Leveraging Lightweight Facial Models and Textual Modality in Audio-visual Emotional Understanding in-the-Wild

Andrey V. Savchenko[1,2]
[1]ITMO University
Saint Petersburg, Russia
[2]Sber AI Lab
Moscow, Russia
avsavchenko@hse.ru

Lyudmila V. Savchenko
HSE University
Moscow, Russia
lsavchenko@hse.ru

Abstract

This article presents our results for the eighth Affective Behavior Analysis in-the-Wild (ABAW) competition. We combine facial emotional descriptors extracted by lightweight pre-trained models from our EmotiEffLib library with acoustic features and embeddings of texts recognized from speech. The frame-level features are aggregated and fed into simple classifiers, e.g., multi-layered perceptron (feed-forward neural network with one hidden layer), to predict ambivalence/hesitancy and facial expressions. In the latter case, we also use the pre-trained facial expression recognition model to select high-score video frames and prevent their processing with a domain-specific video classifier. The video-level prediction of emotional mimicry intensity is implemented by simply aggregating frame-level features and training a multi-layered perceptron. Experimental results for four tasks from the ABAW challenge demonstrate that our approach significantly increases validation metrics compared to existing baselines. As a result, our solutions took first place in the expression classification and Ambivalence/Hesitancy recognition challenges, and third place in emotional mimicry intensity estimation and action unit detection tasks.

1. Introduction

Emotion recognition is crucial in advancing human-computer interaction, psychological research, and AI applications [8, 13]. Understanding human emotionality in real-world scenarios enables the development of empathetic AI systems, enhances mental health monitoring, and improves human-centric technologies such as virtual assistants, social robots, and affective computing systems [17, 22]. However, analyzing affective behavior in practical applications presents unique challenges due to the complexity and variability of human emotions, which are often subtle, ambiguous, and context-dependent. For instance, in mental health, accurately identifying ambivalence or hesitancy [30] in a patient's behavior can uncover conflicting emotions and intentions related to behavioral change interventions and provide early indicators of psychological distress or decision-making struggles. Emotional mimicry intensity (EMI), which reflects the degree to which individuals unconsciously mirror the emotions of others [37, 51], is a critical marker of social bonding and empathy. Facial Expression Recognition (FER), on the other hand, is a foundational component of emotion AI, enabling applications such as personalized education, customer behavior analysis, and human-robot interaction [9, 39]. However, achieving high accuracy in these tasks requires robust methods to handle real-world data's inherent noise and variability, such as poor lighting, occlusions, and diverse cultural expressions [19, 23]. This underscores the need for innovative approaches integrating multimodal data and leveraging advanced techniques to improve performance.

The series of ABAW (Affective Behavior Analysis in-the-Wild) competitions [15, 20, 24, 26, 29] based on Aff-Wild [54] and Aff-Wild2 [18] datasets has been instrumental in pushing the boundaries of emotion recognition by providing researchers with a challenging and dynamic benchmark for evaluating models in real-world, in-the-wild conditions. The eighth edition of the ABAW competition (ABAW-8) [21, 30] introduces a novel task that focuses on Ambivalence/Hesitancy (AH) and continues studies of measuring EMI and video-based frame-wise Expression (EXPR) recognition, Action Unit (AU) detection, etc., from the sixth ABAW challenge [28], reflecting the complexity and nuances of human emotional expression. These tasks are essential for developing systems that can accurately interpret human emotions in naturalistic settings, paving the way for more empathetic and responsive technologies.

In this paper, we present the results of our team in the ABAW-8 competition. We propose a novel framework that integrates multiple modalities, visual, acoustic, and linguistic, to enhance the accuracy and robustness of affective behavior recognition. Our approach combines facial emotional descriptors extracted using the EmotiEffLib library [44] with acoustic emotional features [1, 11] and text embeddings [32] obtained from speech recognition [38]. We employ simple yet effective classifiers for each feature, such as linear feed-forward neural networks, and aggregate their predictions in a late fusion technique. Additionally, we employ a pre-trained FER model to filter video frames with high confidence scores, reducing redundant processing and optimizing computational resources. Our experimental results demonstrate the effectiveness of this approach, highlighting its potential to advance the field of affective computing and emotion understanding.

2. Related Works

In this section, we briefly review the results of participants of previous ABAW-6 tasks [28], namely, EXPR classification and EMI estimation.

2.1. EXPR Classification

One of the most widely studied ABAW challenges is the frame-wise uni-task expression classification in video [15, 26, 28]. It is required to assign one of eight basic emotions to each video frame using audio-visual information and models pre-trained on external datasets. Top-performing solutions often employ pre-trained deep learning models fine-tuned on this dataset. For instance, the usage of CLIP was proposed in [31], where the visual features are fed into a trainable multilayer perceptron (MLP) enhanced with Conditional Value at Risk. The HSEmotion team utilized lightweight architectures like MobileViT and EfficientNet, trained in multi-task scenarios [16] to recognize facial expressions, valence, and arousal on static photos [45, 46]. To predict facial expressions, these models extracted frame-level features fed into simple classifiers, such as linear feed-forward neural networks. This approach significantly improved validation metrics compared to existing baselines.

Yu et al. [49] explored semi-supervised learning (SSL) techniques [4, 12, 34] to address the limited size of labeled FER datasets. By generating pseudo-labels for unlabeled faces and employing an encoder to capture temporal relationships between neighboring frames, their approach achieved third place in the ABAW-6 competition. The second place was achieved by Masked Autoencoders (MAE) [57], which has been pre-trained on external datasets and then fine-tuned on the EXPR training set. It was also proposed that continuous emotion recognition be improved by integrating Temporal Convolutional Network and Transformer Encoder modules. Finally, the winners [56] also employed MAE for high-quality facial feature extraction, leading to superior performance across multiple affective behavior analysis tasks. Moreover, they proposed integrating multi-modal knowledge using a transformer-based feature fusion module to combine emotional information from audio signals, visual images, and transcripts. In addition, they divided the dataset division based on scene characteristics and trained a separate classifier for each scene.

2.2. EMI Estimation

EMI estimation task is a multi-label video classification problem, in which it is required to predict the emotional intensity of the video by selecting from a range of six predefined emotional categories [28]. Successful approaches integrated multimodal features, including visual, auditory, and textual cues, to capture a comprehensive emotional profile. For example, Yu et al. [51] presented a third-place solution focusing on efficient feature extraction and a late fusion strategy [41] for audiovisual EMI estimation. They extracted dual-channel visual features and single-channel audio features, averaging the predictions of visual and acoustic models to achieve a more accurate analysis. The second pace took a technique with pure audio analysis [10]. A pre-trained Wav2Vec2.0 [1] was used with a Valence-Arousal-Dominance (VAD) module. After extracting the acoustic features and the VAD predictions compute a global feature vector and fuse the temporal features in a recurrent neural network. They also demonstrated that visual features in this challenge are not very useful. Qiu et al. [37] developed a language-guided multi-modal framework for EMI estimation, leveraging textual modality to enhance the estimation process. Their experiments demonstrated they achieved top performance in the competition.

These studies collectively highlight the advancements in FER and EMI estimation, emphasizing the importance of integrating multimodal data, leveraging pre-trained models, multi-task learning [25, 27] and employing innovative fusion strategies to enhance the accuracy and robustness of affective behavior analysis systems.

3. Proposed Approach

In this paper, we introduce the novel approach to audio-visual affective behavior analysis (Fig. 1) based on lightweight facial emotion recognition models from EmotiEffLib [46], a library of efficient neural networks optimized for real-time emotion recognition. Our approach is designed to leverage multimodal inputs, including visual, acoustic, and textual features, to enhance the recognition of affective behavior. The pipeline comprises several key components: feature extraction, multimodal fusion, and classification. The goal is to predict FER, EMI Estimation, and AH

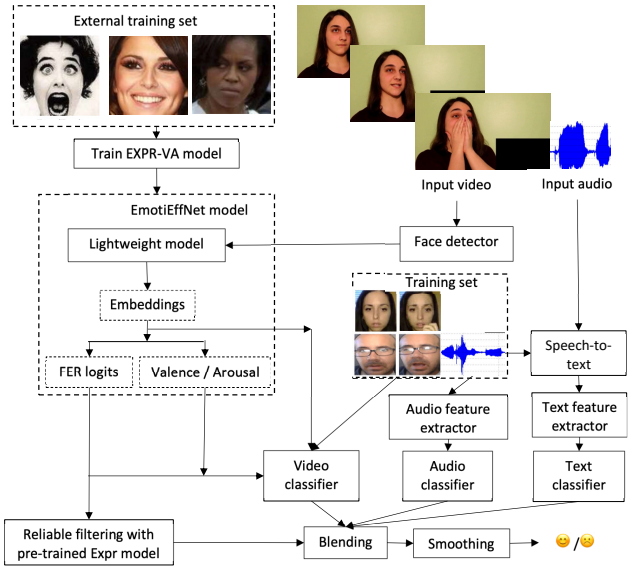


Figure 1. Proposed approach.

Recognition efficiently while maintaining computational efficiency and robustness.

At first, facial regions are cropped in each video frame using RetinaFace [7] or mediapipe landmark detector [33]. At the feature extraction stage, we employ our pre-trained EmotiEffLib neural networks to extract facial emotional descriptors from video frames. Our approach processes individual frames using lightweight neural network architectures [5], such as EmotiEffNet and MobileViT, pre-trained on large-scale affective behavior datasets. These models output embeddings from the penultimate layer [42] and scores (logits of probability distributions) over eight expressions from AffectNet [35], including anger, disgust, fear, happiness, sadness, surprise, neutral, and contempt.

Additionally, acoustic features are extracted using wav2vec 2.0 [1] or HuBERT [11] models. Moreover, audio modality is fed into Speech-to-text module [47] to recognize speech using, e.g., OpenAI's Whisper small.en model [38]. The text embeddings are generated from recognized speech using open-source models, e.g., RoBERTa [32] trained on GoEmotions dataset [6] or commercial ChatGPT and GigaChat embeddings. These modalities capture complementary information, ensuring a comprehensive understanding of affective states.

Each modality is classified separately using specially trained MLP. The fusion and classification step aggregates the extracted frame-wise features across time, creating video-level representations. We employ a simple yet effective linear feed-forward neural network to classify facial videos, audio, and recognized text. We perform blending of these classifiers to obtain the final predictions [40]. To improve accuracy and stability, we apply temporal smoothing to the frame-wise predictions, reducing noise, preventing abrupt fluctuations in classification outputs, and capturing the natural transitions of facial expressions over time [44]. Below, we provide details on the methodology for each task.

3.1. Frame-wise Emotion understanding

In this paper, we consider two tasks for video-based frame-wise emotion understanding: expression classification and AU detection. In the first task, participants recognized eight categories of facial expressions: seven basic emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral) plus an 'other' category. In the second task, a multi-label classification is solved with 12 classes of AUs. The competitions utilize the Aff-Wild2 dataset [18], comprising 548 videos with approximately 2.7 million frames annotated for these expressions.

To extract facial embeddings, we leverage EmotiEffNet-B0 models, which have shown the best performance among EmotiEffib models [46]. It extracts 1280-dimensional embeddings, which we fed into trainable MLP with 128 units in hidden layers. A similar MLP is trained on top of wav2vec 2.0 acoustic features. We do not use speech-to-text and text embeddings here, as most videos do not contain large samples of recognizable speech. The bending of video and audio MLPs outputs (class posterior probabilities) is performed using a weight hyperparameter $w \in [0, 1]$.

To enhance computational efficiency and quality for expression classification, we integrate the same pre-trained EmotiEffNet-B0 model that selects frames with high confidence scores, reducing unnecessary processing. We estimate the posterior probabilities of basic emotions by computing softmax to the logits at the output of the final layer. If the maximal probability is relatively high (greater than a predefined threshold t), we return this expression without using trained MLP for video and audio classifiers. Such a filtering mechanism selects frames with high confidence scores, preventing redundant or misleading frames from affecting the final classification.

3.2. Emotional Mimicry Intensity Estimation

The EMI Estimation challenge aimed to assess the intensity of participants' emotional mimicry in response to 'seed' videos displaying specific emotions (stimulus). Participants were required to predict the intensities of six self-reported emotions for the entire video: admiration, amusement, determination, empathic pain, excitement, and joy. The multimodal Hume-Vidmimic2 dataset, consisting of over 15,000 videos totaling more than 30 hours of audiovisual content, was employed for this task.

We extract multimodal features for facial expressions, voice, and speech content. For visual features, we utilize MT-EmotiMobileFaceNet or MT-EmotiMobileViT [45] to capture fine-grained emotional expressions. We used the fa-

cial images officially provided by the organizers. The best results are obtained for the classification of logits at the output of the final layer. We compute STAT (statistical) features (component-wise mean, std, min, max) [2] of frame-wise logits to obtain a 28-dimensional video descriptor.

Audio signals are processed using wav2vec 2.0 [1] and HuBERT [11], which generate embeddings that reflect prosody, tone, and intonation. The text modality, extracted using RoBERTa, ChatGPT (text-embedding-3-small), or GigaChat embeddings, captures the semantic content of spoken words.

The MLPs with six sigmoid outputs are trained for each modality to optimize the weighted Pearson correlations, where the weights are inversely proportional to class counts. The outputs of MLPs are weighted using late fusion techniques, where each modality contributes to the final EMI estimation based on its reliability.

3.3. Ambivalence/Hesitancy Recognition

The BAH (Behavioural Ambivalence/Hesitancy) dataset [30] identifies ambivalence and hesitancy in Q&A videos, which are recorded explicitly for behavioral analysis. The recorded subjects were asked seven questions to elicit a range of emotional responses, including neutral, positive, negative, ambivalent, willing, resistant, or hesitant feelings about their behaviors. The training and validation sets include 84 participants with up to 7 videos, totaling 431 videos with a combined duration of 3.4 hours and approximately 295,500 frames. Each video frame is annotated by the presence (1) or absence (0) of AH.

Our model extracts facial, vocal, and textual features using the general multimodal approach (Fig. 1). A crucial step in our pipeline is the fusion of textual embeddings with visual and acoustic cues, as verbal content often plays a key role in expressing ambivalence. We employ early fusion strategies where the embeddings from RoBERTa trained on GoEmotions [6], wav2vec 2.0 [1], and MT-EmotiMobileFaceNet [45] are concatenated and processed by a feed-forward neural network. Additionally, we experiment with blending techniques, where predictions from individual modalities are averaged to refine the final classification output.

Unfortunately, the acoustic and text features are not aligned with video frames. To perform frame-level predictions, we interpolated the acoustic and text features for the shape of visual features using interp1d from SciPy. Moreover, we examined the possibility of training a video-level classifier to predict if it contains at least one frame with an AH label set to 1. Here, we compute the component-wise mean of text features, train a logistic regression model, and use its output to filter videos for which this classifier predicts an absence of AH.

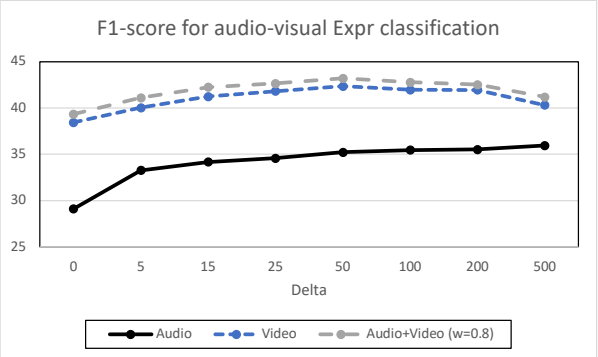


Figure 2. Dependence of F1-score for video Expr recognition on the smoothing kernel size k.

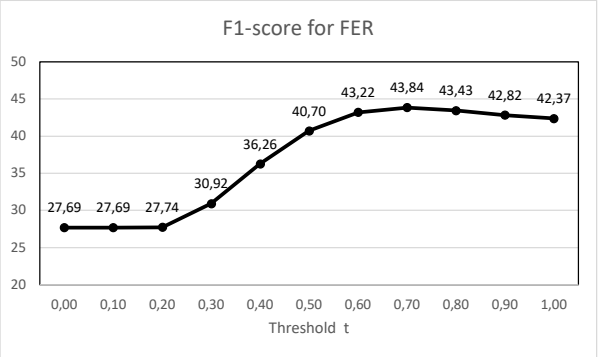


Figure 3. Dependence of F1-score for video Expr recognition on the filtering threshold t.

4. Experimental Results

4.1. Facial Expression Recognition

To evaluate our approach for the EXPR classification task, we tested various models on the Aff-Wild2 validation set. Table 1 summarizes the F1-scores and accuracy achieved by different methods. The baseline VGGFACE (MixAugment) model achieved an F1-score of 25.0, while EfficientNet-B0 improved this score to 40.2.

Our best-performing model, EmotiEffNet with filtering and smoothing, achieved an F1-score of 43.83%, outperforming previous approaches that used visual models from EmotiEffLib [46] up to 1.5%. The combination of wav2vec 2.0 for audio processing [1] and EmotiEffNet for FER [44] further improved the performance, reaching an F1-score of 44.59% if filtering with pre-trained EmotiEffNet-B0 model is applied.

Fig. 2 shows the dependence of the F1-score on the smoothing kernel size k, illustrating the impact of temporal smoothing on expression recognition. The dependence of the F1-score for video modality on the filtering threshold t is presented in Fig. 3. Here, the top F1-score 43.84% is obtained when $t = 0.7$.

Method	Modality	Is ensemble?	F1-score	Accuracy
Baseline VGGFACE (MixAugment) [28]	Faces	No	25.0	-
EfficientNet-B0 [43]	Faces	No	40.2	-
MT-EmotiMobileViT [45]	Faces	No	35.6	46.1
CLIP [31]	Faces	No	36.0	-
SSL + Temporal+ Post-process [49]	Audio/video	Yes	44.43	-
MAE [55]	Audio/video	Yes	49.5	-
MAE + Transformer feature fusion [56]	Audio/video	Yes	55.55	-
wav2vec 2.0	Audio	No	29.09	41.01
wav2vec 2.0, smoothing	Audio	No	35.95	52.36
EmotiEffNet	Faces	No	38.44	49.54
EmotiEffNet, smoothing	Faces	No	42.37	54.34
EmotiEffNet, filtering + smoothing	Faces	No	43.83	54.29
wav2vec 2.0+EmotiEffNet	Audio/video	Yes	40.30	52.03
wav2vec 2.0+EmotiEffNet, smoothing	Audio/video	Yes	43.43	55.67
wav2vec 2.0+EmotiEffNet, filtering + smoothing	Audio/video	Yes	44.59	55.32

Table 1. Expression Classification Results on the Aff-Wild2's validation set.

Model	F1-score
Our EmotiEffNet, filtering + smoothing	34.84
Our wav2vec 2.0 + EmotiEffNet, filtering + smoothing	**36.47**
CtyunAI [58]	36.3
USTC-IAT-United [52]	36.2
AIWELL-UOC [3]	30.6
CAS-MAIS	27.0
Baseline VGGFACE [30]	22.5

Table 2. Leaderboard in the expression classification challenge: F1-score on the test set.

Method	F1-score
Baseline VGGFACE [28]	39.0
CLIP [31]	43.0
AUD-TGN [50]	53.7
MAE+Transformer Fusion [56]	56.94
MAE ViT + TCN [57]	57.62
SimMIM + wav2vec [14]	70.7
EmotiEffNet, smoothing	54.5
EmotiEffNet + LightAutoML, smoothing	55.4

Table 3. Action Unit Detection on the Aff-Wild2's validation set.

Model	F1-score
USTC-IAT-United [52]	51.5
CtyunAI [58]	50.2
Our EmotiEffNet, MLP + LightAutoML	48.8
PR-VSL [36]	48.4
AIWELL-UOC [36]	41.7
Baseline VGGFACE [28]	36.5

Table 4. Leaderboard in the AU detection challenge: F1-score on the test set.

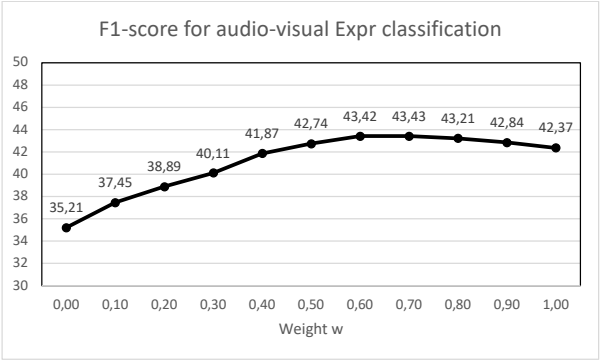


Figure 4. Dependence of F1-score for audio-visual Expr recognition on the weight w.

Fig. 4 further analyzes the effect of blending weight hyperparameter w in the audiovisual recognition pipeline, demonstrating the optimal balance between modalities. Finally, as shown in Fig. 5, the filtering threshold t significantly impacts the classification results, demonstrating the benefits of preprocessing techniques for improving recognition performance. These results confirm that frame filtering, temporal smoothing, and multimodal fusion significantly enhance the accuracy and robustness of FER.

The qualitative results for our filtering procedure with threshold $t = 0.9$ are shown in Fig. 6. Here, we present

Modality	Model	Features	PCC $\overline{\rho}$	Admiration	Amusement	Determination	Empathic Pain	Excitement	Joy
Faces	Baseline ViT [28]		0.09	-	-	-	-	-	-
Audio	wav2Vec2 [28]		0.24	-	-	-	-	-	-
Audio+Video	ViT+wav2Vec2 [28]		0.25	-	-	-	-	-	-
Audio+Video	ResNet18+AUs+Wav2Vec2.0 [51]		0.3288	-	-	-	-	-	-
Audio	Wav2Vec2.0+VAD [10]		0.389	-	-	-	-	-	-
Audio+Video	EmoViT+HuBERT+ChatGLM3 [37]		0.5851	0.7155	0.6159	0.6303	0.3488	0.6174	0.5793
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)	0.1776	0.0619	0.2623	0.1247	0.0773	0.2544	0.2849
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	HuBERT	Embeddings (mean)	0.3045	0.3644	0.3179	0.2987	0.3111	0.2848	0.2499
		Embeddings (STAT)	0.2729	0.3585	0.2454	0.2432	0.3065	0.2566	0.2275
Text	RoBERTa (emotional)	Embeddings (mean)	0.3763	0.4558	0.3328	0.3427	0.4964	0.3201	0.3099
		Embeddings (STAT)	0.3697	0.4501	0.3326	0.3173	0.4843	0.3331	0.3008
Text	SentenceBERT	Embeddings (mean)	0.3756	0.4710	0.3585	0.3589	0.4347	0.3186	0.3119
Text	Sentence all-MiniLM	Embeddings (mean)	0.3537	0.4552	0.3076	0.3371	0.4111	0.3093	0.3023
Text	SentenceE5	Embeddings (mean)	0.3777	0.4719	0.3497	0.3417	0.4553	0.3293	0.3184
Text	OpenAI (small)	Embeddings (mean)	0.4011	0.5113	0.3825	0.3687	0.4847	0.3363	0.3233
Text	GigaChat	Embeddings (mean)	0.4001	0.4888	0.3648	0.3732	0.4862	0.3726	0.3151
Audio + Video	wav2vec 2.0 + MT-EmotiMobileViT		0.2829	0.3011	0.2968	0.2595	0.3074	0.3171	0.2152
	wav2vec 2.0 + MT-EmotiMobileFaceNet		0.2898	0.3041	0.3004	0.2584	0.3148	0.3160	0.2452
Text + Audio	RoBERTa + HuBERT, blending		0.3974	0.4644	0.3727	0.3814	0.4755	0.3477	0.3430
Text + Video	RoBERTa + MT-EmotiMobileViT, blending		0.4028	0.4379	0.3911	0.3580	0.4802	0.3656	0.3840
Text + Video	RoBERTa + MT-EmotiMobileFaceNet, blending		0.4074	0.4408	0.3877	0.3632	0.5064	0.3650	0.3814
Text + Video + Audio	RoBERTa + MT-EmotiMobileViT + HuBERT, blending		0.4192	0.4603	0.4104	0.3844	0.4935	0.3802	0.3864
Text + Video + Audio	RoBERTa + MT-EmotiMobileFaceNet + HuBERT, blending		0.4223	0.4620	0.4095	0.3848	0.4936	0.3876	0.3965
Text + Audio	OpenAI + HuBERT, blending		0.4125	0.5114	0.4006	0.3776	0.4863	0.3645	0.3344
Text + Video	OpenAI + MT-EmotiMobileViT, blending		0.4103	0.4932	0.4090	0.3327	0.4571	0.3771	0.3930
Text + Video	OpenAI + MT-EmotiMobileFaceNet, blending		0.3887	0.4779	0.3593	0.3369	0.4373	0.3452	0.3758
Text + Video + Audio	OpenAI + MT-EmotiMobileViT + HuBERT, blending		0.4451	0.5203	0.4461	0.3938	0.5054	0.4015	0.4036
Text + Video + Audio	OpenAI + MT-EmotiMobileFaceNet + HuBERT, blending		0.4225	0.5073	0.4044	0.3696	0.4801	0.3822	0.3912
Text + Audio	GigaChat + HuBERT, early		0.4011	0.4794	0.3733	0.4182	0.4449	0.3516	0.3395
Text + Audio	GigaChat + HuBERT, blending		0.4106	0.4955	0.3912	0.3881	0.4619	0.3922	0.3348
Text + Video	GigaChat + MT-EmotiMobileViT, blending		0.4103	0.4932	0.4090	0.3327	0.4571	0.3771	0.3930
Text + Video	GigaChat + MT-EmotiMobileFaceNet, early		0.4199	0.4906	0.3988	0.4121	0.4420	0.3876	0.3883
Text + Video	GigaChat + MT-EmotiMobileFaceNet, blending		0.4231	0.4897	0.4107	0.3806	0.4824	0.4131	0.3618
Text + Video + Audio	GigaChat + MT-EmotiMobileViT + HuBERT, blending		**0.4460**	0.5197	0.4491	0.3916	0.4981	0.4192	0.3983
Text + Video + Audio	GigaChat + MT-EmotiMobileFaceNet + HuBERT, early		0.4204	0.4823	0.4013	0.4267	0.4496	0.3765	0.3863
Text + Video + Audio	GigaChat + MT-EmotiMobileFaceNet + HuBERT, blending		0.4338	0.4955	0.4332	0.3923	0.4640	0.4282	0.3898

Table 5. Pearson's correlation for EMI Estimation on the Hume-Vidmimic2's validation set.

3 cases (top part of the figure) for which our pre-trained model corrects the MLP decision and 3 cases (bottom part) for which our filtration leads to incorrect results. As one can notice, the model returned reasonable results even in the latter case, for which the ground-truth label looks questionable.

Finally, the test set results of the ABAW-8 competition are shown in Table 2. Thirty teams submitted their results, out of which five teams scored higher than the baseline. Here, our best model lets us take first place in this competition.

4.2. Action Unit Detection

For AU detection, we utilized the LightAutoML framework [48] to classify outputs of pre-trained EmotiEffNet and perform blending with the outputs of MLP [46]. The main results on validation set are summarized in Table 3, while the leaderboard on the test set is shown in Table 4. Our solution took third place among 27 teams that submitted their results.

4.3. Emotional Mimicry Intensity Estimation

For the EMI Estimation task, we evaluated different feature extraction and fusion methods on the Hume-Vidmimic2 validation set. Table 5 presents Pearson's correlation coefficients (PCC) for various models across different emotions.

The baseline ViT model for facial features alone achieved a correlation of 0.09, while wav2vec 2.0 audio embeddings [1] significantly improved performance to 0.24. Combining visual and audio embeddings (ViT + wav2vec 2.0) increased the correlation to 0.25, demonstrating the importance of multimodal integration.

From our results, one can notice that the impact of visual modality is very small. However, our best model, GigaChat

Model	F1-score
HCAI-VIS	0.71
USTC-IAT-United [53]	0.68
Baseline [28]	0.25
CAS-MAIS	0.15
GigaChat + MT-EmotiMobileFaceNet + HuBERT, train + val, blending	0.5056
GigaChat + MT-EmotiMobileViT + HuBERT, train + val, blending	0.5028
OpenAI + MT-EmotiMobileViT + HuBERT, train + val, blending	0.5005
OpenAI + MT-EmotiMobileFaceNet + HuBERT, train + val, blending	0.4944
RoBERTa + MT-EmotiMobileViT + HuBERT, train + val, blending	0.4646

Table 6. Leaderboard in the EMI estimation challenge: Pearson's correlation on the test set.

Method	Modality	F1-score, %
MT-EmotiMobileFaceNet, Scores	Faces	65.35
MT-EmotiMobileFaceNet, Features	Faces	64.57
wav2vec 2.0	Audio	68.50
HuBERT	Audio	69.29
OpenAI small	Text	70.08
Gigachat	Text	73.23
RoBERTa (emotional)	Text	77.16

Table 7. Video-level Ambivalence/Hesitancy Recognition Accuracy on the BAH's validation set.

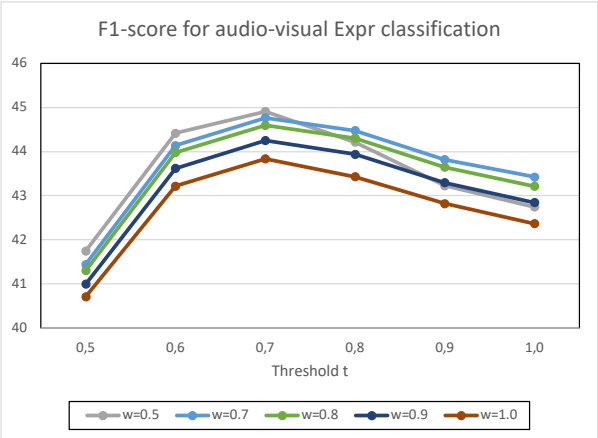
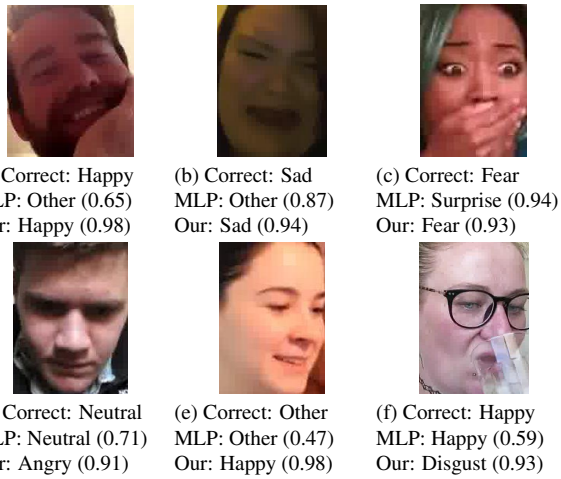


Figure 5. Dependence of F1-score for audio-visual Expr recognition on the filtering threshold t.

text embeddings blended with HuBERT audio [11] and MT-EmotiMobileViT facial video [45] features, achieved a correlation of 0.4460, marking a significant improvement over unimodal approaches. The results of the test set are shown in Table 6. Six teams participated in this competition, and our best solution took third place.



(a) Correct: Happy
MLP: Other (0.65)
Our: Happy (0.98)

(b) Correct: Sad
MLP: Other (0.87)
Our: Sad (0.94)

(c) Correct: Fear
MLP: Surprise (0.94)
Our: Fear (0.93)

(d) Correct: Neutral
MLP: Neutral (0.71)
Our: Angry (0.91)

(e) Correct: Other
MLP: Other (0.47)
Our: Happy (0.98)

(f) Correct: Happy
MLP: Happy (0.59)
Our: Disgust (0.93)

Figure 6. Examples of misclassifications for the proposed expression recognition approach with pre-trained model for robust classifications compared to our MLP model only: (a)-(c) Correct classification; (d)-(e) Mistakes introduced by pre-trained model.

4.4. Ambivalence/Hesitancy Recognition

We evaluated different multimodal strategies on the BAH validation set for the AH Recognition task. At first, we performed intermediate experiments with the global prediction of AH for the entire video. Table 7 contains the F1-score for this task. Here, the top performance is achieved by text modality using RoBERTa emotional embeddings, so we used it in the main experiment if filtering is applied

Table 8 presents the main results, F1-scores achieved using various feature extraction and fusion methods. Baseline approaches using video (MT-EmotiMobileFaceNet), audio (HuBERT), and text embeddings (RoBERTa) achieved scores of 70.59%, 68.57%, and 70.70%, respectively. With smoothing and filtering, the best unimodal visual model, MT-EmotiMobileFaceNet [45], achieved an F1-score of 72.11%.

Our best-performing multimodal model, which combines RoBERTa text embeddings [32] with MT-EmotiMobileFaceNet [45] and smoothing, achieved an F1-score of 73.73%, confirming the effectiveness of integrating textual and visual modalities. The test set results (Table 9) proved that our approach is the best among the nine teams that took part in this challenge.

5. Conclusion

In this paper, we introduced the multimodal approach (Fig. 1) that leverages facial, audio, and textual modalities in a computationally efficient manner. We use pre-trained modes to extract features without the need for their fine-tuning on each task. As a result, we significantly improved the performance of three considered tass (EXPR classifi-

Method	Modality	F1-score, %
Baseline [30]	Faces + Audio + Text	70.0
DDAMFN, Scores	Faces	67.97
DDAMFN, Features	Faces	68.36
EmotiEffNet-B0, Scores	Faces	69.75
EmotiEffNet-B0, Features	Faces	67.66
MT-EmotiMobileViT, Scores	Faces	69.81
MT-EmotiMobileViT, Features	Faces	67.69
MT-DDAMFN, Scores	Faces	68.76
MT-DDAMFN, Features	Faces	68.02
MT-EmotiEffNet, Scores	Faces	68.32
MT-EmotiEffNet, Features	Faces	68.46
MT-EmotiMobileFaceNet, Scores	Faces	70.59
MT-EmotiMobileFaceNet, Features	Faces	68.36
wav2vec 2.0	Audio	67.66
HuBERT	Audio	68.57
RoBERTa (emotional)	Text	70.70
MT-EmotiMobileFaceNet, smoothing	Faces	72.01
MT-EmotiMobileFaceNet, smoothing+ filtering	Faces	72.11
HuBERT+MT-EmotiMobileFaceNet, blending	Audio + Faces	71.51
RoBERTa+MT-EmotiMobileFaceNet, blending	Text + Faces	71.45
RoBERTa + HuBERT + MT-EmotiMobileFaceNet, blending	Text + Audio + Faces	72.12
RoBERTa + HuBERT + MT-EmotiMobileFaceNet, fusion	Text + Audio + Faces	72.01
RoBERTa+MT-EmotiMobileFaceNet, fusion	Text + Faces	72.26
RoBERTa+MT-EmotiMobileFaceNet, smoothing	Text + Faces	73.31
RoBERTa+MT-EmotiMobileFaceNet, smoothing + filtering	Text + Faces	73.73

Table 8. Frame-level Ambivalence/Hesitancy Recognition Challenge Results on the BAH's validation set.

Model	Weighted F1
MT-EmotiMobileFaceNet, train only, smoothing	67.4
MT-EmotiMobileFaceNet, train + val, smoothing	67.4
RoBERTa + HuBERT + MT-EmotiMobileFaceNet, train only, smoothing + filtering	**71.0**
RoBERTa + HuBERT + MT-EmotiMobileFaceNet, train + val, smoothing	70.4
RoBERTa + HuBERT + MT-EmotiMobileFaceNet, train + val, smoothing + filtering	70.7
HCAI-VIS	70.2
Baseline [30]	70.0

Table 9. Leaderboard in the ambivalence/hesitancy recognition challenge: F1-score on the test set.

cation, EMI estimation, and AH recognition) compared to existing baselines [30]. In addition, we used our best models from the ABAW-6 competition in valence-arousal estimation and action unit detection tasks. The source code to reproduce the experiments is publicly available[1]. Our most interesting innovation included filtering of reliable facial expressions obtained by a pre-trained model without refinement on given training sets (Table 1) and usage of global video classifiers and interpolation of acoustic and text embeddings to improve the quality of AH prediction (Table 8). 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. As a result, our solutions took first places in two tasks of the ABAW-8 challenge (video-based frame-wise expression classification and abivalence / hesitancy recognition), and third places in EMI estimation and AU detection tasks.

Our work contributes to the advancement of affective computing research by addressing challenges related to multimodal fusion, domain-specific generalization, and real-time inference. The findings from this study highlight the importance of leveraging multimodal signals and intelligent data selection strategies to achieve state-of-the-art performance in affective behavior analysis. The ability to analyze affective behavior efficiently has significant implications for real-world applications, including emotion-aware virtual agents, adaptive learning systems, and automated mental health assessment tools. In future, it is necessary to apply our approach for other datasets to verify its ability t generalize across different applications with distinct data distributions.

Acknowledgements. This work was supported by R&D Center "Strong Artificial Intelligence in Industry" of ITMO University.

[1] https://github.com/av-savchenko/EmotiEffLib/tree/main/training_and_examples/ABAW/ABAW8

References

[1] 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.

[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] Josep Cabacas-Maso, Elena Ortega-Beltrán, Ismael Benito-Altamirano, and Carles Ventura. Enhancing facial expression recognition through dual-direction attention mixed feature networks and CLIP: Application to 8th ABAW challenge. *arXiv preprint arXiv:2407.12390*, 2024.

[4] Bashar M. Deeb, Andrey V. Savchenko, and Ilya Makarov. Enhancing emotion recognition in speech based on self-supervised learning: Cross-attention fusion of acoustic and semantic features. *IEEE Access*, 13:56283–56295, 2025.

[5] 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.

[6] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4040–4054, 2020.

[7] 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.

[8] Runfang Guo, Hongfei Guo, Liwen Wang, Mengmeng Chen, Dong Yang, and Bin Li. Development and application of emotion recognition technology—a systematic literature review. *BMC psychology*, 12(1):95, 2024.

[9] Xing Guo, Yudong Zhang, Siyuan Lu, and Zhihai Lu. Facial expression recognition: A review. *Multimedia Tools and Applications*, 83(8):23689–23735, 2024.

[10] Tobias Hallmen, Fabian Deuser, Norbert Oswald, and Elisabeth André. Unimodal multi-task fusion for emotional mimicry intensity prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4657–4665, 2024.

[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] Smith K Khare, Victoria Blanes-Vidal, Esmaeil S Nadimi, and U Rajendra Acharya. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information fusion*, 102:102019, 2024.

[14] Jun-Hwa Kim, Namho Kim, Minsoo Hong, and Chee Sun Won. Advanced facial analysis in multi-modal data with cascaded cross-attention based transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7870–7877, 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, Panagiotis Tzirakis, Alan Cowen, Irene Kotsia, UK Cogitat, Eric Granger, Marco Pedersoli, Simon Bacon, Alice Baird, Chunchang Shao, et al. Advancements in affective and behavior analysis: The 8th ABAW workshop and competition.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for multi-task learning of classification tasks: a large-scale study on faces & beyond. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2813–2821, 2024.

[28] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Irene Kotsia, Alice Baird, Chris Gagne, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (ABAW) competition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4587–4598, 2024.

[29] Dimitrios Kollias, Stefanos Zafeiriou, Irene Kotsia, Abhinav Dhall, Shreya Ghosh, Chunchang Shao, and Guanyu Hu. 7th ABAW competition: Multi-task learning and compound expression recognition. *arXiv preprint arXiv:2407.03835*, 2024.

[30] Dimitrios Kollias, Panagiotis Tzirakis, Alan S. Cowen, Stefanos Zafeiriou, Irene Kotsia, Eric Granger, Marco Pedersoli, Simon L. Bacon, Alice Baird, Chris Gagne, Chunchang Shao, Guanyu Hu, Soufiane Belharbi, and Muhammad Haseeb Aslam. Advancements in Affective and Behavior Analysis: The 8th ABAW Workshop and Competition. 2025.

[31] Li Lin, Sarah Papabathini, Xin Wang, and Shu Hu. Robust light-weight facial affective behavior recognition with CLIP. In *Proceedings of 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 607–611. IEEE, 2024.

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[33] 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.

[34] 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.

[35] 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.

[36] Quoc-Tien Nguyen, Hong-Hai Nguyen, and Van-Thong Huynh. Lightweight models for emotional analysis in video. *arXiv preprint arXiv:2503.10530*, 2025.

[37] Feng Qiu, Wei Zhang, Chen Liu, Lincheng Li, Heming Du, Tianchen Guo, and Xin Yu. Language-guided multi-modal emotional mimicry intensity estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4742–4751, 2024.

[38] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 28492–28518. PMLR, 2023.

[39] 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.

[40] Andrey Savchenko, Anton Alekseev, Sejeong Kwon, Elena Tutubalina, Evgeny Myasnikov, and Sergey Nikolenko. Ad lingua: Text classification improves symbolism prediction in image advertisements. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1886–1892, 2020.

[41] Andrey V. Savchenko. Adaptive video image recognition system using a committee machine. *Optical Memory and Neural Networks*, 21:219–226, 2012.

[42] Andrey V. Savchenko. Maximum-likelihood dissimilarities in image recognition with deep neural networks. *Computer Optics*, 41(3):422–430, 2017.

[43] 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.

[44] 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.

[45] Andrey V. Savchenko. Leveraging pre-trained multi-task deep models for trustworthy facial analysis in affective behaviour analysis in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4703–4712, 2024.

[46] Andrey V. Savchenko and Anna P. Sidorova. EmotiEffNet and temporal convolutional networks in video-based facial expression recognition and action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4849–4859, 2024.

[47] Vladimir V. Savchenko and Andrey V. Savchenko. Information-theoretic analysis of efficiency of the phonetic encoding–decoding method in automatic speech recognition. *Journal of Communications Technology and Electronics*, 61:430–435, 2016.

[48] 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.

[49] Jun Yu, Zhihong Wei, Zhongpeng Cai, Gongpeng Zhao, Zerui Zhang, Yongqi Wang, Guochen Xie, Jichao Zhu, Wangyuan Zhu, Qingsong Liu, and Jiaen Liang. Exploring facial expression recognition through semi-supervised

pre-training and temporal modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4880–4887, 2024.

[50] Jun Yu, Zerui Zhang, Zhihong Wei, Gongpeng Zhao, Zhongpeng Cai, Yongqi Wang, Guochen Xie, Jichao Zhu, Wangyuan Zhu, Qingsong Liu, and Jiaen Liang. AUD-TGN: Advancing action unit detection with temporal convolution and gpt-2 in wild audiovisual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4814–4821, 2024.

[51] Jun Yu, Wangyuan Zhu, Jichao Zhu, Zhongpeng Cai, Gongpeng Zhao, Zerui Zhang, Guochen Xie, Zhihong Wei, Qingsong Liu, and Jiaen Liang. Efficient feature extraction and late fusion strategy for audiovisual emotional mimicry intensity estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4866–4872, 2024.

[52] Jun Yu, Yang Zheng, Lei Wang, Yongqi Wang, and Shengfan Xu. Design of an expression recognition solution employing the global channel-spatial attention mechanism. *arXiv preprint arXiv:2503.11935*, 2025.

[53] Jun Yu, Lingsi Zhu, Yanjun Chi, Yunxiang Zhang, Yang Zheng, Yongqi Wang, and Xilong Lu. Dual-stage cross-modal network with dynamic feature fusion for emotional mimicry intensity estimation. *arXiv preprint arXiv:2503.10603*, 2025.

[54] 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.

[55] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Multi-modal 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.

[56] Wei Zhang, Feng Qiu, Chen Liu, Lincheng Li, Heming Du, Tianchen Guo, and Xin Yu. An effective ensemble learning framework for affective behaviour analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4761–4772, 2024.

[57] Weiwei Zhou, Jiada Lu, Chengkun Ling, Weifeng Wang, and Shaowei Liu. Enhancing emotion recognition with pre-trained masked autoencoders and sequential learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4666–4672, 2024.

[58] Weiwei Zhou, Chenkun Ling, and Zefeng Cai. Emotion recognition with CLIP and sequential learning. *arXiv preprint arXiv:2503.09929*, 2025.