EmotiEffNets for Facial Processing in Video-based Valence-Arousal Prediction, Expression Classification 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

Abstract

In this article, the pre-trained convolutional networks from the EmotiEffNet family for frame-level feature extraction are used for downstream emotion analysis tasks from the fifth Affective Behavior Analysis in-the-wild (ABAW) competition. In particular, we propose an ensemble of a multi-layered perceptron and the LightAutoML-based classifier. The post-processing by smoothing the results for sequential frames is implemented. Experimental results for the large-scale Aff-Wild2 database demonstrate that our model is much better than the baseline facial processing using VGGFace And ResNet. For example, our macro-averaged F1-scores of facial expression recognition and action unit detection on the testing set are 11-13% greater. Moreover, the concordance correlation coefficients for valence/arousal estimation are up to 30% higher when compared to the baseline.

1. Introduction

The affective behavior analysis in-the-wild (ABAW) problem is an essential part of many intelligent systems with human-computer interaction [8, 9]. It can be used in online learning to recognize student satisfaction and engagement [28], understand users' reactions to advertisements, analyze online event participants' emotions, video surveillance [29], etc. Despite significant progress in deep learning in image understanding, video-based prediction of human emotions is still a challenging task due to the absence of large emotional datasets without dirty/uncertain labels.

To speed up progress in this area, a sequence of ABAW workshops and challenges has been launched [5, 7, 14]. They introduced several tasks of human emotion understanding based on large-scale AffWild [11, 36] and AffWild2 [12, 13] datasets. The recent ABAW-5 competition [10] contains an extended version of the Aff-Wild2 database for three uni-task challenges, namely, (1) prediction of two continuous affect dimensions, namely, valence and arousal (VA); (2) facial expression recognition (FER); and (3) detection of action units (AU), i.e., atomic facial muscle actions. Refining the model by using only annotations for a given task is strictly required, i.e., the multi-task learning on the VA, FER, and AU labels of the AffWild2 dataset is not allowed. As emotions can rapidly change over time, frame-level predictions are required.

The above-mentioned tasks have been studied in the third ABAW challenge [6]. Hence, there exist several promising solutions for its participants. The baseline for VA prediction is a ResNet-50 pre-trained on ImageNet with a (linear) output layer that gives final estimates for valence and arousal [10]. Much better results on validation and test sets were achieved by EfficientNet-B0 [24] pre-trained on AffectNet [16] from HSEmotion library [25]. An ensemble approach with the Gated Recurrent Unit (GRU) and Transformer [4] combined using Regular Networks (RegNet) [19] let the team PRL take the third place for this task. The runner-up was the FlyingPigs team that proposed a cross-modal co-attention model for continuous emotion recognition using visual-audio-linguistic information based on ResNet-50 for spatial encoding and a temporal convolutional network (TCN) for temporal encoding [37]. Finally, the winning solution of the Situ-RUCAIM3 team utilized two types of encoders to capture the temporal context information in the video (Transformer and LSTM) [15].

The baseline for the FER task is a VGG16 network with fixed convolutional weights, pre-trained on the VGGFACE dataset [10]. The second place was taken by an ensemble of multi-head cross-attention networks (Distract your Attention Network, DAN) from the IXLAB team [2]. A unified transformer-based multimodal framework for AU detection and FER that uses InceptionResNet visual features

and DLN-based audio features [40] took first place in this competition.

Similarly to FER, the baseline for AU detection is a VGGFACE network [10]. A visual spatial-temporal transformer-based model and a convolution-based audio model to extract action unit-specific features were proposed by the STAR-2022 team [32]. An above-mentioned ensemble approach of GRU and Transformer with RegNets [19] took the third place. Slightly better results were achieved by the IResnet100 network of the SituTech team that utilized feature pyramid networks and single-stage headless [3]. The winner is again the InceptionResNet-based audiovisual ensemble of the Netease Fuxi Virtual Human team [40].

The fifth edition of the ABAW challenge [10] significantly improved performance for all these challenges. For example, the team CBCR that took the third place in the VA task fine-tuned the pre-trained VGG model using logmel-spectrogram for the audio part of the visual-audio-linguistic pipeline [38] instead of the previous attempts with VGGfish [37]. Excellent results are obtained by the CtyunAI team using Efficientnet-b2 [28] for EXPR (third place) and AU (6th place) competitions, while a large ensemble of many models worked better for VA estimation (4th place) [42]. The winner of FER and AU competitions (Netease Fuxi AI Lab) used the masked autoencoder and multimodal ensemble [39].

In this paper, we propose a novel pipeline suitable for all three tasks of ABAW in the video. The unified representation of a facial emotion state is extracted by a pre-trained lightweight EmotiEffNet model [22]. These convolutional neural networks (CNN) are tuned on external AffectNet dataset [16], so the facial embeddings extracted by this neural network do not learn any features that are specific to the Aff-Wild2 dataset [12, 13]. Several blending ensembles are studied based on combining embeddings [30] and logits at the output of these models for each video frame [24, 25]. In addition to MLP (multi-layer perceptron), we examine classifiers from the LightAutoML (LAMA) framework [31].

The remaining part of the paper is organized as follows. The proposed workflow is presented in Section 2. Its experimental study for three tasks from the fifth ABAW challenge is provided in Section 3. Finally, Section 4 contains the conclusion and discussion of future studies.

2. Proposed approach

In this Section, the novel workflow for emotion recognition in video is introduced (Fig. 1). At first, the faces are detected with an arbitrary technique, and the representations of affective behavior are extracted from each face by using EfficientNet CNN from HSEmotion library [25], such as EmotiEffNet-B0 [24] or the winner of one of the tasks from ABAW-4, namely, MT-EmotiEffNet-B0 [25]. These models were trained for face identification on the VGGFace2

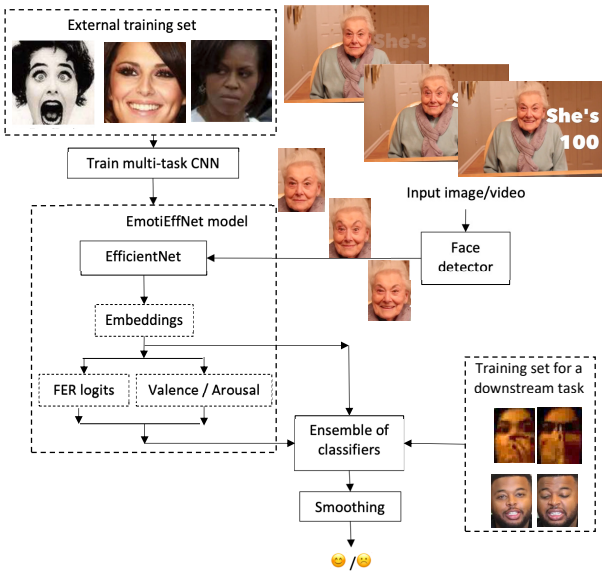

Figure 1. Proposed workflow for the video-based facial emotion analysis.

dataset. Next, they were fine-tuned to recognize facial expression and, in case of multi-task MT-EmotiEffNet model, predict valence/arousal from a static photo by using the AffectNet dataset [16].

For simplicity, let us assume that every t-th frame of the video contains a single facial image $X(t)$, where $t \in \{1, 2, ..., T\}$ and T is the total number of frames [26]. These images are resized and fed into the EmotiEffNet PyTorch models to obtain D-dimensional embeddings $\mathbf{x}(t)$ [30], eight-dimensional logits for 8 facial expressions $\mathbf{l}(t)$ from AffectNet (Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness, Surprise) and valence $V(t) \in [-1; 1]$ (how positive/negative a person is) and arousal $A(t) \in [-1; 1]$ (how active/passive a person is) [10].

Next, these facial representations are used to solve an arbitrary downstream task. In this paper, we examine three problems from the ABAW-5 competition, namely (1) VA prediction (multi-output regression); (2) FER (multi-class classification); and (3) AU detection (multi-class multi-label classification).

The supervised learning case is assumed where a training set of $N > 1$ pairs $(X_n, y_n), n = 1, 2, ...N$ is available. Here, a facial image X_n from the video frame and associated with corresponding labels y_n. Here are the details about each task:

1. VA estimation data contains a training set with 356 videos and 1653757 frames, and a validation set with 76 videos and 376323 frames

2. Training and validation sets for FER contain 248 videos (585317 frames) and 70 videos (280532

frames), respectively. Each frame is associated with one of eight imbalanced classes (Neutral, Anger, Disgust, Fear, Happiness, Sadness, Surprise, or Other).

3. AU detection challenge consists of a training set (295 videos, 1356694 frames) and a validation set (105 videos, 445836 frames) with 12 highly-imbalanced labels: AU 1 (inner brow raiser, AU 2 (outer brow raiser), AU 4 (brow lowerer), AU 6 (cheek raiser), AU 7 (lid tightener), AU 10 (upper lip raiser), AU 12 (lip corner puller), AU 15 (lip corner depressor), AU 23 (lip tightener), AU 24 (lip pressor), AU 25 (lips part), and AU 26 (jaw drop) [10].

At first, every training example X_n is fed into the same CNN to obtain embeddings $\mathbf{x}_n$ [24, 30], FER logits $\mathbf{l}_n$ and valence/arousal V_n, A_n. In this paper, the following classifiers are trained: MLP and ensemble models trained via the LAMA library [31]. The latter tries to find the best pre-processing, classifiers, and their ensembles, and post-processing for an arbitrary classification or regression task. Due to computational complexity and poor metrics obtained after 10 minutes of AutoML search, we do not process embeddings $\mathbf{x}(t)$ here. The input of LAMA is a concatenation of logits $\mathbf{l}(t)$, valence $V(t)$, and arousal $A(t)$ at the output of the last layer of EmotiEffNets.

The MLP is trained with the TensorFlow 2 framework similarly to [24]. VA is better predicted using only logits and valence/arousal by an MLP without a hidden layer and two outputs with $tanh$ activation functions trained to maximize the mean estimate of the Concordance Correlation Coefficient (CCC) for valence CCC_V and arousal CCC_A.

FER and AU detection are solved similarly by feeding embeddings or logits into the MLP with one hidden layer. In the former case, eight outputs with softmax activations were added, and the weighted sparse categorical cross-entropy is used to fit the classifier. In addition, we examined the possibility to fine-tune the whole EmotiEffNet CNN on the training set of this challenge using PyTorch source code from the HSEmotion library.

The output layer for the AU detection task contains 12 units with sigmoid activation functions, and the weighted binary cross-entropy loss was optimized. The final prediction is made by matching the outputs with predefined thresholds. It is possible to either set a fixed threshold (0.5 for each unit) or choose the best threshold for each action unit to maximize F1-score on a validation set. The classifier predicts the class label that corresponds to the maximal output of the softmax layer.

In all tasks, it is possible to build a simple blending decision rule to combine several classifiers (LightAutoML, MLP, fine-tuned model) and input features (embeddings or logits from pre-trained model). Moreover, pre-trained models were used to make predictions for VA and FER tasks.

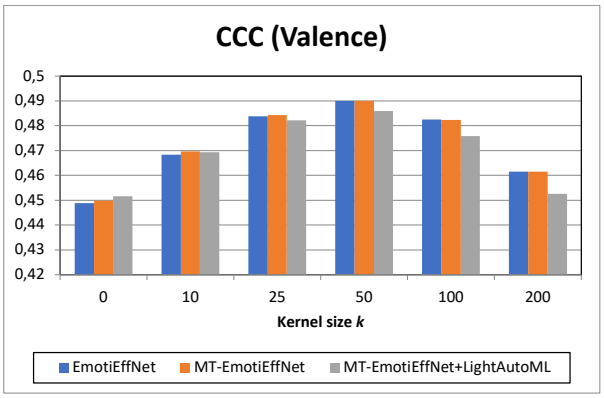

Figure 2. Dependence of CCC for valence prediction on the smoothing kernel size k.

In the former case, the valence $V(t)$ and arousal $A(t)$ predicted by MT-EmotiEffNet [25] were directly used to make a final decision (hereinafter "pre-trained VA only"). In the second case due to the difference in classes, namely, absence of contempt emotion and the presence of the state "Other" in the AffWild2 dataset, we preliminarily apply the MLP classifier to make a binary decision (Other/non-Other). If the predicted class label is not equal to Other, predictions of the pre-trained model from other 7 basic facial expressions are used, i.e., the label that corresponds to the maximal logit (hereinafter "pre-trained logits").

The final decision in the pipeline (Fig. 1) is made by smoothing predictions (class probabilities for classification tasks and predicted valence/arousal for regression problem) [26] for individual frames by using the box filter with kernel size $2k + 1$. Here k is a hyperparameter chosen to maximize performance metrics on the validation set. In fact, we compute the average predictions for the current frame, previous k frames, and next k frames [24].

3. Experimental results

Let us discuss the results of our workflow (Fig. 1) for three tasks from the fifth ABAW challenge [10]. The training source code to reproduce the experiments for the presented approach is publicly available[1].

3.1. Valence-Arousal Prediction

A comparison of our workflow based on EmotiEffNet-B0 and MT-EmotiEffNet-B0 [25] with previous results on the Valence-Arousal estimation challenge is presented in Table 1 and Table 2[2], respectively. A special remark "(train + val)" is added in the latter table, if the regression model

[1] https://github.com/HSE-asavchenko/face-emotion-recognition/tree/main/src/ABAW

[2] Refer https://ibug.doc.ic.ac.uk/resources/cvpr-2023-5th-abaw/ for *

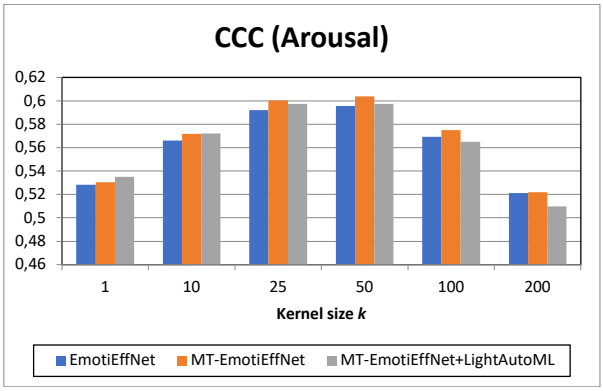

Figure 3. Dependence of CCC for arousal prediction on the smoothing kernel size k.

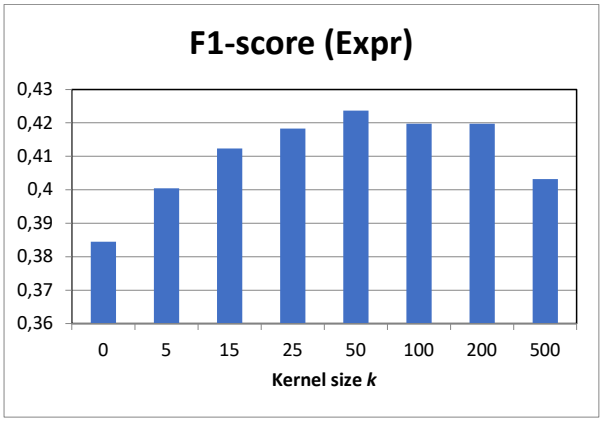

Figure 4. Dependence of F1-score for FER on the smoothing kernel size k.

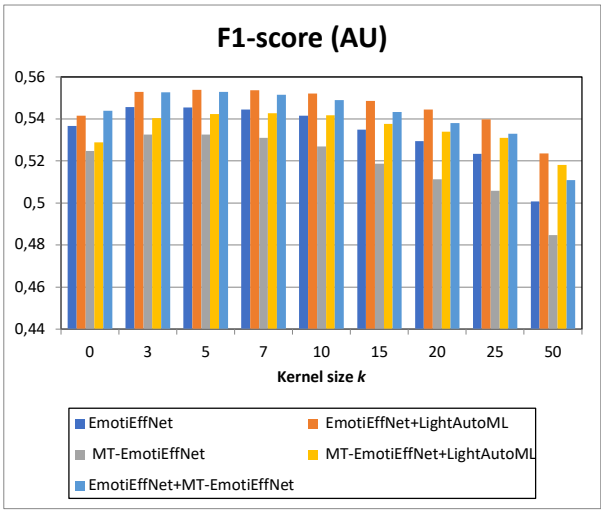

Figure 5. Dependence of F1-score for AU detection on the smoothing kernel size k.

was trained on the concatenation of official training and validation sets. Otherwise, only the training set was utilized. We use official performance metrics from the organizers: CCC for valence, arousal, and their average value $P_{VA} = (CCC_V + CCC_A)/2$. The value of "Is ensemble?" is set to "Yes" for an ensemble of neural networks and "No" for a single model.

Here, we significantly improved the results of the baseline ResNet-50 [10]: our CCC is higher up to 0.19 and 0.53 for valence and arousal, respectively. Moreover, our best ensemble model is characterized by 0.05 greater CCC_V and 0.07 greater CCC_A when compared to the best previous usage of EmotiEffNet models [24]. The LightAutoML classifier is worse than simple MLP, but their blending achieves one of the top results.

One of the most valuable hyperparameters in our pipeline is the kernel size k of the median filter. The dependence of validation CCCs on k for valence and arousal is shown in Fig. 2 and Fig. 3, respectively. As one can notice, the highest performance is reached by rather high values of k (25...50), i.e., 51...101 predictions should be averaged for each frame.

3.2. Facial Expression Recognition

The macro-averaged F1-scores P_{EXPR} and classification accuracy for various FER techniques are shown in Table 3 and Table 4. The value of P_{EXPR} depending on the kernel size k for our several best classifiers is presented in Fig. 4.

In addition to cropped faces provided by the organizers of this challenge, we used here small (112x112) cropped_aligned photos. As one can notice, the quality of FER on the former is much better, so we do not use aligned faces in other experiments. The proposed approach makes it possible to increase F1-score on 20% and 3% when compared to the baseline VGGFACE [10] and the previous usage of EfficientNet models [24]. Again, a large kernel size (50...200) of the mean filter is required to provide the best-possible F1-score (Fig. 4).

Surprisingly, MT-EmotiEffNet [25] is up to 5% worse than EmotiEffNet, though the former was significantly more accurate on the multi-task learning challenge from ABAW-4 competition [5]. It is important to emphasize that the F1-score of our best ensemble (43.3%) is approximately equal to the best single model (43.2%), though the difference in accuracy is significant (55.7% vs 54.6%). Nevertheless, we achieved the greatest validation F1-score, which is 4% higher than the F1-score of the ABAW-3 winner team (Netease Fuxi Virtual Human) [40].

3.3. Action Unit Detection

In the last Subsection, macro-averaged F1-score P_{AU} is estimated for the multi-label classification of action units.

Method	Modality	Is ensemble?	CCC_V	CCC_A	P_{VA}
Baseline ResNet-50 [10]	Faces	No	0.31	0.17	0.24
EfficientNet-B0 [24]	Faces	No	0.449	0.535	0.492
GRU + Attention [19]	Video	Yes	0.437	0.576	0.507
Resnet-50/TCN [37]	Audio/video	Yes	0.450	0.651	0.551
Transformer [15]	Audio/video	Yes	0.588	0.669	0.627
Resnet50/Regnet/EfficientNet [33]	Faces	Yes	0.257	0.383	0.320
Channel Attention Network [38]	Audio/video	Yes	0.423	0.670	0.547
Masked Autoencoder [39]	Audio/video	Yes	0.476	0.644	0.560
Transformer [41]	Audio/video	Yes	0.554	0.659	0.607
TCN [42]	Audio/video	Yes	0.550	0.681	0.615
MT-EmotiEffNet (logits), LightAutoML	Faces	No	0.373	0.433	0.403
MT-EmotiEffNet (logits), MLP	Faces	No	0.444	0.521	0.483
MT-EmotiEffNet (VA only)	Faces	No	0.404	0.248	0.326
MT-EmotiEffNet (logits), MLP + LightAutoML	Faces	Yes	0.447	0.526	0.487
MT-EmotiEffNet (logits), MLP, smoothing	Faces	No	0.490	0.604	0.547
MT-EmotiEffNet (logits), MLP + LightAutoML, smoothing	Faces	Yes	0.486	0.597	0.542
EmotiEffNet (logits), LightAutoML	Faces	No	0.369	0.431	0.400
EmotiEffNet (logits), MLP	Faces	No	0.443	0.519	0.482
EmotiEffNet (logits), MLP, smoothing	Faces	No	0.490	0.596	0.543
EmotiEffNet + MT-EmotiEffNet (logits), MLP	Faces	Yes	0.450	0.530	0.490
EmotiEffNet + MT-EmotiEffNet (logits), MLP, smoothing	Faces	Yes	0.494	0.607	0.550

Table 1. Valence-Arousal Challenge Results on the Aff-Wild2's validation set.

Method	Modality	Is ensemble?	CCC_V	CCC_A	P_{VA}
SituTech*	Audio/video	Yes	0.6193	0.6634	0.6414
Masked Autoencoder [39]	Audio/video	Yes	0.6486	0.6258	0.6372
Channel Attention Network [38]	Audio/video	Yes	0.5526	0.6299	0.5913
TCN [42]	Audio/video	Yes	0.5008	0.6325	0.5666
Transformer [41]	Audio/video	Yes	0.5234	0.5451	0.5342
Regnet/Video Vision Transformer [17]	Faces	No	0.5043	0.4279	0.4661
Transformer [18]	Faces	No	0.4703	0.4578	0.4640
EfficientNet-B0 [24]	Faces	No	0.4174	0.4538	0.4356
Resnet50/Regnet/EfficientNet [33]	Faces	Yes	0.3245	0.2321	0.2783
Baseline ResNet-50 [10]	Faces	No	0.211	0.191	0.201
MT-EmotiEffNet (logits), MLP (train + val), smoothing	Faces	No	0.4818	0.5279	0.5048
MT-EmotiEffNet (logits), MLP, smoothing	Faces	No	0.4771	0.5263	0.5017
EmotiEffNet + MT-EmotiEffNet (logits), MLP, smoothing	Faces	Yes	0.4788	0.5227	0.5007
MT-EmotiEffNet (logits), MLP + LightAutoML, smoothing	Faces	Yes	0.4748	0.5174	0.4961
EmotiEffNet (logits), MLP, smoothing	Faces	No	0.4704	0.5059	0.4882

Table 2. Valence-Arousal Challenge Results on the ABAW-5 test set.

The estimates of performance metrics on the validation set are shown in Table 5 and Fig. 5. The results for the test set are presented in Table 6. In contrast to previous experiments, the kernel size k should be much lower (3...5) to achieve the best performance. Indeed, at least one action unit is rapidly changed in typical scenarios.

The LightAutoML ensemble is again slightly worse than a simple MLP, but their blending leads to excellent results. Our best model is 16% better than the baseline, though we increase the F1-score of EmotiEffNet compared to its previous usage [24] on 1%. However, only the second-place winner team (SituTech) has a higher F1 score on the validation set. Finally, the choice of thresholds can definitely improve the AU detection quality, though it is possible that

Method	Modality	Is ensemble?	F1-score P_{EXPR}	Accuracy
Baseline VGGFACE [10]	Faces	No	0.23	-
RegNetY [20]	Faces	Yes	0.304	-
EfficientNet-B0 [24]	Faces	No	0.402	-
DAN (ResNet50) [2]	Faces	Yes	0.346	-
InceptionResNet [40]	Audio/video	Yes	0.394	-
Meta-Classifier [33]	Faces	Yes	0.302	0.462
TCN [42]	Audio/video	Yes	0.377	-
Transformer [41]	Audio/video	Yes	0.406	-
Masked Autoencoder [39]	Audio/video	Yes	0.495	-
MT-EmotiEffNet (embeddings), aligned faces	Faces	No	0.293	0.403
MT-EmotiEffNet (embeddings), cropped faces	Faces	No	0.336	0.447
EmotiEffNet (embeddings), aligned faces	Faces	No	0.304	0.474
EmotiEffNet (embeddings), cropped faces	Faces	No	0.384	0.495
EmotiEffNet (embeddings), smoothing	Faces	No	0.432	0.546
EmotiEffNet (logits)	Faces	No	0.327	0.426
EmotiEffNet (fine-tuned), cropped faces	Faces	No	0.380	0.484
EmotiEffNet (embeddings + logits), frame-level	Faces	Yes	0.396	0.502
EmotiEffNet (embeddings + logits), smoothing	Faces	Yes	0.431	0.546
EmotiEffNet (pre-trained + fine-tuned), frame-level	Faces	Yes	0.405	0.524
EmotiEffNet (pre-trained + fine-tuned), smoothing	Faces	Yes	0.433	0.557

Table 3. Expression Challenge Results on the Aff-Wild2's validation set.

Method	Modality	Is ensemble?	F1-score P_{EXPR}
Masked Autoencoder [39]	Audio/video	Yes	0.4121
SituTech*	Audio/video	Yes	0.4072
TCN [42]	Audio/video	Yes	0.3532
Transformer [41]	Audio/video	Yes	0.3337
IResnet100 [35]	Faces	Yes	0.3075
Noise aware model [1]	Faces	No	0.3047
EfficientNet-B0 [24]	Faces	No	0.3025
Transformer [18]	Faces	No	0.2949
Baseline VGGFACE [10]	Faces	No	0.2050
EmotiEffNet (embeddings), MLP (train+val), smoothing	Faces	No	0.3292
EmotiEffNet (pre-trained + fine-tuned), smoothing	Faces	Yes	0.3286
EmotiEffNet (embeddings + logits), smoothing	Faces	Yes	0.3171
EmotiEffNet (embeddings), MLP, smoothing	Faces	No	0.3058
EmotiEffNet (fine-tuned)	Faces	No	0.2862

Table 4. Expression Challenge Results on the ABAW-5 test set.

the current choice of a threshold for each action unit using the validation set is not optimal.

4. Conclusion and future works

We proposed the video-based emotion recognition pipeline (Fig. 1) suitable for a wide range of affective behavior analysis downstream tasks. It exploits the pre-trained EmotiEffNet models to extract representative emotional features from each facial frame. Experiments on datasets from the fifth ABAW challenge showed the benefits of our workflow when compared to the baseline CNNs [10] and previous application of EfficientNet models [24]. Though our results are worse when compared to the multimodal ensemble of the winning team (Netease Fuxi AI Lab) [39], our workflow (Fig. 1) is one of the best among all participants if only facial modality is analyzed.

Thus, it is possible to significantly improve the quality

Method	Modality	Is ensemble?	F1-score P_{AU}
Baseline VGGFACE [10]	Faces	No	0.39
InceptionResNet [40]	Audio/video	Yes	0.525
Transformer [32]	Audio/video	Yes	0.523
GRU + Attention [19]	Video	Yes	0.544
EfficientNet-B0 [24]	Faces	No	0.548
IResNet [3]	Faces	Yes	0.735
IResnet100 [35]	Faces	Yes	0.511
TCN [42]	Audio/video	Yes	0.517
Transformer [41]	Audio/video	Yes	0.530
Regnet/Video Vision Transformer [17]	Faces	No	0.540
Masked Autoencoder graph representations [34]	Faces	Yes	0.543
Masked Autoencoder [39]	Audio/video	Yes	0.567
Regnet [33]	Faces	Yes	0.698
MT-EmotiEffNet, LightAutoML	Faces	No	0.472
MT-EmotiEffNet, MLP	Faces	No	0.525
MT-EmotiEffNet, MLP, smoothing	Faces	No	0.533
MT-EmotiEffNet, MLP + LightAutoML	Faces	Yes	0.533
MT-EmotiEffNet, MLP + LightAutoML, smoothing	Faces	Yes	0.543
EmotiEffNet, LightAutoML	Faces	No	0.477
EmotiEffNet, MLP	Faces	No	0.537
EmotiEffNet, MLP, smoothing	Faces	No	0.545
EmotiEffNet, MLP + LightAutoML	Faces	Yes	0.542
EmotiEffNet, MLP + LightAutoML, smoothing	Faces	Yes	0.554
EmotiEffNet + MT-EmotiEffNet	Faces	Yes	0.544
EmotiEffNet + MT-EmotiEffNet, smoothing	Faces	Yes	0.553

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

Method	Modality	Is ensemble?	F1-score P_{AU}
Masked Autoencoder [39]	Audio/video	Yes	0.5549
SituTech*	Audio/video	Yes	0.5422
IResnet100 [35]	Faces	Yes	0.5144
Masked Autoencoder graph representations [34]	Faces	Yes	0.5128
Regnet/Video Vision Transformer [17]	Faces	No	0.5101
TCN [42]	Audio/video	Yes	0.4887
Regnet [33]	Faces	Yes	0.4811
EfficientNet-B0 [24]	Faces	No	0.4731
Transformer [41]	Audio/video	Yes	0.4752
Transformer [18]	Faces	No	0.4563
Baseline VGGFACE [10]	Faces	No	0.365
EmotiEffNet, MLP + LightAutoML, smoothing	Faces	Yes	0.4878
EmotiEffNet + MT-EmotiEffNet, smoothing	Faces	Yes	0.4821
MT-EmotiEffNet, MLP, smoothing	Faces	No	0.4786
EmotiEffNet, MLP, smoothing	Faces	No	0.4722
EmotiEffNet, MLP (train + val), smoothing	Faces	No	0.4687

Table 6. Action Unit Challenge Results on the ABAW-5 test set.

metrics by combining it with audio processing [21, 27, 40] and/or temporal models [15, 37]. Another direction for future research is an increase in decision-making speed by using sequential inference and processing of video

frames [23]. Finally, solutions of the winners [38, 39] proved that it is necessary to use other options to perform the train-test split of the AffWild2 dataset, so it is important to borrow their ideas to increase the overall performance metrics.

Acknowledgements. The work is supported by RSF (Russian Science Foundation) grant 20-71-10010.

References

[1] Darshan Gera, Badveeti Naveen Siva Kumar, Bobbili Veerendra Raj Kumar, and S Balasubramanian. ABAW: Facial expression recognition in the wild. *arXiv preprint arXiv:2303.09785*, 2023. 6

[2] Jae-Yeop Jeong, Yeong-Gi Hong, Daun Kim, Jin-Woo Jeong, Yuchul Jung, and Sang-Ho Kim. Classification of facial expression in-the-wild based on ensemble of multi-head cross attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2353–2358, 2022. 1, 6

[3] 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, 2022. 2, 7

[4] Aleksei Karpov and Ilya Makarov. Exploring efficiency of vision transformers for self-supervised monocular depth estimation. In *Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 711–719. IEEE, 2022. 1

[5] Dimitrios Kollias. ABAW: Learning from synthetic data & multi-task learning challenges. *arXiv preprint arXiv:2207.01138*, 2022. 1, 4

[6] 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) Workshops*, pages 2328–2336, 2022. 1

[7] 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. 1

[8] 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

[9] 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

[10] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. ABAW: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. *arXiv preprint arXiv:2303.01498*, 2023. 1, 2, 3, 4, 5, 6, 7

[11] 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

[12] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-Wild2, multi-task learning and ArcFace. *arXiv preprint arXiv:1910.04855*, 2019. 1, 2

[13] 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, 2

[14] 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. 1

[15] 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, 2022. 1, 5, 7

[16] 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. 1, 2

[17] Vu Ngoc Tu, Van Thong Huynh, and Soo-Hyung Kim. Vision transformer for action units detection. *arXiv preprint arXiv:2303.09917*, 2023. 5, 7

[18] 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. 5, 6, 7

[19] Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. An ensemble approach for facial behavior analysis in-the-wild video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2512–2517, 2022. 1, 2, 5, 7

[20] Kim Ngan Phan, Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. Facial expression classification using fusion of deep neural network in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2507–2511, 2022. 6

[21] Andrey V Savchenko. Phonetic words decoding software in the problem of Russian speech recognition. *Automation and Remote Control*, 74:1225–1232, 2013. 7

[22] Andrey V Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *Proceedings of International Symposium on Intelligent Systems and Informatics (SISY)*, pages 119–124. IEEE, 2021. 2

[23] Andrey V Savchenko. Fast inference in convolutional neural networks based on sequential three-way decisions. *Information Sciences*, 560:370–385, 2021. 7

[24] 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. 1, 2, 3, 4, 5, 6, 7

[25] 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) Workshops*, pages 45–59. Springer, 2023. 1, 2, 3, 4

[26] Andrey V Savchenko and Natalya S Belova. Statistical testing of segment homogeneity in classification of piecewise–regular objects. *International Journal of Applied Mathematics and Computer Science*, 25(4):915–925, 2015. 2, 3

[27] Andrey V Savchenko and Liudmila V Savchenko. Towards the creation of reliable voice control system based on a fuzzy approach. *Pattern Recognition Letters*, 65:145–151, 2015. 7

[28] 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. 1, 2

[29] Anastasiia D Sokolova, Angelina S Kharchevnikova, and Andrey V Savchenko. Organizing multimedia data in video surveillance systems based on face verification with convolutional neural networks. In *Proceedings of Analysis of Images, Social Networks and Texts (AIST)*, pages 223–230. Springer, 2018. 1

[30] Boris Tseytlin and Ilya Makarov. Hotel recognition via latent image embeddings. In *Proceedings of the 16th International Work-Conference on Artificial Neural Networks (IWANN), Part II 16*, pages 293–305. Springer, 2021. 2, 3

[31] 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. 2, 3

[32] Lingfeng Wang, Jin Qi, Jian Cheng, and Kenji Suzuki. Action unit detection by exploiting spatial-temporal and label-wise attention with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2470–2475, 2022. 2, 7

[33] 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. 5, 6, 7

[34] Zihan Wang, Siyang Song, Cheng Luo, Yuzhi Zhou, Shiling Wu, Weicheng Xie, and Linlin Shen. Spatio-temporal AU relational graph representation learning for facial action units detection. *arXiv preprint arXiv:2303.10644*, 2023. 7

[35] Jun Yu, Renda Li, Zhongpeng Cai, Gongpeng Zhao, Guochen Xie, Jichao Zhu, and Wangyuan Zhu. Local region perception and relationship learning combined with feature fusion for facial action unit detection. *arXiv preprint arXiv:2303.08545*, 2023. 6, 7

[36] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1980–1987. IEEE, 2017. 1

[37] Su Zhang, Ruyi An, Yi Ding, and Cuntai Guan. Continuous emotion recognition using visual-audio-linguistic information: A technical report for ABAW3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2376–2381, 2022. 1, 2, 5, 7

[38] Su Zhang, Ziyuan Zhao, and Cuntai Guan. Multimodal continuous emotion recognition: A technical report for ABAW5. *arXiv preprint arXiv:2303.10335*, 2023. 2, 5, 8

[39] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Facial affective analysis based on mae and multi-modal information for 5th ABAW competition. *arXiv preprint arXiv:2303.10849*, 2023. 2, 5, 6, 7, 8

[40] Wei Zhang, Feng Qiu, Suzhen Wang, Hao Zeng, Zhimeng Zhang, Rudong An, Bowen Ma, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2428–2437, 2022. 2, 4, 6, 7

[41] Ziyang Zhang, Liuwei An, Zishun Cui, Tengteng Dong, et al. Facial affect recognition based on transformer encoder and audiovisual fusion for the ABAW5 challenge. *arXiv preprint arXiv:2303.09158*, 2023. 5, 6, 7

[42] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Continuous emotion recognition based on TCN and transformer. *arXiv preprint arXiv:2303.08356*, 2023. 2, 5, 6, 7