2022 competition shows an unbalanced data distribution, so that some emotional categories have far fewer images than others. Therefore, we also applied a weighted data sampling method in which the sampling frequency of data for each class in each batch is adjusted according to the different weights given to the class when loading images during the training process of ensemble models. In other words, by giving a higher weigh to the class with a smaller proportion of data, we draw more samples of that class during the sampling process.

All the experiments were conducted using a GPU server equipped with six NVIDIA RTX 3090 GPUs, 128 GB RAM, and an Intel i9-10940X CPU. We used Pytorch framework for the implementation/modification, training and evaluation of the model.

Table 2. Hyperparameter setting

| Hyper-parameter | Value |
|---|---|
| Batch Size | 1024 |
| Optimizer | ADAM |
| Learning Rate | $1e^{-4}$ |
| Learning Rate Scheduler | Exponential Decay |
| Epochs | 8 |
| Optimizer Weight Decay | $1e^{-4}$ |
| Number of Cross Attention Head | 4 |

### 3.2. Model configurations

Based on the overall architecture discussed in Section 2.2, we built various models with four different configurations and evaluated their initial performances using the Aff-wild2 validation set. The model configuration dimensions we defined in this work are 1) data sampling ratio (i.e., training with sub-sampled data or full dataset?) and 2) training strategy (i.e., training from scratch or fine-tuning with pretrained weights?). In addition to these dimensions, we also applied an ensemble approach for weak models (i.e., models trained with sub-sampled data). As a result, we could obtain the following models to be validated: 1) a single model fine-tuned with full data (DAN TL), 2) a single model trained from scratch with full data (DAN Scratch), 3) ensemble of models each of which fine-tuned with sub-

sampled data (DAN Weak Ensemble TL), and 4) ensemble of models each of which trained from scratch with sub-sampled data (DAN Weak Ensemble Scratch). Finally, we also built an ensemble model which integrates all the configurations described above (DAN Ensemble).

### 3.3. Performance evaluation

First, the performance of our various models was compared with those of a competition baseline (VGG16 network pretrained on the VGGFace dataset) and the DAN baseline (a model pretrained on the AffectNet-8 dataset with a ResNet18 backbone pretrained on the MSCeleb-1M dataset [7]). Table 3 summarizes the average F1 score of each model on the official validation set. The baseline of this year's competition achieved 23% while the DAN baseline [25] scored 22.60%. The average F1 scores from our DAN TL and DAN Scratch methods were 31.7% and 33.3%, respectively, which outperform the baselines. These imply that the extended training dataset and several modifications to the original DAN architecture could contribute to the performance improvement. On the other hand, the ensemble-based methods (i.e., DAN Weak Ensemble TL, DAN Weak Ensemble Scratch, and DAN Ensemble) could also achieve comparable performances, resulting in the average F1-score of 33.23%, 33.47%, and 34.60%, respectively. During the training step, every single weak model was trained with a sub-sample dataset, learning different views of underlying feature representations, thereby increasing the base model diversity which is essential for a robust ensemble architecture. It is also worth noting that DAN Ensemble method (Ensemble using all the DAN variants) performed the best (34.60%), implying that a set of big DAN models and a set of weak ensembles could complement each other, which leads to better performances.

Based on the validation result, we chose the DAN Ensemble method as our final model and the prediction results from this model on the official testset were submitted for the final evaluation. Table 4 shows the leader-board for the expression classification task of the ABAW 2022 competition. The leader-board only includes a list of valid submissions that achieved better performance than the baseline. As shown in the table, the baseline method produced the average F1 score of 20.50 which is similar to that on the validation set. Our approach ranked second with an F1 score of 33.77%, following the method of the Netease Fuxi Virtual Human group with an F1 score of 35.87%. From this result, we could observe that our approach (DAN Ensemble) attempted to reduce generalization errors using multi-faceted features from a set of ensembles, yielding a stable performance for even unseen data. As a result, this result shows the feasibility of the proposed approach for the classification of facial expressions in-the-wild.

Table 3. Average F1 scores on the validation set ("TL" denotes "Transfer Learning", "Scratch" denotes training from scratch, "Weak Ensemble" denotes aggregation of the weak models, and "Ensemble" denotes the result from the aggregation of all models we trained).

| Method | F1(%) |
|---|---|
| baseline | 23.00 |
| DAN (ResNet18) pretrained on AffectNet-8 | 22.60 |
| DAN (ResNet50) TL | 31.70 |
| DAN (ResNet50) Scratch | 33.30 |
| DAN (ResNet50) Weak Ensemble TL | 33.23 |
| DAN (ResNet50) Weak Ensemble Scratch | 33.47 |
| **DAN (ResNet50) Ensemble** | **34.60** |

Table 4. Average F1 scores on the test set

| Method | F1(%) |
|---|---|
| baseline | 20.50 |
| USTC-NELSLIP | 21.91 |
| dgu [9] | 27.2 |
| PRL [21] | 28.6 |
| HSE-NN [22] | 30.25 |
| AlphaAff [26] | 32.17 |
| Netease Fuxi Virtual Human [28] | 35.87 |
| **Ours (DAN Ensemble)** | **33.77** |

## 4. Conclusion

In this paper, we proposed an ensemble approach of multi-head cross attention networks to address the facial expression challenge introduced in ABAW 2022. In addition, we also collected and pre-processed various facial expression-related datasets so that our networks learn more robust feature representations from the images of diverse appearances. Finally, our approach produced promising results with the average F1 score of 34.60 on the validation set and 33.77 on the final test set, which positioned second place on the leader-board. In our future work, we plan to extend our approach to handle the VA estimation task, AU detection task, and multi-task-learning task. Our current approach, however, still suffers from the class imbalance problem, resulting in much lower performance in some emotional classes like "Disgust". Therefore, we will attempt to exploit various generative approaches for artificial facial expression synthesis/generation to tackle this issue. Finally, we will study a method to exploit the temporal relationship between the video frames and utilize multi-modal features during the multi-task learning process to achieve more improved performances.

## 5. Acknowledgement

This work was supported by the NRF grant funded by the Korea government (MSIT) (No.2021R1F1A1059665), by the Basic Research Program through the NRF grant funded by the Korea Government (MSIT) (No.2020R1A4A1017775), and by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0017123, The Competency Development Program for Industry Specialist).

## References

[1] Ai-hub dataset, https://aihub.or.kr/aidata/27716. 2

[2] *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* Series in affective science. Oxford University Press, New York, NY, US, 1997. 1

[3] Leo Bbeiman. Bagging predictors, 1996. 3

[4] Patricia E. G. Bestelmeyer, Sonja A. Kotz, and Pascal Belin. Effects of emotional valence and arousal on the voice perception network. *Social Cognitive and Affective Neuroscience*, 12(8):1351–1358, 04 2017. 1

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. pages 67–74. Institute of Electrical and Electronics Engineers Inc., 6 2018. 2

[6] Yue Cao, Thomas Andrew Geddes, Jean Yee Hwa Yang, and Pengyi Yang. Ensemble deep learning in bioinformatics. *Nature Machine Intelligence*, 2(9):500–508, Sep 2020. 3

[7] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV 2016*, August 2016. 5

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2

[9] Jun-Hwa Kim, Namho Kim, and Chee Sun Won. Facial expression recognition with swin transformer, 2022. 5

[10] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022. 1

[11] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 794–800. 1

[12] 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

[13] 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

[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. 1

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

[16] 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

[17] 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

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 3

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2017.

[20] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(01):18–31, jan 2019. 2

[21] Kim Ngan Phan, Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. Expression classification using concatenation of deep neural network for the 3rd abaw3 competition, 2022. 5

[22] Andrey V. Savchenko. Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices, 2022. 5

[23] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5, 2020. 2

[24] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4, 2021. 2

[25] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. 9 2021. 2, 3, 5

[26] Fanglei Xue, Zichang Tan, Yu Zhu, Zhongsong Ma, and Guodong Guo. Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition, 2022. 5

[27] 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

[28] Wei Zhang, Zhimeng Zhang, Feng Qiu, Suzhen Wang, Bowen Ma, Hao Zeng, Rudong An, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis, 2022. 5

[29] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning social relation traits from face images, 2015. 2