Multi-modal Information Fusion for Action Unit Detection in the Wild

Yuanyuan Deng[1], Xiaolong Liu[1], Liyu Meng[1], Wenqiang Jiang[1], Youqiang Dong[2], Chuanhe Liu[1†]
[1] Beijing Seek Truth Data Technology Co.,Ltd.
[2]School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 100044, China.

Abstract

Action Unit (AU) detection is an important research branch in affective computing, which better understands human emotional intentions and responds more naturally to their needs and desires. In this paper, we present our latest progress techniques in the 5th Affective Behavior Analysis in-the-wild (ABAW) competition, including data balancing by marking, extracting features visual through models trained in face database and audio through deep networks and traditional methods, proposing model structures for mapping multimodal information to a unify multimodal vector space and fusing results from multiple models. These methods are effective on the official validation dataset of the Aff-Wild2. The final F1 in the 5th ABAW competition test dataset achieves 54.22%, 4.33% higher than the best results in the 3rd ABAW competition.

1. Introduction

The explosion of LLM [1] in January 2023 has brought more attention to artificial intelligence and motivate countless researchers. As an essential component of artificial intelligence and human-computer interaction, affective computing has made significant progress in recent years with the deepening of psychological research and the rapid development of deep learning. However, there are still many technologies that need to research. Action Unit (AU) detection, as a technique in emotion computing, helps to understand human emotional needs and intentions and plays an essential role in human-computer interaction, healthcare, marketing, and user research.

The 5th ABAW Competition is a continuation of the Competitions held at ECCV 2022, IEEE CVPR 2022, ICCV 2021, IEEE FG 2020 and CVPR 2017 Conferences, and is dedicated at automatically analyzing affect [2]. It motivates researchers worldwide to implement their latest techniques on the Aff-Wild2 [2–13] database, which multiple experts annotate. It provides us with rich and reliable data resources and makes our experimental results more convincing.We are studying the methods used by ABAW competition winners over the past few years and finding that multimodal fusion performs well. This is inspiring us to explore the multimodal fusion method more deeply.

†Corresponding Author.

2. Related Works

Multimodal training has the significant advantage of leveraging other modalities to improve model performance. However, multimodal training also increases the number of input parameters and the demands on GPUs. We improved the model training method to achieve high-precision output with a low-configuration GPU.

Our method involves two steps to achieve model training. First, we use face-related pre-training models to extract visual and audio features. Then, we combine these features as input to train the model. In this way, we can achieve model training with longer sequences while reducing GPU memory usage to some extent.

Section 2 describes our multimodal information fusion model method for the AU task in the ABAW5 competition. First, to balance the distribution of labels in the training data, we cover the image's upper or lower by face landmarks detectionn [14] on the official dataset and extend them to the training set. Then, we use different depth networks to extract visual and audio features. Last, we propose the model structure of CrossAttention [15] + Transformer [16], Dual Transformer, and TCN [17]+Transformer to train the multimodal features. Section 3 demonstrates the effectiveness of these methods by conducting comparative experiments in the official validation set.

3. Method

A video consists of two components: visual and audio information. Typically, image frames are used to process visual information, while audio signals are converted into digital representations for model training. In our pipeline, we derive visual information from n-dimensional image features extracted from a pre-trained face model, and process audio information using a combination of deep neural net-

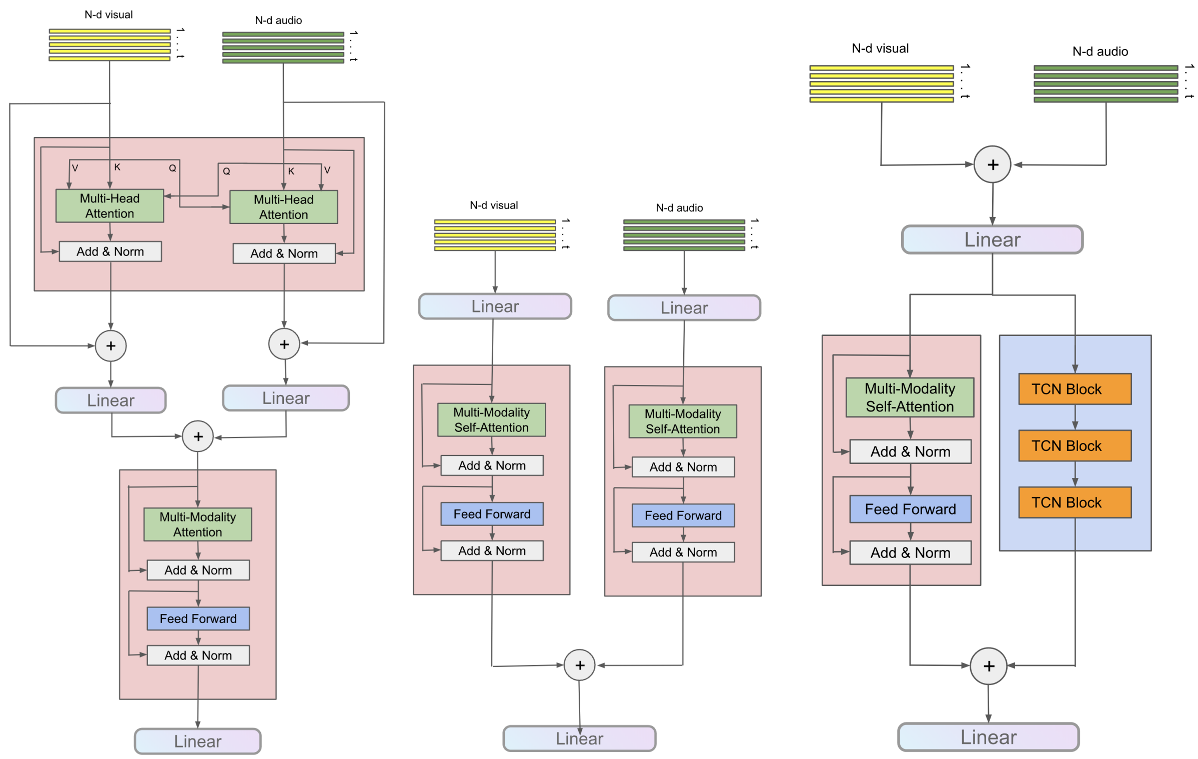

Figure 1. model structure
Figure 1 shows the three model constructs we use in the 5th ABAW competition. The input to the model is a combination of visual and audio features extracted from the Aff-Wild2 dataset. Different feature combinations have different feature dimensions, so N-d is used to represent the input feature dimensions of the model.

works and traditional methods to extract n-dimensional audio features. Our model uses both types of features as inputs for the visual and audio parts, respectively, with a uniform sequence length for the AU task during training. If a video's length is shorter than the sequence length, we replicate the last frame. Additionally, we handle frames without faces by using the nearest valid frame to represent them.

Training features require less time and GPU memory than directly training multi-modal image-audio data. Therefore, we train AU features using various model structures, including single-model Transformer, TCN, GRU [18], BIGRU [19], LSTM [20], BiLSTM [21] and our proposed hybrid models. The three hybrid model structures shown in Figure 1 are the top-performing models in our experiments during the 5th ABAW competition.

3.1. Data balancing

The AU dataset poses a challenge due to its imbalanced distribution of data for multi-label classification. Imbalanced data can hinder models from learning better representations. To address this issue, we propose a novel masking method that covers the upper and lower face of images with a few labels using face landmark detection. These masked images and labels are then extended to the training set. The facial action unit labels covered by the black borders are set to 0, while those not covered retain their original labels. This method is applied only to AUs with below-average data volume (AU1, AU2, AU4, AU5, AU12, AU15, AU23, AU24, AU26) to mitigate data imbalance. Conversely, AUs with above-average data volume (AU6, AU7, AU10, AU25) do not require additional masked data.

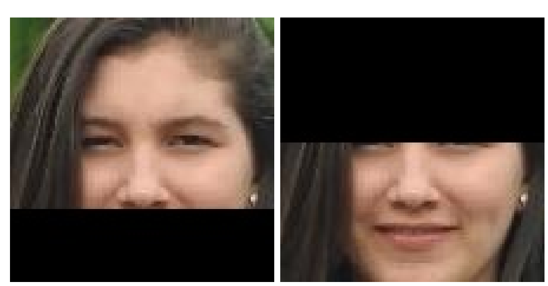

Figure 2. Face with mask generated by face landmarks detection.

3.2. Feature extraction

3.2.1 Visual Feature

Different networks extract features that can have diverse effects on model training. In the 5th ABAW competition, we used both public and private datasets to pre-train various models, from which we then extracted visual features for the Aff-Wild2 dataset.

iResNet100 iResNet100 [22] is a deep learning architecture that combines the strengths of Residual [23] networks and Inception [24] networks, utilizing 100 convolutional layers, inception modules, and residual connections. It offers several advantages over traditional convolutional neural networks, including extracting features at multiple scales and addressing the problem of vanishing gradients in deep networks. iResNet100 has been pre-trained on large-scale datasets and has proven highly effective at feature extraction and transfer learning. We refer to the visual features extracted from the Aff-wild2 dataset using iResNet100 pre-trained on databases from Glint360K [25] and a private commercial FAU database as **ires100**.

MobileNet MobileNet [26] is a convolutional neural network that uses depth-separable convolution to reduce the size and computation of the network while maintaining accuracy. Depth separable convolution involves two operations: deep convolution and point-by-point convolution. The deep convolution is applied to each input channel independently, and the point-by-point convolution combines the output of the deep convolution. This design allows MobileNet to significantly reduce the number of network parameters while achieving comparable accuracy to traditional convolutional neural networks. We refer to the visual features extracted from the Aff-wild2 dataset using MobileNet pre-trained on databases from a private commercial VA database as **mobilenet**.

MAE MAE [27] is a deep network structure that reconstructs an input image by predicting the pixel value of each mask block. MAE pre-training is efficient, simple and does not require any special sparse operation. We refer to the visual features extracted from the Aff-wild2 dataset using MAE pre-trained on databases from DFEW [28], Emotionet [29], FERV39k [30] and a private commercial VA database as **mae**.

DenseNet Traditional convolutional neural networks only receive the output of the previous layer as input, but DenseNet [31] is designed to receive input from all previous layers. This densely connected design enables DenseNet to better leverage the characteristics of previous layers and achieve better performance with fewer layers. Each layer in DenseNet is composed of two sub-layers: the primary sub-layer and the compact connector sub-layer. The primary sub-layer consists of convolution layers, batch normalization layers, and activation function layers, used to extract features. The compact connector sub-layer joins the output of all previous layers and uses it as input to the current layer. This compact connection design allows DenseNet to transfer gradients better, alleviating the vanishing gradients problem and improving training efficiency. We refer to the visual features extracted from the Aff-wild2 dataset using MAE pre-trained on databases from FER+ [32] and AffectNet [33] database as **densenet**.

VIT The VIT [34] model first divides the input image into fixed-size blocks, which are flattened into one-dimensional vectors and passed through a group of Transformer encoders. Each encoder comprises multiple self-attention layers and feedforward neural network layers for feature extraction and encoding. The self-attention mechanism allows the model to focus on relevant areas in the image and capture more visual information. VIT also uses a learnable position embedding vector, representing the position information of each image block, which is concatenated with the feature vector of the image. This embedding method enables VIT to capture spatial information and achieve high performance in image classification tasks. We refer to the visual features extracted from the Aff-wild2 dataset using VIT pre-trained on databases from a private commercial AU database as **vit**.

3.2.2 Audio Feature

Two ways to get audio features; one is the deep network extracting features, including using Wav2vec 2.0 base [35] to extract features which are called **wav2vec**, extracted features on HuBERT [36] are called **hubert**, extracted features on ECAPA-TDNN [37, 38] are called **ecapatdnn**, and the other is using the tradition method Fbank [39] to extract features which are called **fbank**.

Wav2vec 2.0 base Wav2vec 2.0 base divides speech signals into small blocks of fixed length and converts each block into a high-dimensional vector representation. Here we use Mel spectrum features composed of Mel filters and a learnable linear transformation to map the Mel spectrum features to a higher dimensional representation space. Using a large amount of unlabeled speech data, an autoencoder model is trained under the framework of self-supervised learning, with the goal of minimizing the distance between the original speech signal and the reconstructed speech signal. A mask convolution mechanism is used to introduce some random masks between the input and output of the encoder to enhance the robustness of the model. Finally, supervised learning is conducted through fine-tuning and adaptive prediction.

HuBERT HuBERT is a speech recognition model based on mixed input representation. It adopts two different input modes, binaural input and monaural input, and combines convolutional neural network and Transformer network for feature extraction and coding. Specifically, HuBERT employed a set of cross-channel attention mechanisms that weighted and fused the features of binaural and monaural inputs to further improve the model's performance.

ECAPA-TDNN ECAPA-TDNN enhances the architecture of Time Delay Neural Network (TDNN) in multiple ways. It reconstructs the initial frame layer into a one-dimensional Res2Net module with effective skip connections. It aggregates and propagates features from different levels using complementary information with varying complexity in each layer of the neural network. It improves the statistics pooling module with channel-dependent frame attention. This enables the network to focus on different subsets of frames during each of the channel's statistics estimation [38].

Fbank Fbank is a traditional speech feature extraction method, which divides the speech signal into a series of short-time frames, and then carries out a series of filter convolutions for each frame. Finally, the output of each filter is taken logarithm and spliced together to form a fixed dimension feature vector. The advantage of Fbank is that it can compress the speech signal while retaining enough speech information, thus reducing the dimension of the feature vector and improving the efficiency of speech recognition.

3.3. Architectures

CrossAttention+Transformer Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions [16]. To integrate the multi-modal features of vision and audio more effectively, we add CrossAttention before the transformer to handle feature interactions to varying scales using the cross-modal attention mechanism to improve visual and audio feature fusion. The model structure diagram is in Figure 1, and the Multi-head attention formula is below.

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

Dual Transformer Visual and audio features are diverse compared to the same modal features. To make the model distinguish this, we use a dual transformer structure to process audio and visual separately. The model structure diagram is in Figure 1.

TCN+Transformer Taking advantage of TCN's translation invariance and local feature extraction capabilities in sequential data and Transformer's global dependency and context modeling capabilities in sequences, we use a combined model framework training of TCN and Transformer to improve model performance and robustness.

Loss Function AU task is an imbalanced multi-label task. We chose Binary Cross Entropy [40, 41](BCE) Loss in the 5th ABAW competition, creating a criterion that measures the Binary Cross Entropy between the target and the input probabilities, and the formula is below.

$$BCE(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{2}$$

4. Experiments

4.1. Database

To obtain better visual features, we pre-train our models on Glint360K, DFEW, Emotionet, FERV39k, FER+, AffectNet, private commercial AU, and private commercial VA database before extracting visual features. After obtaining these pre-train models, we extract visual features from the AU database in Aff-Wild2.

4.2. Training

In our training for the AU task, we find that longer image sequences lead to better model performance. However, we train on 2 NVIDIA GeForce RTX 3080 Ti GPUs. Due to the limited GPU memory, feeding long image sequences as input is not feasible. Therefore, we first extract features to get visual and audio features and combine them as model input for training which the sequence length is 128.

For feature training, the optimizer is Adam, with a learning rate of 0.0001, and the learning scheduler uses StepLR, which reduces the learning rate by multiplying it by 0.1 every 20 epochs. The total number of training epochs is 50.

4.3. Ablation Study

Data Balancing In Table 1, the performance of the validation set before and after data balancing compare under the same feature combination and network.

Ablation of Features Table 2 shows the single visual feature transformer training experiment that verifies the effectiveness of features. Table 3 shows the fusion of different audio features with the ires100 feature for the training experiment. The different combinations of multi-modal features training experiments and the experimental results are in Table 4.

Ablation of Models The single-model network training experiments use combined features ires100 and ecapatdnn to compare Transformer, TCN, GRU, BIGRU, LSTM, and BiLSTM on the validation set, which results are in Table 5. After that, we select the top-performing Transformer and TCN to design three hybrid model structures for training, which evaluate effects on the validation set shown in Table 5.

Table 1. The performance of data whether balancing on the official validation set.

Data	Features	Model	F1
balancing	ires100,mobilenet	Transformer	0.5508
disbalancing	ires100,mobilenet	Transformer	0.5486

Table 2. The performance of visual features on the official validation set.

Visual Features	Model	F1
ires100	Transformer	0.518
mae	Transformer	0.507
densenet	Transformer	0.504
mobilenet	Transformer	0.488
vit	Transformer	0.486

Table 3. The performance of audio features on the official validation set.

Audio Features	Visual Features	Model	F1
wav2vec	ires100	Transformer	0.531
fbank	ires100	Transformer	0.529
hubert	ires100	Transformer	0.526
ecapatdnn	ires100	Transformer	0.525

Table 4. The performance of different combinations of visual and audio features on the official validation set.

Features	Model	F1
ires100;mae;wav2vec;ecapatdnn	Transformer	0.556
ires100;mobilenet;mae;wav2vec;ecapatdnn	Transformer	0.554
ires100;mae;densenet;wav2vec;ecapatdnn	Transformer	0.554
ires100;mobilenet;mae;wav2vec;fbank;ecapatdnn	Transformer	0.553
ires100;densenet;vit;wav2vec;fbank	Transformer	0.552
ires100;mobilenet;hubert;wav2vec	Transformer	0.551

4.4. Model Ensemble

Model fusion can to some extent avoid model overfitting, increase model robustness, and improve model performance by leveraging the differences between different model structures. After experimenting with different feature combinations and model frameworks, we selected the models with excellent performance on the AU validation set for result fusion, and the fused results are shown in Table 6.

Table 5. The performance of the single models and hybrid models on the official validation set.

Model	Features	F1
Transformer	ires100;ecapatdnn	0.525
TCN	ires100;ecapatdnn	0.519
BiGRU	ires100;ecapatdnn	0.516
BiLSTM	ires100;ecapatdnn	0.515
GRU	ires100;ecapatdnn	0.513
LSTM	ires100;ecapatdnn	0.513
TCN+Transformer	ires100;ecapatdnn	0.531
Dual Transformer	ires100;ecapatdnn	0.526
CrossAttention+Transformer	ires100;ecapatdnn	0.525

Table 6. The performance of the ensemble models on the official validation set.

Model	Features	F1
TCN+Transformer	ires100;mae;wav2vec;ecapatdnn	0.556
Dual Transformer	ires100;mobilenet;mae;wav2vec;ecapatdnn	0.554
Transformer	ires100;mae;densnet;wav2vec;	0.554
CrossAttention+Transformer	ires100;mobilenet;mae;wav2vec;fbank;ecapatdnn	0.553
TCN+Transformer	ires100;densenet;vit;wav2vec;fbank	0.553
Transformer	ires100;mobilenet;mae;wav2vec;fbank;ecapatdnn	0.553
Ensemble		**0.5796**

4.5. K-fold Validation

In order to increase the training data of the model to improve the performance of the model, we used the k-fold method to split the dataset into 7 parts, where the training set was divided into 5 parts and the validation set was divided into 2 parts. The validation results at each fold are shown in Table 7.

Table 7. The performance of the k-fold model on each fold validation set.

Fold	1	2	3	4	5	6	7	avg
F1	0.56	0.53	0.55	0.61	0.56	0.53	0.57	**0.559**

4.6. Test Performance

In this subsection, we describe our final submission strategy for the 5th ABAW competition. We submitted the results five times, the first time using the 12 best performing models on the single AU label for ensemble of model results, the second time using the 6 best performing models on the overall AU for ensemble. The third time is the fusion of seven models by the k-fold method, and the fourth time is the selection of six models for ensemble by train val mix method. The fifth time is to select the 6 models with the best performance on the data set from the 13 models in the train val mix method and the k-fold method for model ensemble.

Finally, on the test set of the 5th ABAW Competition, our K-fold method achieved an F1 score of 54.22%, ranking 2nd in the 5th ABAW Competition. Compared with the best F1 in the 3rd ABAW Competition, this is an increase of 4.33%. The results with other teams are shown in Table 9.

Table 8. The performance of these five strategies on the test set.

Submission	Strategy	F1
1	AU top Ensemble	0.5169
2	Model Ensemble	0.5286
3	K-fold	**0.5422**
4	Train-Val-Mix	0.5352
5	K-fold + Train-Val-Mix	0.5404

Table 9. The performance of top teams perform on the Aff-Wild2 test set. * representing teams in the 3rd ABAW competition.

Teams	F1
Netease Fuxi Virtual Human*	0.4989
SituTech* [Our Team]	0.4982
PRL*	0.4904
Netease Fuxi Virtual Human	0.5549
SituTech [Our Team]	0.5422
USTC-IAT-United	0.5144
SZFaceU	0.5128
PRL	0.5101

5. Conclusion

This paper introduces our Facial Action Unit (AU) recognition task training method in the 5th ABAW competition. Through experiments on the AU dataset of Aff-Wild2, our data balancing method makes minor AUs perform better, combined training of multimodal features extracted from different ways can improve model performance, and proposed hybrid model structures are more effective than single networks. We also learn many training skills and model structure methods [42–46] during this process on previous researchers' contributions to affective computing. We hope this paper inspires more researchers in this field.

References

[1] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[2] 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.

[3] Cvpr 2023: 5th workshop and competition on affective behavior analysis in-the-wild (abaw). `https://ibug.doc.ic.ac.uk/resources/cvpr-2023-5th-abaw/`, 2023.

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

[5] 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.

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

[7] 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.

[8] 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.

[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.

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

[11] 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.

[12] 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.

[13] 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.

[14] Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, 129(12):3174–3194, sep 2021.

[15] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

[16] 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.

[17] Colin Lea, Michael Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[18] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[19] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1317–1325, 2015.

[20] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*, 2014.

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. In *Neural Computation*, volume 9, pages 1735–1780. MIT Press, 1997.

[22] Ionut Cosmin Duta, Li Liu, Fan Zhu, and Ling Shao. Improved residual networks for image and video recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9415–9422. IEEE, 2021.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[25] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021.

[26] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1382–1391. IEEE, 2017.

[27] 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.

[28] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2881–2889, 2020.

[29] Carlos Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martínez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5562–5570. IEEE Computer Society, 2016.

[30] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene

dataset for facial expression recognition in videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20890–20899. IEEE, 2022.

[31] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[32] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM International Conference on Multimodal Interaction (ICMI)*, 2016.

[33] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[34] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3156–3164. IEEE, 2021.

[35] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[36] Jacob Hubert, Ankit Goyal, Erich Elsen, Lucas Theis, William Fedus, Stephan Gouws, and Michael Auli. Leveraging pre-trained checkpoints for sequence generation tasks. *arXiv preprint arXiv:2103.14062*, 2021.

[37] Voiceprintrecognition-pytorch. `https://github.com/yeyupiaoling/VoiceprintRecognition-Pytorch`, 2023.

[38] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

[39] Stephen B Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[40] A proposal for computing the multi-attribute utility of alternatives. *Management science*, 9(3):67–80, 1959.

[41] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. A brief review of the mathematical foundations of gradient-based deep learning. *arXiv preprint arXiv:1804.08838*, 2015.

[42] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. Multi-feature based emotion recognition for video clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 630–634, 2018.

[43] Chuanhe Liu, Wenqiang Jiang, Minghao Wang, and Tianhao Tang. Group level audio-video emotion recognition using hybrid networks. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 807–812, 2020.

[44] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Wenqiang Jiang, Tenggan Zhang, Chuanhe Liu, and Qin Jin. Valence and arousal estimation based on multimodal temporal-aware features for videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2345–2352, June 2022.

[45] Wenqiang Jiang, Yannan Wu, Fengsheng Qiao, Liyu Meng, Yuanyuan Deng, and Chuanhe Liu. Model level ensemble for facial action unit recognition at the 3rd abaw challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2337–2344, June 2022.

[46] Tenggan Zhang, Chuanhe Liu, Xiaolong Liu, Yuchen Liu, Liyu Meng, Lei Sun, Wenqiang Jiang, Fengyuan Zhang, Jinming Zhao, and Qin Jin. Multi-task learning framework for emotion recognition in-the-wild. In *European Conference on Computer Vision*, pages 143–156. Springer, 2023.