EmotiEffNet and Temporal Convolutional Networks in Video-based Facial Expression Recognition and Action Unit Detection

Andrey V. Savchenko[1,2]
[1]Sber AI Lab
Moscow, Russia
[2]HSE University
Laboratory of Algorithms and Technologies for Network Analysis, Nizhny Novgorod, Russia
avsavchenko@hse.ru

Anna P. Sidorova
HSE University
Nizhny Novgorod, Russia
anna.pav.sidorova@gmail.com

Abstract

This paper examines the video-based facial expression recognition and action unit detection tasks. We propose to use pre-trained EmotiEffNet models for frame-level facial feature extraction and feed them into the Temporal Convolutional Networks to take into account the dynamics of facial expressions. In addition, we study the possibility of combining facial processing with audio feature extraction to improve the accuracy of audio-visual expression recognition. Experimental results for two tasks from the sixth Affective Behavior Analysis in-the-Wild challenge demonstrate that our approach lets us significantly improve quality metrics on validation sets compared to existing non-ensemble techniques. As a result, our approach took third place in the action unit detection and fourth place in the expression recognition.

1. Introduction

Emotion recognition has garnered substantial interest due to its potential applications in various fields, including human-computer interaction, virtual reality, mental health assessment, and multimedia content analysis [5, 9, 56]. Conventional machine learning techniques face challenges in accurately categorizing human emotions, highlighting the necessity for advanced approaches to effectively interpret the intricate and subtle nuances of human emotions conveyed through audio-visual cues.

The rapid progress in deep learning and neural network algorithms has created new opportunities to enhance the accuracy and reliability of emotion recognition systems [21, 22]. Utilizing neural network models allows for the automatic extraction of high-level representations from raw audio and visual data, enabling more effective recognition of subtle emotional cues and expressions. Despite the promising potential of deep learning, the research gaps and contextual emotion recognition issues still exist.

Companies, research institutions, and developer teams collaborate to create highly accurate emotion recognition algorithms. For example, the recent studies have shown the benefits of 3D representations [10, 28–30] and multi-modal techniques [54, 57]. One of the most famous competitions to compare the quality of these algorithms is the Affective Behavior Analysis in-the-wild (ABAW) [15, 17, 19, 20, 23–26, 55]. The main tasks in recent ABAW Challenges were frame-level, video-based, facial expression recognition (EXPR) and action unit (AU) detection [16]. Using a model based on EfficientNet [42], precise results were achieved across all tasks [12, 38, 39, 61]. The team that proposed an audio-visual fusion model combining transformers and temporal convolutional networks (TCN) took third place in the EXPR competition [61].

Notably, one of the most promising outcomes was achieved using an ensemble of audio-visual transformer-based models, incorporating fine-tuned masked autoencoder (MAE) technology [58]. However, most previously proposed techniques are expensive in terms of both memory and running time. Some researchers did not fully appreciate the unique characteristics of transformers, specifically regarding the difficulty in accounting for the relationship between frames in frame-level classification [34], which leads

to a decrease in generalization and the need for retraining.

In this paper, we describe a novel methodology to utilize the lightweight models from the EmotiEffNet-family [44], as well as the MobileViT [32], MobileFaceNet [4], and DDAMFN [56]. We also discuss various architectures and methodologies employed to address these challenges. Additionally, we present a successful implementation of a sequential classification method based on the Temporal Convolutional Network (TCN). The training methodology played a significant role. Specifically, applying weighting to the loss function to avoid overfitting and accounting for contextual distortions due to box transformations (i.e., smoothing of predictions) were essential considerations.

The remaining part of the paper is organized as follows. In Section 2, we review the methodologies used by participants in the Emotion Classification and Action Unit Detection Challenge, identifying common approaches and analyzing their results. Section 3 describes our approach, detailing the models used to extract features and the design of methods employed to solve the given tasks. Additionally, we address the nuances of learning and predicting outcomes. Following this, we compare and analyze the experimental results of our approach with existing competitors in Section 4, concluding the evident advantages and limitations of each technique in Section 5.

2. Related work

The trend towards using sequence processing techniques (e.g., transformers and competitive neural networks) is growing. Some teams are expanding their efforts by incorporating additional modalities, improving their methods by applying new techniques, or increasing the depth of their models. This indicates that, despite the availability of well-established methods at present, there is still significant potential for further performance improvement. Let us briefly describe the details of two tasks from the sixth ABAW competition: facial expression recognition and AU detection.

2.1. Facial Expression Recognition

This task is a multi-class classification problem, which aims to accurately recognize one of eight emotions at a given moment (on a frame-by-frame basis). Kollias et al. [26] provided information on the number of instances in each class (Table 1). Here, the disparity in class levels makes the task very challenging.

Compared to the results from ABAW 2023 [16], the metric's performance has significantly [43] improved (F1 score: 0.4094 $\rightarrow$ 0.5005). The winners are Zhang et al. [59], who use audio and video modalities. Their concept has been maintained: the authors employ the MAE encoder that has been fine-tuned on a large dataset to highlight visual features and VGGish to extract audio-specific features. An encoder transformer is also used to take contextual information into account. The model is based on ensemble learning, and a Gaussian filter has been used to reduce the impact of outliers. Although this method is considered the most accurate, it is still rather expensive.

Table 1. Expression Classification Challenge: Number of Annotated Images for each Expression

Expressions	Number of frames
Neutral	468,069
Anger	36,627
Disgust	24,412
Fear	19,830
Happiness	245,031
Sadness	130,128
Surprise	68,077
Other	512,262

Zhou et al. [62] achieved the top two positions in this task. They followed a similar approach to last year, using TCN. However, their methodology changed in visual sign recognition, and inspired by the work of Zhang et al. [59], the researchers decided to adapt MAE to identify visual signs and enhance the metric indicator.

The distinctive features of TCNs include 1) employing causal convolutions to prevent information leakage from future to past and 2) enabling the processing of sequences of varying lengths, akin to RNNs, while also emphasizing strategies for achieving extended effective history sizes through the use of deep networks with residual layers and dilated convolutions [3].

Jun et al. [53] proposed a slightly different method for solving this problem. They used semi-supervised learning [28] to do so. The key difference between this approach and others lies in selecting data to ensure an even distribution of examples and using data augmentation methods to enhance robustness.

2.2. Action Unit Detection

The action unit detection task is a multi-label classification problem. The main challenge lies not only in the balancing of the data but also in the rapidity of changes in facial expressions. In total, there are 12 different classes, with a specific distribution that is illustrated by the organizers of this challenge [26] (Table 2).

The methods listed in Subsection 2.1 are also used to solve this task. This indicates the variability and correlation among approaches to emotion understanding. However, several papers have concentrated only on AU detection. For example, the fifth team in the ABAW-5 competition used the Regnet backbone, which was implemented in the video Vision Transformer [46]. The local region percep-

Table 2. Action Unit Detection Challenge: Distribution of AU Annotations in Aff-Wild2

Action Unit #	Action	Total Number of Activated AUs
AU 1	inner brow raiser	301,102
AU 2	outer brow raiser	139,936
AU 4	brow lowerer	386,689
AU 6	cheek raiser	619,775
AU 7	lid tightener	964,312
AU 10	upper lip raiser	854,519
AU 12	lip corner puller	602,835
AU 15	lip corner depressor	63,230
AU 23	lip tightener	78,649
AU 24	lip pressor	61,500
AU 25	lips part	1,596,055
AU 26	jaw drop	206,535

tion was introduced in [52] based on a graph neural network relational learning module and IResnet100 feature extractor pre-trained on pre-trained on Glint360K. The high accuracy was obtained by the spatio-temporal representation learning [50] with MAE and temporal graph embeddings [31]. Finally, an ensemble of six transformer-based models that analyze various visual (IresNet-100, MobileNet, MAE, etc.) and audio (wav2vec, fbank, etc.) features was presented in [8]. Though such complex models may reach state-of-the-art results, their practical usage is limited due to high runtime and space complexity [37]. The following section introduces an approach based on lightweight visual models.

3. Methodology

3.1. Overview

In this paper, the facial features are extracted from each frame by using the lightweight neural networks from HSEmotion library [7, 13], such as EmotiEffNet [39] and several models pre-trained in multi-task (MT) fashion, namely, MT-EmotiEffNet [40], MT-EmotiMobileFaceNet, MT-DDAMFN and MT-EmotiMobileViT [41]. While this method may not be the most advanced approach for a specific dataset, it highlights the critical need for a model capable of emotion analysis that can function in real-world uncontrolled environments. As a result, feed-forward neural networks similar to those used in this study are employed for facial expression recognition and AU detection in the ABAW Challenge using the Aff-Wild2 dataset [18]. These models are trained on all annotated frames from the competition training sets.

Another approach is based on the research conducted by CtyunAI [61], which has been at the forefront of the field for several years. This approach utilizes the TCN model for continuous emotion recognition.

In this paper, we proposed to combine EmotoiEffNet features and TCN. The complete model is presented in Fig. 1. Here, we compute several audio features and combine them with EmotiEffNet-based facial embeddings in several TCN models. Their outputs are concatenated and fed into a multilayer perceptron (MLP) to predict facial expressions or action units. Let us describe the details of our approach.

3.2. Audio features

We employ the wav2vec 2.0 model [2] to extract audio features and utilize this embedding for our EmotiEffNet model. The wav2vec 2.0 applies a mask to the speech input in latent space and solves a contrastive learning task defined over the quantized latent representations, which are learned jointly.

We also identify a range of additional acoustic features by utilizing OpenL3 [1, 6], wav2vec2-hubert [51] and wav2vec2-large-robust-emotion [47] models.

OpenL3 is an embedding trained through self-supervised learning of audio-visual correspondence in videos instead of other embeddings requiring labeled data. This framework can potentially produce powerful out-of-the-box embeddings for downstream audio classification tasks. Still, several unexplained design choices may impact the embeddings' behavior [6]. It has several hyperparameters that can be tuned to highlight specific features. One of these is the version of the model. The version depends on the data the model was trained on, specifically whether it was trained on a music or environmental subset. The environmental subset, selected for its inclusion of human sounds, animal sounds, and other natural acoustic environments, provides a comprehensive representation for various applications such as sound analysis and immersive experiences. The embedding size is 512, and the input representation is a Mel spectrogram with 128 filters.

The base model for wav2vec2-hubert is hubert-base-ls960 [11], which was trained for 960 hours using 16 kHz sample speech audio. Wav2vec2-Hubert was trained on the IEMOCAP dataset, which contains neutral, happy, sad, and angry speech recordings.

The wav2vec2-large-robust-emotion [47] is a model for dimensional speech emotion recognition based on the wav2vec 2.0. The model was developed by fine-tuning a pre-trained wav2vec2-large-robust model, trained initially on the MAP-Podcast dataset. The authors reduced the number of transformer layers from 24 to 12 before fine-tuning them to optimize the model's performance.

3.3. Video features

In this paper, the models of the EmotiEffNet family were utilized to highlight various visual attributes [39, 42]. In particular, the EmotiEffNet-B0 [38] was employed to emphasize visual features for the TCN model (across cropped

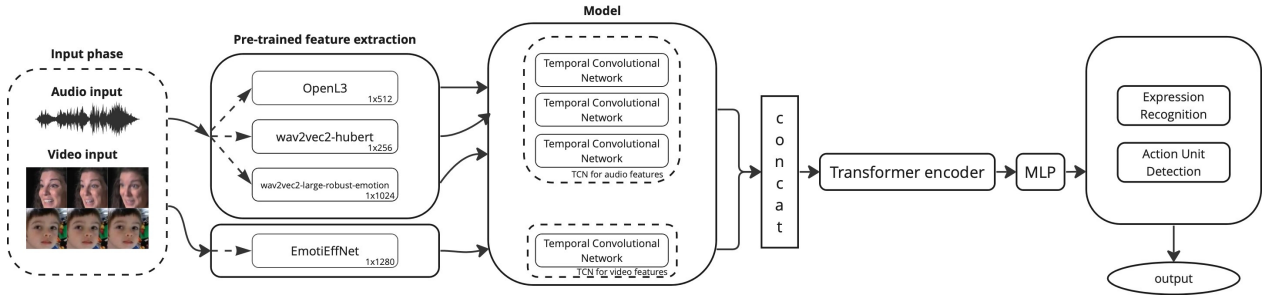

Figure 1. Proposed EmotiEffNet + TCN model

and aligned frames). This model was pre-trained to identify faces from the VGGFace2 dataset and later fine-tuned to predict facial expressions from the AffectNet dataset.

In addition, we opted to employ a variety of widely recognized lightweight neural network architectures, including MobileViT [32], MobileFaceNets [4], and DDAMFN [56]. Similarly to EmotiEffNet, these models have been pre-trained to recognize faces. However, the MT training was used to simultaneously predict facial expressions, valence, and arousal from AffectNet. The resulting models are MT-EmotiMobileViT, MT-EmotiMobileFaceNet, and MT-DDAMFN [41]. Each labeled video frame [26] is fed into these models, and the embeddings $\mathbf{x}$ are extracted at the output of the penultimate layer of EmotiEffNet network [39].

3.4. Frame-level Facial Emotion Classification

To make a final decision for given embeddings $\mathbf{x}$ for both EXPR and AU tasks, we employed an MLP with a single hidden layer containing 128 units as our classifier. Additionally, we used the wav2vec 2.0 model to calculate embeddings from the audio corresponding to video data [2] in conjunction with our models. Since the number of audio and video frames differ, we aligned features of an acoustic frame with the closest video frame [41].

For optimization, we utilized Adam across all scenarios, given that our models typically deal with high-resolution facial images. We processed the directory containing cropped faces provided officially by the organizers of the 6th ABAW competition [26].

3.4.1 EXPR classification

For EXPR classification, we use softmax activations and weighted categorical cross-entropy:

$$\mathfrak{L}_{EXPR}(y, \hat{\mathbf{z}}) = -w_y \cdot \log(\hat{\mathbf{z}}_y), \quad (1)$$

where y is the ground-truth emotional category, $\hat{\mathbf{z}}$ is the vector at the output of the last softmax layer, $\hat{\mathbf{z}}_y$ is the y-th component of vector $\hat{\mathbf{z}}$, and the weights w_y for each facial expression y are inversely proportional to the number of frames of this class in the training set (Table 1).

The training dataset consists of 585,317 annotated images extracted from cropped facial regions within 248 videos supplied by the organizers. Validation testing was conducted on 280,532 images from 70 of these videos, using a model that had been trained for ten iterations.

3.4.2 AU detection

Sigmoid activations and multi-class weighted binary cross-entropy were utilized for AU detection:

$$\mathfrak{L}_{AU}(\mathbf{y}, \hat{\mathbf{y}}) = \\ -\sum_{k=1}^{K} w_k \cdot (y_k \cdot \log(\hat{y}_k) + (1 - y_k) \cdot \log(1 - \hat{y}_k)), \quad (2)$$

where $K = 12$ is the total number of AUs in this challenge, $\mathbf{y} = [y_1, ..., y_K]$ is the K-dimensional vector of 0-1 ground-truth AU labels y_k, $\hat{\mathbf{y}} = [\hat{y}_1, ..., \hat{y}_K]$ is the output of last K sigmoid units, and weights w_k are inverse proportional to the number of k-th AU units with positive ($y_k = 1$) and negative ($y_k = 0$) labels.

The training dataset comprises 1,356,694 labeled facial images extracted from 295 video files. The validation dataset contains 445,836 images derived from the remaining 105 video files.

3.5. Temporal Convolutional Networks approach

Four vectors are accepted as input for both tasks, namely, embeddings of EmotiEffNet-B0 as visual features and three sets of audio features (OpenL3, wav2vec2-hubert, and wav2vec2-large-robust-emotion). The TCN layers are applied to each of these sets of features to produce the final output:

1. EmotiEffNet-B0 visual features:

$$\mathbb{R}^{(b,1280,w)} \rightarrow \mathbb{R}^{(b,512,w)} \rightarrow \mathbb{R}^{(b,256,w)} \rightarrow \mathbb{R}^{(b,128,w)}$$

2. wav2vec2-large-robust-emotion:

$$\mathbb{R}^{(b,1024,w)} \rightarrow \mathbb{R}^{(b,512,w)} \rightarrow \mathbb{R}^{(b,256,w)} \rightarrow \mathbb{R}^{(b,128,w)}$$

3. OpenL3:

$$\mathbb{R}^{(b,512,w)} \rightarrow \mathbb{R}^{(b,256,w)} \rightarrow \mathbb{R}^{(b,128,w)}$$

4. wav2vec2-hubert audio features:

$$\mathbb{R}^{(b,256,w)} \rightarrow \mathbb{R}^{(b,128,w)}$$

.

Here b is the batch size, and w is the TCN window size. Since TCN analyzes the sequence, it is necessary to determine the window size w and shift hyperparameters. We have chosen a window size of 300 and a change of 200. This captures 100 values from each previous window, considering dependencies and increasing the training data set.

Then, these representations are fed into the transformer encoder and MLP, after which a softmax function is applied to produce a probabilistic output. The exact losses described in Section 3.4 were utilized (1) and (2) for EXPR and AU, respectively. The Adam optimizer was used during training for both tasks.

3.6. Final Predictions

We studied several post-processing techniques to make final predictions with our pipeline. First, we used the pre-trained EmotiEffNet that predicts seven basic facial expressions from the AffectNet dataset. If the model is reliable, i.e., the maximal score at the output of the softmax layer is greater than a fixed threshold $t \in [0.5, 1]$, the predicted expression is used for the frame. Otherwise, the MLP with a single hidden layer trained on the AffWild2 dataset from ABAW-6 is applied. Second, in addition to MLP, we used the LightAutoML library [45] to classify outputs of pre-trained model [39].

Finally, creating model predictions generally involves the steps described below. Due to the fact that decisions are made frame by frame, noise can be introduced into the outputs of trained feed-forward neural networks. To address this issue, a smoothing technique is applied to the predictions for each frame within a short time window [39], using a box filter. It calculates the arithmetic mean of the predicted values for all frames within the specified window. Given that action units (AU) representing facial muscle movements can exhibit rapid variations, five frames (2 frames before and two frames after the current frame) are used. On the contrary, facial expressions, which have a longer duration, are more likely to be persistent. For our investigation, we utilized windows comprising 50 frames.

4. Experimental Results

4.1. Evaluation metric

To gauge the model's performance in each track, the ABAW established distinct evaluation metrics for each challenge. Let's look at them.

	w2v2-hub	w2v2-large	OpenL3	all
Accuracy	0.339	**0.354**	0.308	0.348
F1-score	0.235	0.249	0.26	**0.276**

Table 3. The results of applying the MLP model to the described features

Model	all classes		w/o "Other"	
	F1-score	Accuracy	F1-score	Accuracy
EmotiEffNet-B2	0.229	0.282	0.320	0.443
DDAMFN	0.244	0.315	0.362	0.502
MT-DDAMFN	0.245	**0.340**	0.366	**0.547**
MT-EmotiMobileViT	0.248	0.287	0.330	0.434
MT-EmotiMobileFaceNet	0.250	0.325	0.354	0.513
MT-EmotiEffNet	0.254	0.324	0.381	0.522
EmotiEffNet-B0	**0.257**	0.325	**0.383**	0.522

Table 4. Results of pre-trained facial expression recognition models on the Aff-Wild2's validation set. The best result is marked in bold.

EXPR Recognition For the evaluation of emotion recognition, the macro F1 score metric is used:

$$P_{EXPR} = \frac{1}{8}\sum_{i=1}^{8} F1_i. \tag{3}$$

Here the F1-score $F1_i$ for each expression category i is defined as follows

$$\begin{aligned} F1_i &= \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}, \\ Precision_i &= \frac{TP_i}{TP_i + FP_i}, \\ Recall_i &= \frac{TP_i}{TP_i + FN_i}, \end{aligned} \tag{4}$$

where TP_i, FP_i, FN_i and TN_i are True Positives, False Positives, False Negatives, and True Negatives statistics for i-th emotional class, respectively.

AU Detection The evaluation of performance involves computing the average F1-score across all 12 categories, which is expressed as:

$$P_{AU} = \frac{\sum_{i=1}^{12} F1_i}{12}, \tag{5}$$

where $F1_i$ for each AU is defined similarly to Eq. 4.

4.2. Training Details

All experiments were conducted on two GPU NVIDIA GeForce RTX 2080 Ti devices with 12 GB of memory. The TensorFlow 2 framework was utilized to train the models,

Method	Modality	F1-score P_{EXPR}	Accuracy
Baseline VGGFACE (MixAugment) [26]	Faces	0.25	-
EfficientNet-B0 [38]	Faces	0.402	-
Meta-Classifier [49]	Faces	0.302	0.462
TCN [61]	Audio/video	0.377	-
Transformer [60]	Audio/video	0.406	-
MAE [58]	Audio/video	0.495	-
TCN+MLP	Audio	0.151	0.412
wav2vec 2.0	Audio	0.291	0.410
wav2vec 2.0, smoothing	Audio	0.355	0.521
DDAMFN	Faces	0.308	0.433
EmotiEffNet-B2	Faces	0.320	0.438
MT-EmotiMobileFaceNet	Faces	0.327	0.462
MT-EmotiEffNet	Faces	0.336	0.447
MT-DDAMFN	Faces	0.351	0.469
MT-EmotiMobileViT	Faces	0.356	0.461
EmotiEffNet	Faces	0.384	0.495
EmotiEffNet, smoothing	Faces	0.424	0.543
EmotiEffNet, pre-trained + MLP	Faces	0.395	0.4977
EmotiEffNet, pre-trained + MLP, smoothing	Faces	0.434	0.5463
wav2vec 2.0+EmotiEffNet	Audio/video	0.403	0.520
wav2vec 2.0+EmotiEffNet, smoothing	Audio/video	0.434	0.557
TCN (aligned frames) + MLP	Audio/video	0.353	0.536
TCN (cropped frames) + MLP	Faces	0.403	0.523
EmotiEffNet (aligned) + TCN + transformer	Faces	0.338	0.51
EmotiEffNet (cropped) + TCN + transformer	Audio/video	0.375	0.52
EmotiEffNet (aligned)+ TCN + transformer	Audio/video	0.422	0.55

Table 5. Expression Challenge Results on the Aff-Wild2's validation set.

while the visual features were extracted using PyTorch pre-trained EmotiEffNet models.

The optimizer's learning rate in the TCN approach was set to 1e-4. The training process consisted of between 100 and 200 training epochs. The transformer encoder consisted of four multi-attention heads and eight layers.

The Adam optimizer with a learning rate 1e-3 was used to train MLP classifiers. The models were trained on ten epochs for validation purposes, and the model from the top-performing epoch on the validation set was analyzed. It was experimentally found that the best EXPR MLP classifier is obtained right after the first epoch, while the AU detector typically needs five epochs. Hence, the test set predictions are obtained by the model trained on concatenated training and validation sets on 1 and 5 epochs for EXPR and AU tasks, respectively.

4.3. EXPR Classification

4.3.1 Audio Features

The selected audio was divided into K segments of 90- and 60-second durations to extract the features. For audio clips of 90 seconds in length, features were identified using pre-trained models wav2vec2-hubert and wav2vec2-large-robust-emotion. For clips of 60 seconds, OpenL3 was used. The choice is based on the frequency of fluctuating emotions, which is approximately one minute or a half.

The results of applying an MLP to the selected audio features are presented in Table 3. In practice, OpenL3 embeddings have proven to be more effective. Nevertheless, all three feature sets were used for the TCN.

4.3.2 Visual Pre-trained Models

Initially, we assessed the performance of pre-trained models for emotion classification (Table 4) without utilizing the training set. A significant challenge arises from differences in class names: AffectNet includes the "Contempt" category, whereas Aff-Wild2 contains numerous instances labeled as "Other".

We investigated two approaches for matching classes: assigning all "Contempt" predictions as "Other", or excluding the "Contempt" class from predictions and "Other" from the validation dataset. The EmotiEffNet-B0 model

Model	P_{EXPR}
EmotiEffNet+TCN (train+val)	0.3043
EmotiEffNet, audio+vid	0.3137
EmotiEffNet+TCN+audio	0.3221
EmotiEffNet (train+val)	0.3200
EmotiEffNet+TCN	0.3207
EmotiEffNet+TCN	0.3221
EmotiEffNet+TCN, smoothing	0.3244
wav2vec 2.0+EmotiEffNet (train+val)	0.3301
EmotiEffNet, pre-trained + MLP (train+val)	**0.3414**

Table 6. Expression Challenge Results on the Aff-Wild2's test set: the diversity of our approaches.

Model	P_{EXPR}
Netease Fuxi AI Lab [59]	0.5005
CtyunAI [62]	0.3625
USTC-IAT-United [54]	0.3534
Ours	**0.3414**
M2-Lab-Purdue [27]	0.3228
KBS-DGU [14]	0.3005
SUN_CE [36]	0.2877
AIOBT[35]	0.2797
CAS-MAIS [48]	0.2650
IMLAB [33]	0.2296
Baseline VGGFACE [26]	0.2250

Table 7. Expression Challenge Results on the Aff-Wild2's test set: Leaderboard.

produced the highest F1-score for emotion recognition and outperformed the other models. Its embeddings demonstrated outstanding classification performance (Table 5). However, the MT-DDAMFN approach achieved the best accuracy for seven classes, surpassing the initial DDAMFN [56] result by 4%.

4.3.3 Audio-Visual Models

Table 5 provides the validation outcomes for traditional tasks. We have compared our findings with the baseline models supplied by the challenge organizers [26], as well as several papers presented at the ABAW CVPR 2023 workshop. It has been noted that there are no significant improvements when using features from models that were trained in a multi-task fashion. Interestingly, the wav2vec 2.0 embeddings technique demonstrated notable performance within the EXPR challenge, achieving top performance through the simple combination of predictions made by leading visual (EmotiEffNet-B0) and audio models.

We have taken the cropped and aligned embeddings presented in the ABAW competitions for comparison purposes. Table 5 presents such experimental results (noted in parentheses with *cropped* or *aligned* text).

The results in the overall leaderboard are shown in Table 7. Here, our approach is much better than the baseline but still worse than the top-performing team. As a result, our solution took fourth place in the EXPR competition. All or results in the test set are presented in Table 6.

The ablation study of our best model that utilizes pre-trained EmotiEffNet scores is shown in Fig. 2. Here, we demonstrate the dependence of the F1 score on the threshold for blending pre-trained scores and the output of the trained MLP classifier. As one can notice, the best results are obtained for a threshold lower than 1.0, which means the efficiency of pre-trained scores. We re-trained the MLP classifier on the union of training and validation set after the challenge and obtained the top F1 score of 0.3629. If

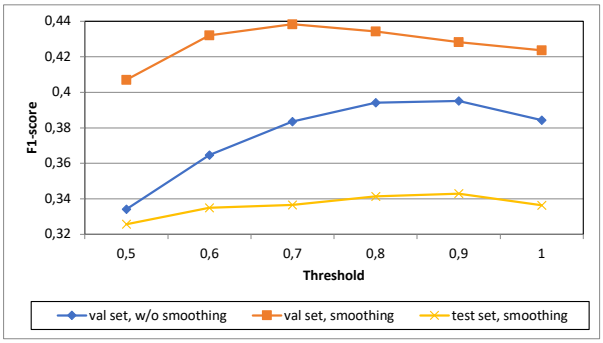

Figure 2. Dependence of F1-score for blending of pre-trained scores and output of MLP classifier on threshold, EXPR classification challenge

we succeeded in submitting the results for this model to the official competition, we would take second place.

4.4. AU Detection

We have also explored the basic approaches to solving the problem for this task. These approaches are presented in Table 8. The RegNet, Masked Autoencoder, and Transformer models achieved the best results.

Given that Transformers are one of the most widely used and stable methods, we have incorporated them into our approaches. The results of these models are also presented in Table 8. We have also compared the performance of our implemented techniques with the baseline provided and some of the outcomes from the ABAW 2023 CVPR workshop.

The best approach we have implemented regarding the F1 Score on the validation set is using EmotiEffNet as a visual representation for LightAutoML. The final F1 Score on the validation dataset was 0.554, representing a 41% improvement over the baseline model.

Test values were obtained and are presented in Table 9. As a result, our solutions based on LightAutoML and TCN

Method	Modality	F1-score P_{AU}
Baseline VGGFACE [26]	Faces	0.39
IResnet100 [52]	Faces	0.511
TCN [61]	Audio/video	0.517
Transformer [60]	Audio/video	0.530
Regnet/Video Vision Transformer [46]	Faces	0.540
MAE graph representations [50]	Faces	0.543
MAE [58]	Audio/video	0.567
Regnet [49]	Faces	0.698
wav2vec 2.0	Audio	0.313
DDAMFN	Faces	0.500
MT-EmotiMobileFaceNet	Faces	0.512
MT-DDAMFN	Faces	0.519
EmotiEffNet (aligned)+TCN+transformer	Video	0.525
MT-EmotiEffNet	Faces	0.525
EmotiEffNet (aligned)+TCN+transformer	Audio/video	0.528
EmotiEffNet	Faces	0.537
EmotiEffNet, smoothing	Faces	0.545
EmotiEffNet + LightAutoML	Faces	0.542
EmotiEffNet + LightAutoML, smoothing	Faces	0.554

Table 8. Action Unit Challenge Results on the Aff-Wild2's validation set.

Model	P_{AU}
EmotiEffNet	0.4726
MT-DDAMFN (train+val)	0.4763
TCN+EmotiEffNet+audio	0.4817
TCN+EmotiEffNet	0.4866
EmotiEffNet + LightAutoML	**0.4878**

Table 9. Action Unit Challenge Results on the Aff-Wild2's test set: the diversity of our approaches.

ranked among the top three in this competition (Table 10).

5. Conclusion

To conclude, we have introduced approaches for solving the Expression Recognition and Action Unit Detection tasks using TCN and EmotiEffNet (Fig. 1). Using the data from the sixth ABAW competition [26], we explored these methods in terms of visual and audio modalities. Our results notably outperformed the baseline results. For example, the best-performing models achieved the following results on the official validation sets: macro-averaged F1-score for facial expression recognition $P_{EXPR} = 0.434$, representing a 0.19 improvement over the baseline VGGFace model with MixAugment, Table 5). AU detection also performed better, achieving $P_{AU} = 0.545$, exceeding the baseline value by 0.15 points, Table 8). The code required for replicating these experiments is publicly available[1]. As a result, we took third place in AU detection and fourth place in EXPR classification challenges. Moreover, as we mentioned in Subsection 4.3.3, by using another MLP trained on the union of train and validation set, we obtained the F1-score that would make it possible to rank in top-2 of the EXPR challenge, proving the potential of our approach.

Model	P_{AU}
Netease Fuxi AI Lab [59]	0.5601
CtyunAI [62]	0.4941
Ours	**0.4878**
USTC-IAT-United [54]	0.484
KBS-DGU	0.4652
M2-Lab-Purdue [27]	0.3832
Baseline VGGFACE [26]	0.365

Table 10. Action Unit Challenge Results on the Aff-Wild2's test set: Leaderboard.

However, our models have specific drawbacks. For instance, in methods using transformers or transformer encoders, it has been observed that the models do not fully consider the influence and interconnection between frames within a video, as also mentioned by Zhou et al. [61] Regarding TCN, it should be noted that due to its use of a sliding window, it can exhibit weaknesses in terms of accurately detecting "fast" emotional states (e.g., surprise or fear) or is unstable in terms of action unit detection, as people's expressions can be pretty dynamic. Additionally, it has also been observed that if an emotion is predominant in the audio input (e.g., due to an external factor such as music or a film, or if just one person within a frame is expressing a particular emotion and the other is expressing the opposite emotion), then the TCN model may overfit and classify a more intense emotional state that is present in the audio segment. TCNs are also sensitive to the length of input sequences, so selecting the window size and shifting carefully is essential. All these disadvantages should be carefully taken into account in the future.

Acknowledgements. The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE).

[1] https://github.com/av-savchenko/face-emotion-recognition/blob/main/src/ABAW/ABAW6/abaw6_affwild2.ipynb

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017.

[2] 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.

[3] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[4] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices. In *Proceedings of the 13th Chinese Conference on Biometric Recognition (CCBR)*, pages 428–438. Springer, 2018.

[5] M Kalpana Chowdary, Tu N Nguyen, and D Jude Hemanth. Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications*, 35(32):23311–23328, 2023.

[6] Aurora Linh Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856, 2019.

[7] Polina Demochkina and Andrey V Savchenko. MobileEmotiFace: Efficient facial image representations in video-based emotion recognition on mobile devices. In *Proceedings of ICPR International Workshops and Challenges on Pattern Recognition, Part V*, pages 266–274. Springer, 2021.

[8] Yuanyuan Deng, Xiaolong Liu, Liyu Meng, Wenqiang Jiang, Youqiang Dong, and Chuanhe Liu. Multi-modal information fusion for action unit detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5855–5862, 2023.

[9] Berat A Erol, Abhijit Majumdar, Patrick Benavidez, Paul Rad, Kim-Kwang Raymond Choo, and Mo Jamshidi. Toward artificial emotional intelligence for cooperative social human–machine interaction. *IEEE Transactions on Computational Social Systems*, 7(1):234–246, 2019.

[10] Ruian He, Zhen Xing, Weimin Tan, and Bo Yan. Unsupervised disentangling of facial representations with 3D-aware latent diffusion models. *arXiv preprint arXiv:2309.08273*, 2023.

[11] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[12] Ilia Indyk and Ilya Makarov. MonoVAN: Visual attention for self-supervised monocular depth estimation. In *Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 1211–1220. IEEE, 2023.

[13] AS Kharchevnikova and AV Savchenko. Neural networks in video-based age and gender recognition on mobile platforms. *Optical Memory and Neural Networks*, 27:246–259, 2018.

[14] Jun-Hwa Kim, Namho Kim, Minsoo Hong, and Cheesun Won. Cca-transformer: Cascaded cross-attention based transformer for facial analysis in multi-modal data. *Preprints*, 2024.

[15] 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 (CVPR)*, pages 2328–2336, 2022.

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

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

[18] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-Wild2, multi-task learning and ArcFace. *arXiv preprint arXiv:1910.04855*, 2019.

[19] 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.

[20] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second ABAW2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3652–3660, 2021.

[21] 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.

[22] 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.

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

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

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

[26] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (ABAW) competition. *arXiv preprint arXiv:2402.19344*, 2024.

[27] Li Lin, Sarah Papabathini, Xin Wang, and Shu Hu. Robust light-weight facial affective behavior recognition with CLIP. *arXiv preprint arXiv:2403.09915*, 2024.

[28] Albert Luginov and Ilya Makarov. Swiftdepth: An efficient hybrid CNN-transformer model for self-supervised monocular depth estimation on mobile devices. In *Proceedings of International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 642–647. IEEE, 2023.

[29] Ilya Makarov, Alisa Korinevskaya, and Vladimir Aliev. Sparse depth map interpolation using deep convolutional neural networks. In *Proceedings of the 41st International Conference on Telecommunications and Signal Processing (TSP)*, pages 1–5. IEEE, 2018.

[30] Ilya Makarov, Dmitrii Maslov, Olga Gerasimova, Vladimir Aliev, Alisa Korinevskaya, Ujjwal Sharma, and Haoliang Wang. On reproducing semi-dense depth map reconstruction using deep convolutional neural networks with perceptual loss. In *Proceedings of the 27th ACM International Conference on Multimedia (ACMMM)*, pages 1080–1084, 2019.

[31] Ilya Makarov, Andrey Savchenko, Arseny Korovko, Leonid Sherstyuk, Nikita Severin, Dmitrii Kiselev, Aleksandr Mikheev, and Dmitrii Babaev. Temporal network embedding framework with causal anonymous walks representations. *PeerJ Computer Science*, 8:e858, 2022.

[32] Sachin Mehta and Mohammad Rastegari. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[33] Seongjae Min, Junseok Yang, Sangjun Lim, Junyong Lee, Sangwon Lee, and Sejoon Lim. Emotion recognition using transformers with masked learning. *arXiv preprint arXiv:2403.13731*, 2024.

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

[35] Bach Nguyen-Xuan, Thien Nguyen-Hoang, and Nhu Tai-Do. Emotic masked autoencoder with attention fusion for facial expression recognition. *arXiv preprint arXiv:2403.13039*, 2024.

[36] Elena Ryumina, Maxim Markitantov, Dmitry Ryumin, Heysem Kaya, and Alexey Karpov. Audio-visual compound expression recognition method based on late modality fusion and rule-based decision. *arXiv preprint arXiv:2403.12687*, 2024.

[37] Andrey Savchenko. Facial expression recognition with adaptive frame rate based on multiple testing correction. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 30119–30129. PMLR, 2023.

[38] Andrey V. Savchenko. Video-based frame-level facial analysis of affective behavior on mobile devices using EfficientNets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2359–2366, 2022.

[39] Andrey V Savchenko. EmotiEffNets for facial processing in video-based valence-arousal prediction, expression classification and action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5715–5723, 2023.

[40] Andrey V. Savchenko. MT-EmotiEffNet for multi-task human affective behavior analysis and learning from synthetic data. In *Proceedings of the European Conference on Computer Vision (ECCV 2022) Workshops*, pages 45–59. Springer, 2023.

[41] Andrey V Savchenko. HSEmotion team at the 6th ABAW competition: Facial expressions, valence-arousal and emotion intensity prediction. *arXiv preprint arXiv:2403.11590*, 2024.

[42] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13 (4):2132–2143, 2022.

[43] Vladimir V Savchenko and Andrey V Savchenko. Criterion of significance level for selection of order of spectral estimation of entropy maximum. *Radioelectronics and Communications Systems*, 62(5):223–231, 2019.

[44] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 6105–6114. PMLR, 2019.

[45] Anton Vakhrushev, Alexander Ryzhkov, Maxim Savchenko, Dmitry Simakov, Rinchin Damdinov, and Alexander Tuzhilin. LightAutoML: Automl solution for a large financial services ecosystem. *arXiv preprint arXiv:2109.01528*, 2021.

[46] Ngoc Tu Vu, Van Thong Huynh, Trong Nghia Nguyen, and Soo-Hyung Kim. Ensemble spatial and temporal vision transformer for action units detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5770–5776, 2023.

[47] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[48] Paul Waligora, Osama Zeeshan, Haseeb Aslam, Soufiane Belharbi, Alessandro Lameiras Koerich, Marco Pedersoli, Simon Bacon, and Eric Granger. Joint multimodal transformer for dimensional emotional recognition in the wild. *arXiv preprint arXiv:2403.10488*, 2024.

[49] 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. *arXiv preprint arXiv:2303.09145*, 2023.

[50] Zihan Wang, Siyang Song, Cheng Luo, Yuzhi Zhou, Shiling Wu, Weicheng Xie, and Linlin Shen. Spatial-temporal graph-based AU relationship learning for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5899–5907, 2023.

[51] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech

processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.

[52] Jun Yu, Renda Li, Zhongpeng Cai, Gongpeng Zhao, Guochen Xie, Jichao Zhu, Wangyuan Zhu, Qiang Ling, Lei Wang, Cong Wang, Luyu Qiu, and Wei Zheng. Local region perception and relationship learning combined with feature fusion for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5785–5792, 2023.

[53] Jun Yu, Zhihong Wei, and Zhongpeng Cai. Exploring facial expression recognition through semi-supervised pretraining and temporal modeling. *arXiv preprint arXiv:2403.11942*, 2024.

[54] Jun Yu, Jichao Zhu, and Wangyuan Zhu. Compound expression recognition via multi model ensemble. *arXiv preprint arXiv:2403.12572*, 2024.

[55] 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 Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1980–1987. IEEE, 2017.

[56] Saining Zhang, Yuhang Zhang, Ye Zhang, Yufei Wang, and Zhigang Song. A dual-direction attention mixed feature network for facial expression recognition. *Electronics*, 12(17): 3595, 2023.

[57] Su Zhang, Ziyuan Zhao, and Cuntai Guan. Multimodal continuous emotion recognition: A technical report for ABAW5. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5764–5769, 2023.

[58] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Multimodal facial affective analysis based on masked autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5793–5802, 2023.

[59] Wei Zhang, Feng Qiu, Chen Liu, Lincheng Li, Heming Du, Tiancheng Guo, and Xin Yu. Affective behaviour analysis via integrating multi-modal knowledge. *arXiv preprint arXiv:2403.10825*, 2024.

[60] Ziyang Zhang, Liuwei An, Zishun Cui, Ao Xu, Tengteng Dong, Yueqi Jiang, Jingyi Shi, Xin Liu, Xiao Sun, and Meng Wang. ABAW5 challenge: A facial affect recognition approach utilizing transformer encoder and audiovisual fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5725–5734, 2023.

[61] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Leveraging TCN and transformer for effective visual-audio fusion in continuous emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5756–5763, 2023.

[62] Weiwei Zhou, Jiada Lu, Chenkun Ling, Weifeng Wang, and Shaowei Liu. Boosting continuous emotion recognition with self-pretraining using masked autoencoders, temporal convolutional networks, and transformers. *arXiv preprint arXiv:2403.11440*, 2024.