Leveraging Pre-trained Multi-task Deep Models for Trustworthy Facial Analysis in Affective Behaviour Analysis in-the-Wild

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

This article presents our results for the sixth Affective Behavior Analysis in-the-wild (ABAW) competition. To improve the trustworthiness of facial analysis, we study the possibility of using pre-trained deep models that extract reliable emotional features without the need to fine-tune the neural networks for a downstream task. In particular, we introduce several lightweight models based on MobileViT, MobileFaceNet, EfficientNet, and DDAMFN architectures trained in multi-task scenarios to recognize facial expressions, valence, and arousal on static photos. These neural networks extract frame-level features fed into a simple classifier, e.g., linear feed-forward neural network, to predict emotion intensity, compound expressions, and valence/arousal. Experimental results for three tasks from the sixth ABAW challenge demonstrate that our approach lets us significantly improve quality metrics on validation sets compared to existing non-ensemble techniques. As a result, our solutions took second place in the compound expression recognition competition.

1. Introduction

The ability to accurately analyze human emotions is crucial for developing human-centered technologies [6, 9, 52]. Contemporary research on affective behavior analysis in unconstrained environments reached a high level of maturity [54]. It is incredibly challenging for the in-the-wild domain, where conventional approaches often struggle due to variations in lighting, pose, and expression intensity [17, 20]. Nowadays, researchers are focused on 3D facial expression analysis [11, 30, 31, 56], multi-modalities [25, 54, 55] and new datasets. One of the well-known benchmarks for evaluating progress in this field is the sequence of Affective Behavior Analysis in-the-wild (ABAW) competitions [14, 18, 21, 23, 24, 29, 51].

One of the traditional tasks in the previous editions of ABAW is the frame-level video-based prediction of valence/arousal (VA). A simple fusion of Resnet50, Regnet, and EfficientNet backbones significantly improves the baseline [46]. Very high quality was obtained by several teams [31, 38, 40, 58] that used EfficientNet-based model [41] pre-trained on the AffectNet dataset [33]. An affine module was proposed [57] to align the features to the same dimension in the audio-visual transformer. The fusion of audio and video modalities using channel attention is examined in [53]. The top-performing team implemented the facial processing with the fine-tuned masked autoencoder (MAE) [54].

Another task introduced in the previous ABAW challenge is the Emotional Mimicry Intensity (EMI) estimation [23], which is the multi-task regression problem that should be solved for the entire video from the Hume-Reaction dataset. The third place was obtained by the above-mentioned MAE [54]. In the paper of the second place winner [48], the spatial attention mechanism and the Mel-Frequency Cepstral Coefficients were used to extract visual and acoustic features, respectively. At the same time, the temporal dynamics were modeled using TCN and a transformer encoder. Finally, the top performance was achieved by studying, analyzing, and combining diverse models and tools to extract multimodal features [25]. However, the organizers of the current ABAW challenge [24] decided to significantly reduce the training and validation sets in this task to make it much more complicated.

The dataset has been significantly modified in the 2024 edition of the EMI competition. The USTC-AC team proposed implementing the late fusion of the dual-channel visual features from ResNet-18 and Wav2Vec2.0 audio features [50]. The pre-trained valence-arousal-dominance module from the Wav2Vec 2.0 let the USTC-

IAT-United team take the second place [10]. Finally, the best-performing team is Netease Fuxi AI Lab with their transformer-based multimodal fusion [55].

One major challenge in emotion understanding is the need for extensive fine-tuning of deep neural networks for specific tasks. For example, the winner of the previous challenge led to excellent results by three times fine-tuning the MAE encoder on the image frames from the Aff-wild2 dataset for each task individually [54]. This procedure leads to obtaining state-of-the-art models for particular datasets but can be computationally expensive and limit the generalizability of models. Thus, the emotion analysis in-the-wild primary goal is to construct single models [19] that are fair, explainable, trustworthy, and privacy-conscious, achieving high performance while enhancing generalization in real-world scenarios.

To encourage the research of this final goal, the sixth ABAW competition introduces the new task for unsupervised Compound Expression (CE) recognition [15] without the labeled training set, so the high-accurate pre-trained models should be utilized in the domain adaptation/self-supervised/zero-shot learning techniques. The audio-visual dynamic model was trained on the concatenation of several video datasets by the SUN_CE team [36]. Promising results were obtained by the USTC-IAT-United team [49] who implemented three expression classification models based on convolutional networks, Vision Transformers, and multiscale local attention networks pre-trained on the union of AffectNet and RAF-DB. Again, audio-visual transformers with the MAE facial features from the Netease Fuxi AI Lab team showed the top results [55].

To achieve the objectives mentioned above, we introduce a novel methodology centered around lightweight models derived from architectures such as MobileViT (Mobile Vision Transformer) [32], MobileFaceNet [4], EfficientNet [43], and DDAMFN (Dual-Direction Attention Mixed Feature Network) [52]. These models are trained in a multi-task scenario, enabling them to predict facial expressions, valence, and arousal from static photographs. By extracting frame-level features, we feed these neural networks into a straightforward classifier, such as a linear feed-forward neural network. Our approach facilitates the prediction of emotion intensity, compound expressions, and valence/arousal, thereby offering a holistic analysis of affective behavior. Our method aims to extract robust emotional features directly applicable to various tasks within the ABAW challenge.

2. Methodology

2.1. Multi-Task Learning of Emotional Descriptors

This paper uses the following approach to train neural networks that extract emotional embeddings [41, 54]. At first, the neural network is pretrained on a face recognition task. In particular, we use the VGGFace2 dataset [3] with 3,067,564 photos of 9131 persons while the validation set contains 243,722 remaining images. During ten epochs, Adam optimized the conventional softmax cross-entropy and sharpness-aware minimization [27]. In contrast to studies of face identification [1], we use the cropped facial regions without any margins and alignment to concentrate on the central part of the face.

Next, we fine-tune the model to recognize emotions on static images from the AffectNet dataset [33] with $C_{Expr} = 8$ emotional classes corresponding to Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness and Surprise, and values of Valence/Arousal from the range between -1 and 1. The official training set contains 287651 manually labeled photos, while the validation set consists of 4000 images (500 per class). We leverage the multi-task (MT) learning [12, 16, 22] and minimize the CCC (Concordance Correlation Coefficient) for valence/arousal and weighted cross-entropy for facial expressions to mitigate the class imbalance in the training set [39]:

$$L\left(X, y_{Expr}, y_V, y_A\right) = 1 - \\ - \log\left(softmax(z_{y_{Expr}}) \cdot \max_{y \in \{1,\dots,C_{Expr}\}} N_y / N_{y_{Expr}}\right) - \\ - 0.5\left(CCC(z_V, y_V) + CCC(z_A, y_A)\right), \quad (1)$$

where X is the facial image, y_V, y_A and $y_{Expr} \in \{1, ..., C_{Expr}\}$ are its valence, arousal and facial expression label, respectively, N_y is the total number of training examples of the y-th expression, z is the logits at the output of last fully connected layer.

Due to privacy issues, it is desirable to implement facial analysis on the edge/mobile device [6, 7, 13]. Thus, it is important to use fast video recognition algorithms [37] and lightweight architectures of neural networks, such as MobileViT [32], MobileFaceNet [4], and DDAMFN [52]. The resulting models are called MT-EmotiMobileViT, MT-EmotiMobileFaceNet, and MT-DDAMFN. We make the weights publicly available. in the repository with our previous EmotiEffNet models [39, 41].

2.2. Video-based Valence-Arousal Estimation

The main idea of this paper is to study the usage of pre-trained deep neural networks without fine-tuning them on every downstream task (Fig. 1). Though such an approach cannot produce state-of-the-art results for a concrete dataset, it reflects the practically essential requirement for an emotion analysis model that can be used in unconstrained environments. Hence, similar feed-forward neural networks are used for the VA estimation task of the ABAW challenge with the Aff-Wild2 dataset [16]. The models are trained using all labeled frames from the training set.

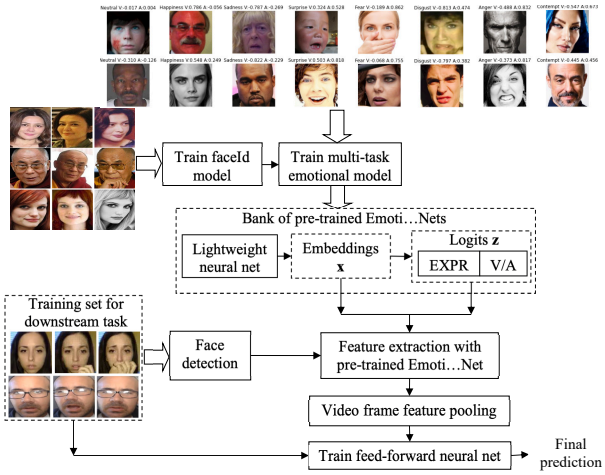

Figure 1. Proposed approach

It was experimentally found that valence and arousal are more accurately predicted by the linear model without hidden layers using the logits $\mathbf{z}$ at the output of the last layer [40]. The Adam optimizer was leveraged in all experiments. As our models typically deal with high-resolution facial images, we processed the folder of cropped faces officially released by the organizer of the 6th ABAW competition. The tanh activations were added to the output layer, and CCC loss was optimized for 20 epochs. The training set contains 1,653,930 cropped faces from 356 videos, while validation is performed on 376,332 other images from 76 videos [24].

As the decision is made in a frame-level manner, some noise may be introduced in the outputs of trained feed-forward neural networks. Thus, we smoothed the predictions for each frame in a short window [40] using a box filter, i.e., the arithmetic mean of predicted scores for 50 frames in a window.

2.3. Compound Expression Recognition

The new task of the 6th ABAW competition is the frame-wise CE recognition on videos from the C-EXPR database [15]. It is required to assign each frame of 56 videos into one of 7 classes, namely, Fearfully_Surprised, Happily_Surprised, Sadly_Surprised, Disgustedly_Surprised, Angrily_Surprised, Sadly_Fearful, Sadly_Angry. It is the most complicated task as no labeled validation set compares different solutions. To somehow choose the candidate submissions, we measured the class balance. The authors of the C-EXPR [15] presented the number of frames for each compound class: 14445, 24915, 10780, 10637, 10535, 10112, and 8878. Hence, we choose the Kullback-Leibler divergence between the frequencies of each class at the output of our models and the frequencies of classes computed from these number of frames. Such an

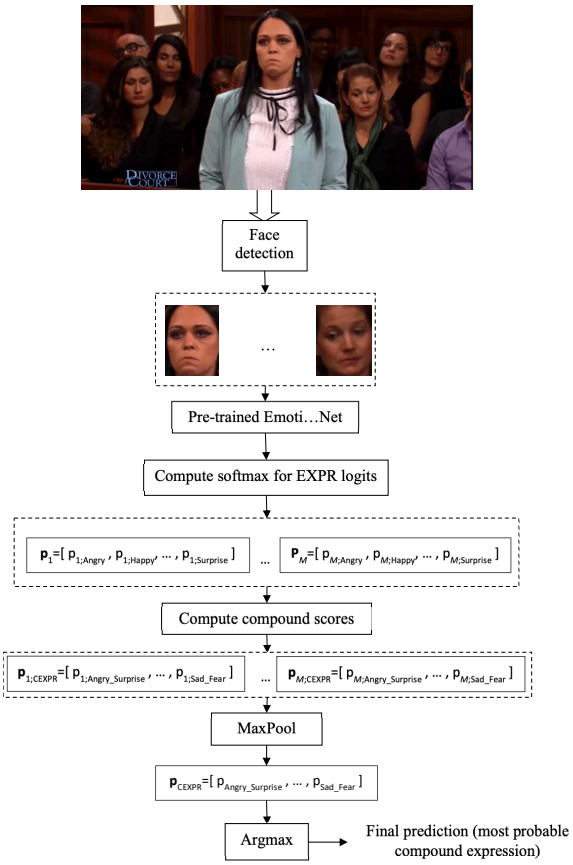

Figure 2. Frame-level compound expression recognition

approach has at least two issues: a) it is a bit of a cheat, in the sense of data leakage (these numbers also include the test set); and b) C-EXPR-DB is way bigger and only a small part was shared for the CE Challenge (thus using the above numbers may have biased the model in the wrong way). However, we decided to compare different models with this imperfect metric.

The proposed pipeline is shown in Fig. 2. Here, the faces from each frame were extracted with the RetinaFace model [8]. If it cannot detect the face on a particular frame, it feeds into the input of the face detector from the Mediapipe framework [28]. As a result, we obtained 22,641 frames with at least one video. Several faces may be detected for a frame, so the total number of detected faces equals 32,329. We analyze all of them with our emotional models. Next, we predict $C_{Expr} = 8$ AffectNet's basic expressions for every detected face, compute probabilities from logits $\mathbf{z}$, and summarize the probability scores for two classes from the compound expressions. Predictions for several faces inside one frame are aggregated with the arithmetic mean. The final prediction is the compound class label corresponding to the maximal summary score.

In addition, we decided to use clustering of embeddings

extracted from video and audio frames. Simple K-means clustering with 7 clusters is utilized. To choose each cluster's label, we compute the average scores of compound classes for all frames from each cluster. The scores are calculated as described in the previous paragraph (Fig. 2).

2.4. Emotional Mimicry Intensity Estimation

The EMI estimation is a multi-output regression problem with six categories (Admiration, Amusement, Determination, Empathic Pain, Excitement, and Joy. In contrast to previous tasks, one label per whole video is available. Hence, it is necessary to obtain a single descriptor for an entire video, given the facial features of every frame. The official training set contains 8072 videos, while 4588 videos are available for validation.

In this paper, we used simple STAT (statistical) features that have previously shown excellent performance in EmotiW (Emotion recognition in-the-Wild) challenges [2, 7]. In particular, we compute component-wise mean, standard deviation, minimum, and maximum of logits/embeddings at the output of our model and concatenate them into a single descriptor. The latter is fed into a linear classifier (feed-forward neural network without hidden units) with six outputs and sigmoid activation functions. The weighted Pearson Correlation Coefficient (PCC) ρ loss is minimized for logits $\mathbf{z}$ by the Adam with 100 epochs. If the embeddings of our models or wav2vec 2.0 features are estimated, they are fed into a multi-layer perceptron with one hidden layer and 128 units. The output scores of the trained neural net without any post-processing are directly used as the final predictions.

3. Experimental Results

3.1. Facial Emotion Analysis for Static Photos

In the first experiment, we demonstrate the efficiency of our models for the official validation part of AffectNet [33]. Table 1 contains the RMSE (Root Mean Square Error) and CCC for predicted valence/arousal and accuracy for facial expression recognition. In the latter case, we compute two metrics traditionally used with this dataset, namely, 8-Acc (Accuracy for all eight classes) and 7-Acc (Accuracy for seven basic categories: Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise). Existing papers usually train two different models with 8 and 7 outputs and report the performance of each model separately. However, our primary goal is to study the universality of our models. Hence, we train only one model with eight emotional categories and remove the logits corresponding to Contempt to obtain the outputs for seven basic classes.

As one can see, our models trained with the multi-task loss (1) show very high performance. The state-of-the-art DDAMFN [52] is only slightly more accurate when compared to our MT-DDAMFN (and this difference is insignificant [42]). However, our main objective was to obtain the models that can serve as reliable feature extractors for multiple downstream tasks. Let us demonstrate their advantages using data from the ABAW competition in the following subsections.

3.2. VA Estimation

Table 2 shows the VA prediction task validation results. We compare our results with the baselines of the challenge organizers [24] and several papers from the ABAW CVPR 2023 workshop.

First, we significantly improved the metrics compared to the previous attempts with EfficientNet features [40]. MT-DDAMFN achieves the top performance. It is important to emphasize that it has 2$ greater mean CCC when compared to the initial DDAMFN, thus showing the benefits of our training procedure (Subsection 2.1). It is also remarkable that the quality of the largest EmotiEffNet-B2 is the worst among our models, though it also reaches very high accuracy on AffectNet (Table 1). This result highlights the need to verify that the facial analysis model works in various domains and cross-dataset environments.

In addition, we decided to demonstrate the quality of the pre-trained models for VA prediction (Table 3) without using the training set. Here, the best quality of VA estimation is obtained by the MT-MobileFaceNet model, which is 1.1% greater when compared to other models.

Finally, the test set results of the ABAW-6 competition are shown in Table 4. Forty teams submitted their results, out of which ten teams scored higher than the baseline. Our solution has a much higher total score compared to the baseline of the organizers (0.519 vs 0.201) by simple replacement of ResNet-50 to our pre-trained models. As a result, we took the sixth place in this competition.

3.3. CE Recognition

As mentioned in Subsection 2.3, the CE prediction task has no baselines or direct metrics. Hence, we measure the indirect metric of class balance. In addition to our initial models, we used the neural networks trained to predict facial expressions using data from the EXPR (expression recognition) challenge (hereinafter, "(EXPR_ft)"). We used either maximal scores at the output of our models or performed clustering of 1) scores from the final layer, 2) embeddings from the penultimate layer, and 3) wav2vec 2.0 audio embeddings.

The Kullback-Leibler (KL) divergence between actual and predicted class probabilities for all our models is shown in Table 5. As one can notice, the KL divergence in several cases is relatively high, caused by a significant class imbalance of our predictions. It seems that MT-MobileFaceNet

Model	Facial expressions		Valence		Arousal	
	8-Acc., % (↑)	7-Acc., % (↑)	RMSE (↓)	CCC (↑)	RMSE (↓)	CCC (↑)
AlexNet [33]	58.0	-	0.394	0.541	0.402	0.450
SSL inpanting-pl [34]	61.72	-	-	-	-	-
Distract Your Attention [47]	62.09	65.69	-	-	-	-
ViT-base + MAE [26]	62.42	-	-	-	-	-
Static-to-Dynamic [5]	63.06	66.42	-	-	-	-
DDAMFN [52]	64.25	67.03	-	-	-	-
EmotiEffNet-B0	61.32	64.57	-	-	-	-
MT-EmotiEffNet	61.93	64.97	0.434	0.594	0.387	0.549
MT-EmotiMobileFaceNet	62.32	65.17	0.447	0.577	0.387	0.547
MT-EmotiMobileViT	62.50	66.46	0.423	0.599	0.371	0.565
EmotiEffNet-B2 [41]	63.03	66.29	-	-	-	-
EmotiEffNet-B2	63.13	66.51	-	-	-	-
MT-DDAMFN	64.20	67.00	0.363	0.729	0.341	0.643

Table 1. Results for the AffectNet validation set (high Accuracy and CCC are better, low RMSE is better)

Method	CCC_V	CCC_A	P_{VA}
Baseline ResNet-50 [24]	0.24	0.20	0.22
EfficientNet-B0 [38]	0.449	0.535	0.492
Resnet50 + Regnet + EfficientNet [46]	0.257	0.383	0.320
Audio/video Channel Attention Network [53]	0.423	0.670	0.547
Audio/video MAE [54]	0.476	0.644	0.560
Audio/video Transformer [57]	0.554	0.659	0.607
Audio/video TCN [58]	0.550	0.681	0.615
EmotiEffNet-B2	0.423	0.498	0.464
EmotiEffNet	0.443	0.519	0.482
EmotiEffNet, smoothing	0.490	0.596	0.543
DDAMFN	0.438	0.523	0.481
DDAMFN, smoothing	0.485	0.598	0.541
MT-EmotiEffNet	0.444	0.521	0.483
MT-EmotiEffNet, smoothing	0.490	0.604	0.547
MT-EmotiMobileViT	0.445	0.525	0.485
MT-EmotiMobileViT, smoothing	0.493	0.612	0.552
MT-EmotiMobileFaceNet	0.439	0.532	0.486
MT-EmotiMobileFaceNet, smoothing	0.483	0.610	0.547
MT-DDAMFN	0.468	0.537	0.502
MT-DDAMFN, smoothing	0.519	0.616	0.568

Table 2. Valence-Arousal Challenge Results on the Aff-Wild2's validation set.

Model	CCC_V	CCC_A	P_{VA}
MT-DDAMFN	0.412	0.230	0.321
MT-EmotiMobileViT	0.403	0.244	0.324
MT-EmotiMobileFaceNet	**0.413**	**0.266**	**0.339**
MT-EmotiEffNet	0.404	0.248	0.326

Table 3. Results of pre-trained Valence-Arousal prediction models on the Aff-Wild2's validation set. The best result is marked in bold.

Model	CCC_V	CCC_A	P_{VA}
Netease Fuxi AI Lab [55]	0.6873	0.6569	0.6721
DeepAVER [35]	0.5418	0.6196	0.5807
CtyunAI [59]	0.5223	0.6057	0.564
SUN_CE [36]	0.5355	0.5861	0.5608
USTC-IAT-United [49]	0.5208	0.5748	0.5478
KBS-DGU	0.4836	0.5318	0.5077
HSE-NN-SberAILab [40]	0.4818	0.5279	0.5048
ETS-LIVIA [44]	0.4198	0.4669	0.4434
CAS-MAIS [44]	0.4245	0.3414	0.3830
Baseline ResNet-50 [24]	0.211	0.191	0.201
DDAMFN	0.4805	0.5373	0.5089
MT-EmotiMobileViT	0.4807	0.5375	0.5091
MT-EmotiMobileFaceNet	0.4961	0.5264	0.5113
MT-DDAMFN	0.4921	0.5481	0.5202

Table 4. Valence-Arousal Challenge Results on the Aff-Wild2's test set.

achieves the best balance, though the results of the MT-EmotiEffNet-B0 are also very low [39].

To better compare our predictions, we present Cohen's kappa coefficient, which typically measures the inter-rater reliability (Fig. 3). The most consistent with other mod-

Model	Scores	Scores	Clustering Embeddings	Audio
DDAMFN	0.1847	0.1516	0.2362	0.1792
EmotiEffNet-B0	0.1186	0.8419	1.0230	0.1681
EmotiEffNet-B0 (EXPR_ft)	0.5643	**0.0489**	1.1008	0.2105
EmotiEffNet-B2	0.4968	0.1674	0.1791	0.1849
MT-EmotiEffNet	0.0830	0.2161	0.5211	0.1681
MT-EmotiEffNet (EXPR_ft)	0.1934	0.1715	0.3206	**0.0368**
MT-EmotiMobileFaceNet	**0.0603**	0.3316	0.5846	0.1394
MT-EmotiMobileViT	0.1689	0.1265	**0.0975**	0.1646
MT-DDAMFN	0.2687	0.2037	0.2408	0.1503

Table 5. The Kullback-Leibler divergence between real and predicted class probabilities for CE recognition.

Model	F1-score
Netease Fuxi AI Lab [55]	0.5526
USTC-IAT-United [49]	0.2240
SUN_CE [36]	0.2201
USTC-AC [45]	0.1845
Audio clustering + MT-EmotiMobileFaceNet	0.1232
Audio clustering + MT-EmotiEffNet (EXPR_ft)	0.1468
EmotiEffNet-B0 (EXPR_ft)	0.1719
EmotiEffNet-B2	0.1800
MT-EmotiMobileViT	0.2009
MT-DDAMFN	0.2077
MT-EmotiEffNet	0.2341
DDAMFN	0.2395
MT-EmotiEffNet (EXPR_ft)	0.2580
EmotiEffNet-B0	0.2625
MT-EmotiMobileFaceNet	0.2708

Table 6. F1-score of CE recognition on the test set.

els are EmotiEffNet-B0, MT-EmotiEffNet-B0, and MT-MobileFaceNet. Moreover, the clustering results seem inconsistent with other models, so we do not expect this approach to be as accurate as other models.

Table 6 shows the test set's results. Here, 17 Teams submitted their results, and 5 made valid submissions. We took second place, and the gain over the third F1-score [49] is 5%. The difference with the leader [55] is too high. However, in contrast to the winner, the weights of our models are publicly available, so the reproducibility of our results should not be too complicated.

3.4. EMI Estimation

As the previous edition of EMI at the ABAW-5 [23] used much more training data, our results are not directly compared with participants of that challenge. We can only compare with the audio/visual baselines obtained by ViT (Visual Transformer) and wav2vec 2.0 features. The results of our ablation experiments for the EMI task are presented in Table 7.

Here, our facial models are 6-8% more accurate than

Figure 3. Kappa Cohen scores for CE predictions

the ViT baseline. However, the audio features are classified much better. We used a more straightforward approach for processing acoustic features, thus leading to 1% less macro-averaged Pearson correlation $\overline{\rho}$. However, our best ensemble is 4% more accurate. Like most previous experiments, multi-task learning loss (1) leads to a 0.5% better MT-DDAMFN model. The 40-dimensional scores (logits) from the final layer of our model are recognized as not worse than high-dimensional embeddings from the penultimate layer. Finally, STAT features are typically better than the traditional average pooling of frame-level features.

The results of the test set are presented in Table 8. Finally, the test set results of the ABAW-6 competition are shown in Table 4. Forty teams submitted their results, out of which ten teams scored higher than the baseline. Our solution has a much higher total score than the organizers' baseline (0.52 vs. 0.20) by simple replacement of ResNet-50 to our pre-trained MT-DDAMFN model. As a result, we took the sixth place in this competition.

4. Conclusion

To conclude, we introduce several novel lightweight models trained in the multi-task framework (1) to simultaneously predict facial expression, valence, and arousal on a static photo. The neural network weights and the training source code to reproduce the experiments for the presented approach are publicly available[1].

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

Modality	Model	Features	PCC $\overline{\rho}$	Admiration	Amusement	Determination	Empathic Pain	Excitement	Joy
Faces	Baseline ViT [24]	Embeddings	0.09	-	-	-	-	-	-
Audio	Wav2Vec2 [24]	Embeddings	0.24	-	-	-	-	-	-
Audio+ Video	ViT+ Wav2Vec2 [24]	Embeddings	0.25	-	-	-	-	-	-
Faces	MobileFaceNet (VggFace2)	Embeddings (mean)	0.0734	0.0235	0.0542	0.0645	0.0837	0.1053	0.1093
		Embeddings (STAT)	0.0972	0.0374	0.1008	0.0981	0.0972	0.1320	0.1175
Faces	DDAMFN	Embeddings (mean)	0.1619	0.0139	0.2515	0.1211	0.0841	0.2373	0.2641
		Embeddings (STAT)	0.1603	0.0595	0.2169	0.1355	0.0687	0.2245	0.2565
		Scores (mean)	0.1640	0.0174	0.2462	0.1257	0.0740	0.2438	0.2770
		Scores (STAT)	0.1684	0.0354	0.2461	0.1304	0.0634	0.2426	0.2927
Faces	EmotiEffNet -B0	Embeddings (mean)	0.1647	0.0472	0.2387	0.1272	0.1017	0.2225	0.2508
		Embeddings (STAT)	0.1658	0.0596	0.2308	0.1318	0.0743	0.2373	0.2611
		Scores (mean)	0.1597	0.0163	0.2342	0.1315	0.0708	0.2281	0.2765
		Scores (STAT)	0.1645	0.0186	0.2477	0.1277	0.0787	0.2278	0.2863
Faces	MT-EmotiEffNet -B0	Embeddings (mean)	0.1632	0.0162	0.2336	0.1239	0.1001	0.2339	0.2715
		Embeddings (STAT)	0.1673	0.0349	0.2318	0.1379	0.0877	0.2428	0.2683
		Scores (mean)	0.1584	0.0275	0.2115	0.1258	0.0805	0.2273	0.2776
		Scores (STAT)	0.1590	0.0188	0.2335	0.1150	0.0729	0.2312	0.2828
Faces	MT-EmotiMobile-ViT	Embeddings (mean)	0.1644	0.0379	0.2314	0.1387	0.0781	0.2334	0.2672
		Embeddings (STAT)	0.1683	0.0433	0.2459	0.1347	0.0779	0.2382	0.2699
		Scores (mean)	0.1642	0.0321	0.2484	0.1490	0.0674	0.2399	0.2481
		Scores (STAT)	0.1727	0.0621	0.2548	0.1430	0.0624	0.2398	0.2738
Faces	MT-DDAMFN	Embeddings (mean)	0.1628	0.0289	0.2385	0.1281	0.0761	0.2363	0.2689
		Embeddings (STAT)	0.1723	0.0613	0.2319	0.1282	0.1064	0.2446	0.2610
		Scores (mean)	0.1682	0.0408	0.2333	0.1387	0.0825	0.2429	0.2710
		Scores (STAT)	0.1703	0.0289	0.2450	0.1298	0.0878	0.2410	0.2895
Faces	MT-Emoti-MobileFaceNet	Embeddings (mean)	0.1518	0.0215	0.2288	0.1140	0.0692	0.2299	0.2476
		Embeddings (STAT)	0.1646	0.0557	0.2380	0.1303	0.0703	0.2325	0.2605
		Scores (mean)	0.1667	0.0276	0.2367	0.1336	0.0807	0.2516	0.2699
		Scores (STAT)	0.1732	0.0285	0.2498	0.1318	0.097	0.2543	0.2776
Audio	wav2vec 2.0	Embeddings (mean)	0.1514	0.2153	0.11760	0.1834	0.1426	0.1275	0.1219
		Embeddings (STAT)	0.2311	0.3006	0.1659	0.2559	0.3198	0.1844	0.1602
Audio + Video	wav2vec 2.0 +	MT-DDAMFN	0.2767	0.2993	0.3079	0.2230	0.2672	0.3008	0.2546
		MT-EmotiMobileViT	0.2829	0.3011	0.2968	0.2595	0.3074	0.3171	0.2152
		MT-EmotiMobileFaceNet	**0.2898**	0.3041	0.3004	0.2584	0.3148	0.3160	0.2452

Table 7. Pearson's correlation for EMI Estimation on the Hume-Vidmimic2's validation set.

Model	F1-score
Netease Fuxi AI Lab [55]	0.7185
USTC-IAT-United [10]	0.5536
USTC-AC [50]	0.3594
wav2vec 2.0 + MT-EmotiMobileFaceNet (train)	0.3201
wav2vec 2.0 + MT-EmotiMobileFaceNet (train+val)	0.3285
MT-EmotiMobileFaceNet (train+val)	0.1786
wav2vec 2.0 + MT-EmotiMobileViT (train+val)	**0.3316**
wav2vec 2.0 + MT-DDAMFN (train+val)	0.3139
Baseline [24]	0.25

Table 8. EMI Estimation Pearson's correlation on the Hume-Vidmimic2's test set.

We experimentally demonstrated that our models reach near state-of-the-art results on conventional AffectNet benchmark (Table 1). Moreover, these models extract emotional features that can be used in various downstream tasks. We demonstrated the results for the five functions from the sixth ABAW challenge [24], which are essentially better when compared to baselines. For example, our best models achieved the following quality on official validation sets: CCC for VA estimation $P_{VA} = 0.568$ (0.35 greater than baseline VGGFACE, Table 2). In addition, the best facial model for EMI estimation reaches macro-averaged Pearson correlation $\overline{\rho} = 0.173$ (0.08 better than baseline ViT, Table 7). As a result, our solutions took second place at the CE recognition competition, fourth place in the EMI contest, and sixth place in the VA estimation task.

It is important to emphasize that our approach does not require to fine-tune the model on a new dataset, so only a simple feed-forward neural network should be trained on top of our features. Though this can lead to less accurate models on concrete datasets, we believe that obtaining the facial models that analyze affective behavior in unconstrained environments for various datasets is essential.

Moreover, it is essential to extend our techniques to 3D video representations that go beyond 2D factors and pixel-level consistency [11, 12, 29, 56].

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

References

[1] Xiang An, Jiankang Deng, Jia Guo, Ziyong Feng, XuHan Zhu, Jing Yang, and Tongliang Liu. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial FC. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4042–4051, 2022.

[2] Sarah Adel Bargal, Emad Barsoum, Cristian Canton Ferrer, and Cha Zhang. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*, pages 433–436, 2016.

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 67–74. IEEE, 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] Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *arXiv preprint arXiv:2312.05447*, 2023.

[6] 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.

[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] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5203–5212, 2020.

[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] Tobias Hallmen, Fabian Deuser, Norbert Oswald, and Elisabeth André. Unimodal multi-task fusion for emotional mimicry prediciton. *arXiv preprint arXiv:2403.11879*, 2024.

[11] 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.

[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] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] Jia Li, Yin Chen, Xuesong Zhang, Jiantao Nie, Ziqiang Li, Yangchen Yu, Yan Zhang, Richang Hong, and Meng Wang. Multimodal feature extraction and fusion for emotional reaction intensity estimation and expression classification in videos with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5838–5844, 2023.

[26] J Li, J Nie, D Guo, R Hong, and M Wang. Emotion separation and recognition from a facial expression by generating the poker face with vision transformers. *arXiv preprint arXiv:2207.11081*, 2023.

[27] Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12360–12370, 2022.

[28] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[29] 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.

[30] 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.

[31] 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.

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

[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] Mahdi Pourmirzaei, Gholam Ali Montazer, and Farzaneh Esmaili. Using self-supervised auxiliary tasks to improve fine-grained facial representation. *arXiv preprint arXiv:2105.06421*, 2021.

[35] R Gnana Praveen and Jahangir Alam. Recursive cross-modal attention for multimodal fusion in dimensional emotion recognition. *arXiv preprint arXiv:2403.13659*, 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. MT-EmotiEffNet for multi-task human affective behavior analysis and learning from synthetic data. In *Proceedings of European Conference on Computer Vision (ECCV) Workshops*, pages 45–59. Springer, 2022.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] Jiahe Wang, Jiale Huang, Bingzhao Cai, Yifan Cao, Xin Yun, and Shangfei Wang. Zero-shot compound expression recognition with visual language model at the 6th ABAW challenge. *arXiv preprint arXiv:2403.11450*, 2024.

[46] 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.

[47] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics*, 8(2):199, 2023.

[48] Jun Yu, Jichao Zhu, Wangyuan Zhu, Zhongpeng Cai, Guochen Xie, Renda Li, Gongpeng Zhao, Qiang Ling, Lei Wang, Cong Wang, Luyu Qiu, and Wei Zheng. A dual branch network for emotional reaction intensity estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5811–5818, 2023.

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

[50] Jun Yu, Wangyuan Zhu, and Jichao Zhu. Efficient feature extraction and late fusion strategy for audiovisual emotional mimicry intensity estimation. *arXiv preprint arXiv:2403.11757*, 2024.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[57] 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.

[58] 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.

[59] 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.