Multi-modal Arousal and Valence Estimation under Noisy Conditions

Denis Dresvyanskiy[1,2,*], Maxim Markitantov[3,*], Jiawei Yu[4], Heysem Kaya[4], and Alexey Karpov[3]

[1]Ulm University, Germany
[2]ITMO University, Russia
[3]St. Petersburg Federal Research Center of the Russian Academy of Sciences, St. Petersburg, Russia
[4]Department of Information and Computing Sciences, Utrecht University, The Netherlands

denis.dresvyanskiy@uni-ulm.de, {markitantov.m, karpov}@iias.spb.su, {j.yu, h.kaya}@uu.nl

Abstract

Automatic emotion recognition has gained significant attention over the past two decades due to the central role that emotions play in human communication. While multi-modal systems demonstrate high performances on laboratory-controlled data, their validity on non-lab-controlled, namely 'in-the-wild' data, remains a challenge. This work investigates audio-visual deep learning approaches for emotion recognition in-the-wild, with a particular focus on the effectiveness of architectures based on fine-tuned Convolutional Neural Networks (CNN) and Public Dimensional Emotion Model (PDEM) for video and audio modality, respectively. We explore and compare various temporal modeling techniques (e.g., transformer architectures) and fusion strategies by leveraging the embeddings from developed multi-stage trained modality-specific Deep Neural Networks (DNN). The results are reported on the AffWild2 dataset following the Affective Behavior Analysis in-the-Wild 2024 (ABAW'24) challenge protocol. Our investigation highlights the complexities of robust multi-modal emotion recognition in an unconstrained environment, providing insights into the usage of various deep learning architectures for tackling this challenging task.

1. Introduction

This paper presents our contribution to the 2024 edition [36] of the Affective Behavior Analysis in-the-Wild (ABAW) challenge series [27–29, 31, 32, 34, 69]. To replicate the results of our work, the reader is kindly referred to the GitHub repository[1].

The challenges in the field of affective computing have boosted the development of state-of-the-art methods, while also ensuring the reproducibility and comparability of the developed methods under a common experimental protocol. In ABAW 2024[2], the sub-challenges include 8-class categorical emotion recognition (Expression Challenge-EXPR), featuring Ekman's six basic emotions, plus *neutral* and *others* classes as well as emotion primitives (arousal and valence) regression challenge (VA), on which we report the result of this work. The challenge data and baseline system are introduced in [36] and former editions of ABAW [25, 26, 30, 33, 35]. This year, the organizers introduced a novel Compound Expression Recognition (CER) challenge. Our team also participated in CER challenge [55]. We next present a summary of the leading works that participated in the VA challenge.

2. Related Work

In the 6th ABAW Competition, numerous DNN-based Emotion Recognition (ER) approaches have been proposed, with a primary focus on visual and audio modalities.

The baseline system [36] used cropped and aligned face images resized to 112×112 resolution with normalized pixel values. While for the VA challenge a ResNet architecture with 50 layers was employed, in the EXPR challenge, a VGG16 architecture was used. Additionally, MixAugment [52] was applied for the EXPR task.

In the work [71], the authors introduced an ER methodology that effectively integrates emotional cues from multimodal data sources. Distinct feature encoders were employed to extract salient representations from each modality. Specifically, a Masked Auto-Encoder [16] pre-trained on a large amount of data is used for the visual modality, while the VGGish [6] model is exploited for the audio pipe. Subsequently, an ensemble of Transformer Encoders, trained on different subsets of the AffWild2 dataset, is employed to

[*] These authors contributed equally to this work.
[1] https://github.com/DresvyanskiyDenis/ABAW_2024
[2] https://affective-behavior-analysis-in-the-wild.github.io/6th/

fuse the outputs of feature encoders, reaching the top performance for all challenges of the ABAW'24 competition.

In the work [51], Praveen and Alam focused on the VA estimation task, integrating features from visual, audio, and text modalities. In the visual domain, a ResNet-50 pretrained on MS-CELEB-M and FER+ datasets [15] was combined with Temporal Convolutional Networks (TCNs) to effectively capture spatial and temporal cues. Similarly, VGG architecture was used to extract audio features from spectrograms, with TCNs employed to catch temporal dependencies in vocal signals. In the text modality, BERT embeddings followed by TCNs were utilized. Multi-modal fusion was achieved by a recursive cross-modal attention mechanism, refining feature representations iteratively.

Kim et al. [24] employed a consistent feature and fusion strategy across VA and EXPR tasks. The facial features were extracted using a fine-tuned SimMIM model [66] pre-trained on facial expression data. Audio features were directly extracted using pre-trained Wav2Vec model. Subsequently, a cascaded cross-attention mechanism was applied to fuse features from these two modalities.

In [13], Dresvyanskiy et al. proposed a multi-modal ER approach that combined the outputs of audio and visual systems at the decision level. Although the visual system exhibited the best performance on both the EXPR and VA development sets in comparison with other uni-modal systems, fusing the Transformer-based modality-specific models with a functionals-based ELM method led to further gains in recognition performance on these subsets.

Yu et al. [68] proposed a multimodal system using TCN to capture temporal and spatial correlations between features, followed by a fusion of modality-specific feature representations via a Transformer Encoder. Similar to [71], this work employs VGGish [6] to extract audio features, however, augments them with 39D Mel-Frequency Cepstral Coefficients. As a visual feature encoder, the work uses IResNet-50 [3]. Additionally, after IResNet-50, the authors used an LA-SE module (composed of a LANet [67] and a SENet [18]) to better capture local image information, improve channel selection, and suppression.

Savchenko [57] introduced several lightweight deep learning models based on MobileViT [48], MobileFaceNet [7], and DDAMFN [70] architectures for the multi-task ER using static facial frames. Developed models extract frame-level features, predicting facial expression, valence, and arousal in a multi-task setting, reaching near state-of-the-art results on conventional ER datasets with a notable enhancement on 6th ABAW development sets.

Zhou et al. [73] presented a novel approach to enhance continuous ER by using pre-trained Masked Auto-Encoder [16] on facial datasets, followed by fine-tuning on the AffWild2 dataset. The study integrated TCNs and Transformer Encoder into the framework, showcasing a significant improvement in recognition performance.

In [65], the authors proposed a novel Joint Multi-modal Transformer framework for audio-visual ER. To capture spatio-temporal information in the video, the authors used an R(2+1)D network [20] pre-trained on the Kinetics-400 dataset [40]. For audio, a ResNet18 model with Gated Recurrent Unit (GRU) was used. These models were used as backbones for visual and acoustic feature extraction in conjunction with the joint audio-visual feature representation extracted by a fully connected layer. Finally, the fusion model employed three aforementioned encoders to fuse their outputs using multi-self-attention layers.

Min et al. [49] used Visual Transformer (ViT) [11] pre-trained on the facial dataset to extract facial features. Subsequently, a transformer-like model was applied to the obtained features. Additionally, the authors introduced a learning technique through random frame masking, enhancing the performance of ER models in-the-wild setting.

3. Methodology

The pipeline of the implemented emotion recognition system is schematically presented in Fig. 1. In this section, we elaborate on each element of our ER pipeline.

3.1. Acoustic Emotion Recognition System

We proposed three slightly different models. The backbone of all models is based on the Public Dimensional Emotion Model (PDEM) that is the first publicly available transformer-based dimensional Speech Emotion Recognition (SER) model [64] and designed for predicting arousal, valence, and dominance characteristics. The PDEM builds upon the pre-trained wav2vec2-large-robust model, which is one of the variants of Wav2Vec 2.0 [1].

On top of each model, we stack two GRU layers with 256 neurons (AudioModelV1) or two transformer layers with self-attention mechanisms, each with 32 or 16 heads (AudioModelV2 and AudioModelV3). After the last transformer layer, we aggregate the information along the temporal axis using 1D Convolutional Neural Networks (CNN) and apply two subsequent Fully Connected Layer (FCL) for feature compression and prediction generation. We fine-tune all the layers from the top to the last two (AudioModelV1 and AudioModelV2) or four (AudioModelV3) encoding layers of the backbone model.

3.2. Visual Emotion Recognition System

Visual modality is the most important one in Affective Computing as the human face and body express an immense amount of affective information. Therefore, we tried to use as much available visual information as possible.

To obtain the visual ER system, we have done several steps. First of all, we selected several efficient frame-level

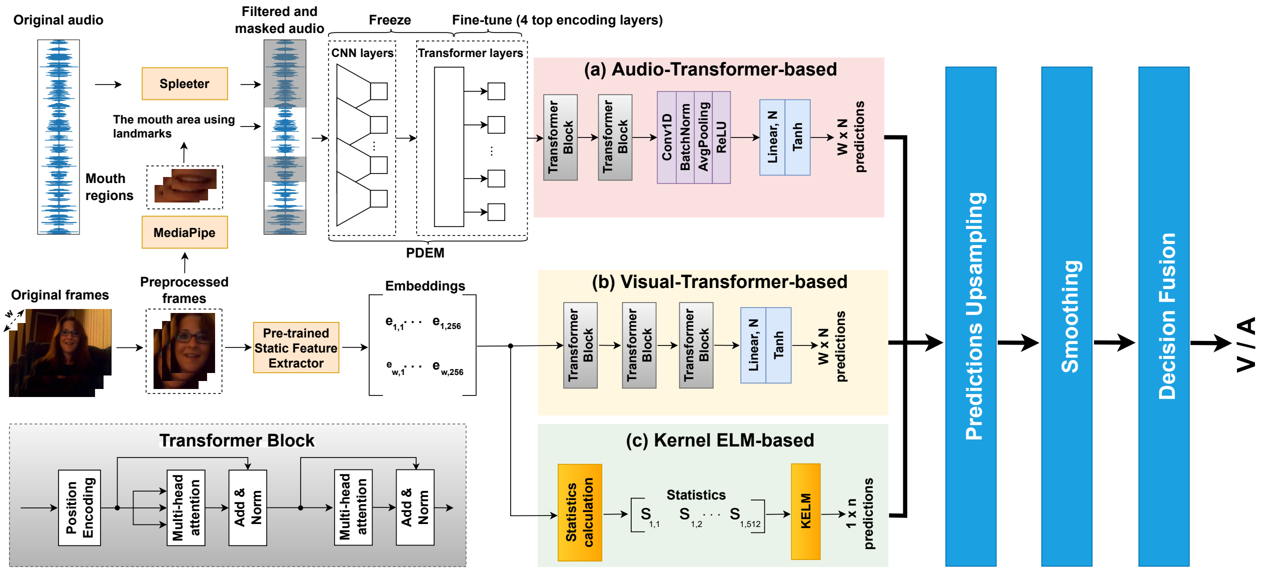

Figure 1. Pipeline of the developed ER system: (a) the Audio-Transformer-based dynamic ER system, (b) the Visual-Transformer-based dynamic ER system, and (c) – the Kernel ELM-based ER system. W – the temporal window size (in the number of frames), N – the number of neurons in the decision-making head (2 for regression task).

static models (EfficientNet-B1 [59], -B4, ViT-B-16 [11], and HRNet [58]), modified them and pre-trained on the data introduced in Sec. 4.1. Next, to further enhance the efficacy and robustness of the static models, they were fine-tuned on the AffWild2 dataset. Finally, the fine-tuned static models have been frozen and used as feature extractors that provide valuable affective features for consecutive temporal aggregation within the visual dynamic ER model. In the next subsections, we provide a detailed description of both static and dynamic visual ER systems.

3.2.1 Static Models

To construct an accurate ER model, especially for the visual data, a robust and efficient feature extractor is needed. We call such models *static* since they are trained on frame-level data and provide emotion predictions per-frame, ignoring the temporal context. In the context of ER, such state-of-the-art models are based either on CNN or recently introduced ViT neural network architectures. We experimented with both approaches, as various models can demonstrate different performances given in-the-wild nature of the data. Specifically, we employed the EfficientNet [59] (B1 and B4 versions, comprising 7.8M and 19.3M parameters, respectively) and Visual Transformer-B-16 [11] architectures that are pre-trained on ImageNet [9, 54]. Additionally, to process the body language and gestures, we adapted the HRNet [58] model that is pre-trained on COCO [44] dataset.

However, before the pre-training of those models on various ER datasets (including fine-tuning on AffWild2), we have slightly modified them as depicted in Fig. 2. Thus, we removed the last layer responsible for the classification and stacked on top of it several new layers responsible for the prediction. In case of HRNet, apart from the regression head, several ResNet-like layers have been added to process the heatmaps produced by the HRNet backbone. As a final activation function, we utilized the *Tanh*.

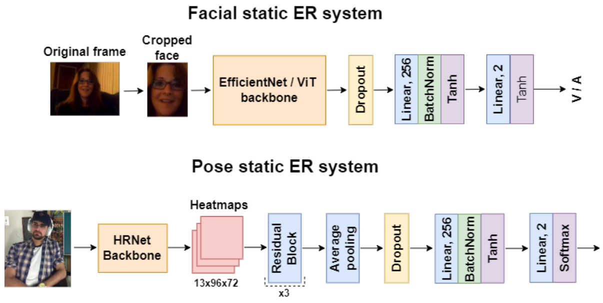

Figure 2. The NN architecture of modified frame-level ER models.

We should note that, since we fixed the number of embeddings by modifying respective static models, the feature extractors always output 256 features per frame.

3.2.2 Dynamic Models

It is well-known that emotions are temporal phenomena that last for a certain period of time. To exploit this aspect, we developed dynamic ER models that take into account a temporal context during the decision-making. An overview of these model architectures is given in Fig. 1 (b and c).

In the ER literature, there are many different approaches for temporal information aggregation, including functionals-aggregation (calculation of statistics over a period of time), recurrent neural networks [12] (RNNs), and recently introduced Transformer-based architectures [5, 43, 63]. Although RNNs have been most popular in ER domain so far, the Transformers-based architectures are taking the lead in the last years.

To leverage the most effective architectures, we employed the Transformer-based temporal aggregation method as well. The implemented ER dynamic model is schematically depicted in Fig. 1 (c). Thus, the dynamic model consists of a static feature extractor and the temporal part of three consecutive Transformer-encoder layers inspired by [62]. Lastly, the regression head completes the decision-making process.

For the comparison and as an alternative, we developed a simpler temporal aggregation method: the statistical-based model that calculates functionals over a fixed period. Here, based on former research [12], we fix the analysis window to 2 seconds. We apply mean, minimum, and maximum functional statistics to non-overlapping 2-second windows and utilize the Kernel Extreme Learning Machine (KELM) [19]. KELM aims to solve a regularized least squares regression problem between a kernel (instance similarity) matrix $\mathbf{K}$ and a target vector (or matrix) $\mathbf{T}$ from the training dataset, and hence is very fast to train given the kernel. The set of weights (β) in KELM is calculated via:

$$\beta = (\mathbf{I}/C + \mathbf{K})^{-1}\mathbf{T}, \tag{1}$$

where $\mathbf{I}$ is the identity matrix, and C is the regularization coefficient optimized via cross-validation on the challenge development set. The prediction for a test instance x is obtained via $\hat{y} = K(\mathbf{D}, x)\beta$, where $K()$ and $\mathbf{D}$ denote the kernel function and the training dataset, respectively.

3.3. Fusion Schemes

Fusion, particularly to leverage multi-modal information, is an important stage in ER systems. Here, we experimented with late and model-based fusion strategies. In the latter, the features from audio and video models are combined through trainable cross-attention mechanism, complementary leveraging the strengths of every modality.

For late fusion, we experimented with two schemes. First, we used Dirichlet-based Random Weighted Fusion (DWF), where fusion matrices containing weights per model-VA combination are randomly sampled from the Dirichlet distribution. A large pool of such matrices is generated and the best one in terms of the challenge measure is selected for the test set submission. This approach is shown to generalize well to in-the-wild data [12, 23].

The second decision fusion approach is based on Random Forests (RF) [4], where the concatenated probability vectors from the base models are stacked to RF as in [22]. To avoid over-fitting, out-of-bag predictions are probed to optimize the number of trees.

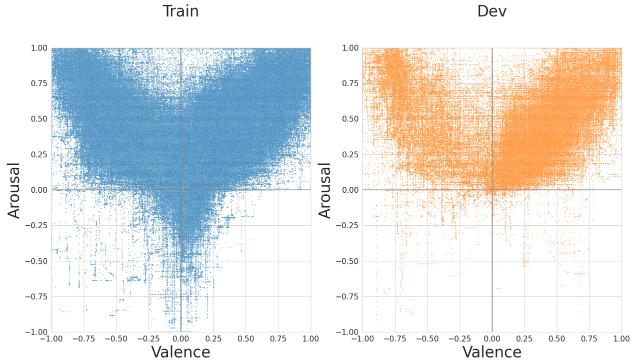

Figure 3. VA distribution of the AffWild2 train and dev sets.

4. Experimental Setup

4.1. Experimental Data

For all experiments presented in this work, the AffWild2 dataset and corresponding labels from the 6th ABAW challenge [36] were used. The AffWild2 dataset is an audio-visual in-the-wild corpus, that serves as a comprehensive benchmark for multiple affective behavior analysis tasks. Comprising 594 videos with approximately 3M frames from 584 subjects, it is annotated in terms of Valence and Arousal emotional continuous labels in the [-1, 1] range. Additionally, a subset of 548 videos is annotated for expression recognition across 8 emotional classes.

We should note, however, that the Valence-Arousal annotations are not evenly distributed as depicted in Fig. 3. As we can see, the labels have a high bias towards the positive value of Arousal, posing additional challenges for the deep learning models' training.

To pre-train our visual static ER models, we used a range of publicly available Facial Expression Recognition (FER) datasets. Those datasets were combined into one large mixed dataset that has been used for the pre-training. Train, development, and test splitting have been done in a speaker-independent way. As labels, Ekman's six basic emotions were selected from the aforementioned datasets along with Valence and Arousal values. We summarized the information about all used pre-training corpora in Tab. 1.

After the pre-training phase, the static ER models have been fine-tuned on the AffWild2 dataset.

4.2. Data Preprocessing

4.2.1 Audio

Before training an audio model, in addition to extracting audio signals from multimedia files, we perform voice activ-

Table 1. Summary of pre-training corpora used in this work (only the volume of data utilized in this work is presented).

Dataset	Modality	Data volume	Annotations	Conditions
RECOLA [53]	Audio, Visual	3:50 hours	A, V	Lab.
SEWA [38]	Audio, Visual	9:10 hours	A, V	In-the-wild
SEMAINE [47]	Audio, Visual	6:30 hours	A, V	Lab.
AFEW-VA [37]	Visual	30,000 images	A, V	In-the-wild
AffectNet [50]	Visual	420,299 images	C, A, V	In-the-wild
SAVEE [21]	Audio, Visual	≈24 minutes	C	Lab.
EMOTIC [39]	Visual	23,571 images	C, A, V	In-the-wild
ExpW [72]	Visual	91,793 images	C	In-the-wild
FER+ [15]	Visual	35,887 images	C	In-the-wild
RAF-DB [42]	Visual	29,672 images	C	In-the-wild

A – Arousal, V – Valence, C – Categories, Lab. - Laboratory conditions

ity detection. Due to the specific nature of the acoustic data provided by the ABAW 2024 challenge organizers, audio data may include background noise and multiple speakers, making it difficult to identify the target speaker. Therefore, methods based only on audio analysis are not suitable for this dataset. That is why, for the appropriate usage of audio modality, we rely on video modality by analyzing the visual data frame by frame. For this purpose, facial landmarks are extracted using the MediaPipe framework [46]. Then, mouth landmarks are detected, and the corresponding region of interest is extracted. Obtained information is used to determine whether the target speaker's mouth is open or closed. In parallel, we separate the speech signal from noise (including music) using Spleeter by deezer.[3]

Next, 4-second windows with a step of two seconds are formed on the filtered voice segments. We downsampled the annotations to 5 Frames Per Second (FPS) for all videos.

To enhance the generalizability of the audio models, we employ several augmentation techniques, including polarity inversion, the addition of white noise, or variation in audio volume. These techniques help to reduce the confidence level of the models in their emotion predictions.

4.2.2 Video

Depending on the model type (static or dynamic), several preprocessing steps have been applied as depicted in Fig. 4. For the static models, we first detect faces and crop them, adding 15 pixels to all bounding box boundaries to include the chin and other human facial features. We have utilized the RetinaFace model [8, 10] based on the MobileNet [17] architecture, namely the *MobileNet-0.25* version. We utilized this model because it is one of the most effective face recognition models known nowadays, yet very computationally efficient, since it has only around 1.7 million parameters. The next step in the static data preprocessing pipeline is to resize the image and normalize the pixel values. In this work, we employed the pre-defined image resolutions and normalization values provided by the authors of corresponding models (EfficientNet [59], ViT-B-16 [11], and HRNet [58]). Finally, to improve the performance and robustness of the deep learning models, the following data augmentation techniques were applied: random image padding, changing of brightness, contrast, saturation, and hue of the image, Gaussian noise addition, random rotation, cropping, image posterization, changing of sharpness, equalization, and flipping. All augmentations were applied to every image with probability of 0.05, resulting in approximately 46% of images augmented every training epoch.

The data preprocessing for dynamic models closely mirrors the static methodology except for one important step: we use additional normalization applied to embeddings to avoid the gradient explosion that can arise in early stages of training. Two different normalization methods were tried: MinMax and Per-Video-MinMax scalings. The difference in methods is that MinMax scaling computes the corresponding min and max values across the whole training set (and then applied to every instance), while the Per-video-MinMax scaling does it for every video separately, applying normalization values only within the corresponding video.

It is well-known that Transformer-based models can operate with sequences of arbitrary length. However, different FPS of videos and varying lengths of the sequences can significantly harm the training process. Therefore, to stabilize it and ensure convergence, we downsampled all videos to 5 FPS and fixed the window length during training. It is experimentally shown that the size of a temporal window can significantly influence the efficacy of the ER model [12]. That is why we have done experiments with different temporal context lengths, namely: 1, 2, 3, 4, 6, and 8 seconds. Finally, the model with the highest CCC score was used for test set submissions.

[3] https://github.com/deezer/spleeter

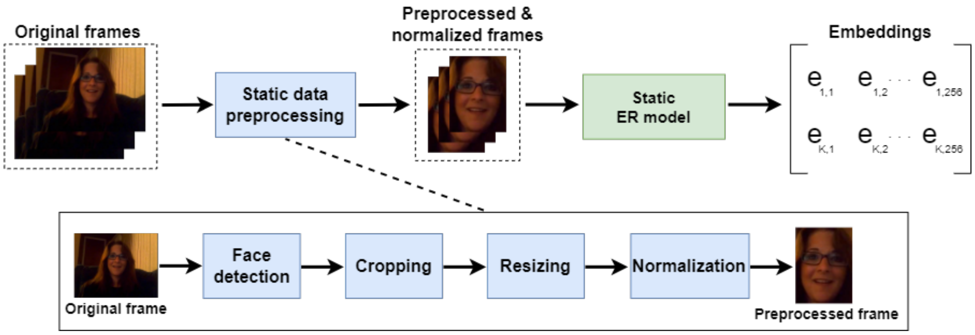

Figure 4. Pipeline of preprocessing of the video data for dynamic emotion recognition modeling. Note that in the case of the pose static model, face detection is not applied.

4.3. Post-processing

After getting Valence and Arousal values for every frame, several post-processing steps were applied. Since dynamic models have been trained using reduced FPS, we first upsampled the models' predictions to align them with the actual video FPS on the development and test sets. We used linear interpolation, filling in missing values between two consecutive predictions. After interpolation, the upsampled values were smoothed using Hamming window [61]. The size of the window has been chosen to be 0.5 seconds.

4.4. Training Hyperparameters

For the training of both static and dynamic ER models, we used an AdamW [45] optimizer with the learning rate (LR) set to 0.005. Moreover, a linear LR warmup was used with a starting value of 0.00005 for the first 100 training steps to avoid the gradient explosion. Additionally, the cyclic LR scheduler was applied with a minimum LR value of 0.0001 and an annealing period of 5. We set the early stopping number of epochs to 10 epochs.

For static models, we also applied two various fine-tuning techniques called discriminative learning and gradual unfreezing. Discriminative learning is a training technique that fine-tunes pre-trained neural networks by setting different learning rates for each layer, helping to increase model performance on specific tasks. The main idea is that earlier layers extract more general features and should undergo minimal changes, while deeper layers are more task-specific and require significant adjustments, so learning rates are gradually decreased from deep to early layers. We applied a 0.9 factor for every LR of consecutive layers starting from newly initialized ones. On the other hand, gradual unfreezing is a technique that unfreezes Deep Neural Networks (DNN) layers over training epochs to prevent overfitting or knowledge loss due to large initial gradients. Typically, only the added layers are unfrozen first, with more layers (often in blocks) being unfrozen in subsequent epochs, allowing the model to adjust gradually to the target task. We should note that we tried all possible combinations of those techniques for every static ER model.

Table 2. Best development set CCC results per acoustic base-model on the VA challenge.

Model	Valence	Arousal	Avg.
AudioModelV3	0.290	0.400	**0.345**
AudioModelV1	0.282	0.377	0.329
AudioModelV2	0.241	0.375	0.308

For the visual Transformer-based ER models, we set the number of heads equal to 8 and dropout to 0.1. Additionally, the positional encoding employed in [62] is applied to embeddings.

5. Experimental Results

The challenge measure for the Valence-Arousal Estimation challenge is set to be the Concordance Correlation Coefficient (CCC). CCC is recently popularly used in regression tasks over the *Pearson's Correlation* (PC), as it also considers the difference in means [41]:

$$CCC = \frac{2 \cdot \sigma_{t,p}}{\sigma_t^2 + \sigma_p^2 + (\mu_t - \mu_p)^2}, \qquad (2)$$

where μ_t and μ_p denote the averaged ground truth and predicted scores for all test clips, respectively; σ_t and σ_p denote the respective standard deviations; $\sigma_{t,p}$ is the covariance between t and p.

In the next subsections, we report the results obtained for each modality, concluding with the recognition performance of the multi-modal systems submitted for the test part of the VA challenge.

5.1. Audio-based Models

For the acoustic modality, we obtained three different models via fine-tuning of the modifications of the PDEM model. All of the top approaches used data augmentation and

Spleeter for background noise separation. The best results on the development set for the VA challenge are presented in Tab. 2. Note that the results here are reported for 4-second windows excluding the silent (absence of voice) segments, rather than frame-wise over which the ground truth annotations are provided. As we can see, the best performance is demonstrated by the AudioModelV3 for both Valence and Arousal. Therefore, we utilized this model to generate the test set predictions for decision-level fusion.

5.2. Video-based Models

For the video modality, as described in Sec. 3.2, several training steps were implemented. First, we pre-trained various static visual models (EfficientNet-B1, EfficientNet-B4, ViT, and HRNet) on a large amount of ER data and then fine-tuned them on the AffWild2 dataset. As we trained models in static mode, ignoring the temporal axis, we have chosen the LogCosh (Eq. (3)) as the loss function due to its statistical properties [56] and the Root Mean Squared Error (RMSE) as the development measure. The LogCosh loss $L(y, \hat{y})$ between the ground truth y and the prediction set $\hat{y}$ is defined as:

$$L(y, \hat{y}) = \sum_{i=1}^{n} log(cosh(\hat{y}_i - y_i)). \tag{3}$$

All combinations of applying gradual unfreezing and discriminative learning techniques were experimented. Tab. 3 shows the best results per deep learning model used. As we can see, the best fine-tuning techniques for every static model type turned out to be the combination of discriminative learning and gradual unfreezing. This can be due to the fact that both EfficientNets and ViT are already pre-trained on many ER datasets, providing them with strong initial representations. Such a combination of fine-tuning techniques enables the smooth usage of those representations, while simultaneously avoiding gradient explosion and overfitting. One more interesting finding is that the most compact model has shown the best recognition performance, pointing out that it is not always necessary to have the heaviest model in such domains as Affective Computing. Thus, we have chosen the trained EfficientNet-B1 model as the visual feature extractor for further experiments.

Next, we experimented with Functionals-based approach. In this model, the extracted by EfficientNet-B1 model embeddings are summarized over 2-second non-overlapping windows to make a single prediction for the whole window. Such an approach demonstrated decent results on the VA Estimation challenge (see Tab. 4). Even though the combination of suprasegmental features summarized using min and mean functionals performed slightly better on the development set, we opted to train the VA prediction model using the combination of mean, min, and max functionals as it performed significantly better in another (EXPR) ABAW'24 challenge. Using this combination of video features with KELM, the best development set CCC of 0.489 (average over two dimensions) was obtained, with corresponding CCC performances of 0.398 and 0.581 for Valence and Arousal, respectively.

Table 3. Best development set RMSE results for visual static ER models.

Model	DL	GU	Valence	Arousal	Avg.
EN-B1	+	+	0.3593	0.2392	0.2993
EN-B4	+	+	0.3611	0.2387	0.2999
ViT-B-16	+	+	0.3805	0.2516	0.3111
HRNet	N/A	N/A	0.4173	0.2804	0.3489

DL – Discriminative Learning, **GU** – Gradual Unfreezing **EN** – EfficientNet, **N/A** – Not Applicable

Table 4. Best development set CCC results per functionals combination using KELM with EN-B1 embeddings on the VA challenge.

Functionals	Valence	Arousal	Avg.
mean, min, max	0.398	0.581	0.489
mean, max	0.393	0.583	0.488
mean, min	0.411	0.580	0.495

Extensive experiments with the uni-modal visual dynamic models (Fig. 1 (b)) showed the sensitivity of these models to extracted embeddings. Interestingly, the dynamic model based on the embeddings extracted by HRNet (body language) demonstrated development CCC scores of 0.0984 and 0.0415 for Valence and Arousal, respectively. Such results can indicate poor generalization ability of those features, especially in the temporal context, and can be caused by the frequent disappearance of participants' bodies in the course of the video. On the contrary, the best E2E facial-based model (Fig. 1 (b), EfficientNet-B1 embeddings) reached an average CCC performance of 0.574, with corresponding arousal and valence CCC scores of 0.626 and 0.523, demonstrating high robustness and generalization.

Thus, as the visual uni-modal components of the final multi-modal ER pipeline, we employed the top performing End-to-End (E2E) and functional-based ER systems, specifically: (1) the Kernel ELM based on mean, minimum, and maximum functionals, and (2) the face-based E2E model based on EfficientNet-B1 embeddings.

5.3. Multi-modal Models and Test Submissions

We selected the best-performing audio and visual models, extracted embeddings, and experimented with the fusion of modality-specific features based on a cross-attention mechanism inspired by [2, 60]. Surprisingly, this approach could not outperform the face-based E2E system on the development set. Therefore, instead of intermediate fusion, we decided to use the decision-based fusion schemes described

Table 5. Development and Test set CCC performances of the submitted systems for the VA challenge.

			Development			Test		
Sys #	Modality	Method	Valence	Arousal	Avg.	Valence	Arousal	Avg.
1	Visual	Face-based E2E model	0.523	0.626	0.574	**0.5355**	**0.5861**	**0.5608**
2	Audio-visual	DWF (Sys1 + Audio)	0.532	0.649	0.591	0.5307	0.5792	0.5549
3	Audio-visual	DWF (Sys1 + Audio + Func.)	0.528	0.641	0.584	0.5259	0.5624	0.5441
4	Audio-visual	RF (Sys1 + Audio + Func.)	0.738	0.787	0.763	0.4014	0.4635	0.4324

DWF – Dirichlet-based Random Weighted Fusion, **RF** – Random Forest-based fusion, and **Func.** – Functionals of visual embeddings fed to Kernel ELM.

earlier. For our test set probes, we used one uni-modal (best face-based E2E system) and three multi-modal systems, based on the development set performances.

Tab. 5 reports the challenge development and test set performances of the VA challenge. We observe an interesting pattern concerning the performance of the E2E uni-modal visual model and the multi-modal models. Overall, contrary to our expectations, the multi-modal systems that show better performance over the video-only model do not generalize well to the test set. The E2E dynamic visual model used in the first submission not only yields the best test set performance among the four submissions, but also has the most similar development and test set average CCC scores.

The inclusion of the audio modality improves the arousal prediction performance on the development set, however, reduces the test set performance on the same emotion primitive. This may partly be attributed to the covariance shift and the label distribution gap between training, validation, and the tests. The ML learns not only the mapping from the inputs to the outputs, but also the pattern of the output distribution, bringing major challenges in cross-corpus or "in-the-wild" acoustic ER [14]. The VA covariance structure depicted in Figure 3 tells us that while the majority of the instances on both the training and validation sets have positive arousal, the validation set has lower negative arousal samples. This may be further exacerbated by the gap in the covariance structure of the test set labels, which are currently not accessible by the competitors.

The test set CCC performance of the top competitors along with our system and the baseline system is shown in Table 6. Among the 60 teams that participated, 23 made submissions and 10 of them surpassed the baseline. On the valence prediction task, we rank third, and on the overall (average of arousal and valence) our system ranks fourth. It is important to remind that this performance is reached via a single E2E dynamic visual model.

Table 6. Comparison of test set CCC scores of top systems in the ABAW 2024 Competition and our work.

System	Valence	Arousal	Avg.
Zhang et al. [71]	0.6873	0.6569	0.6721
Praveen and Alam [51]	0.5418	0.6196	0.5807
Zhou et al. [73]	0.5223	0.6057	0.5640
Our contribution [13]	0.5355	0.5861	0.5608
Yu et al. [68]	0.5208	0.5748	0.5478
Savchenko [57]	0.4925	0.5461	0.5193
Kim et al. [24]	0.4836	0.5318	0.5077
Waligora et al. [65]	0.4198	0.4669	0.4434
CAS-MAIS Team	0.4245	0.3414	0.3830
Min et al. [49]	0.2912	0.2456	0.2684
Baseline [36]	0.2110	0.1910	0.2010

6. Conclusion and Future Work

The results of our research highlight the potential of deep learning models for audio-visual emotion recognition in unconstrained, "in-the-wild" settings. The face-based end-to-end dynamic models, leveraging salient embeddings extracted by the EfficientNet-B1 model, achieved a competitive efficacy, outperforming the traditional functional-based approaches. However, optimizing video models still remains computationally very demanding, posing additional challenges in deploying these solutions for "in-the-wild" scenarios. While our experiments on the development set suggested that combining audio and video modalities through fusion techniques could significantly enhance performance, the acoustic modality's arousal prediction performance did not generalize well to the test set. Improving the robustness of the acoustic model for more accurate arousal prediction as well as usage of other contextual information (background sound, linguistics, etc.) will constitute our future work. Ultimately, progress in this field holds promise for enabling the naturality of human-computer interaction.

7. Acknowledgements

This research was partially supported by RSF in the framework of the project No. 22-11-00321 (M. Markitantov and A. Karpov)

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

[2] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, 2018. Association for Computational Linguistics. 7

[3] Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International conference on machine learning*, pages 573–582. PMLR, 2019. 2

[4] Leo Breiman. Bagging predictors. *Machine learning*, 24: 123–140, 1996. 4

[5] Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna, and Pier Luigi Mazzeo. Vitfer: facial emotion recognition with vision transformers. *Applied System Innovation*, 5(4):80, 2022. 4

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725, 2020. 1, 2

[7] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition*, pages 428–438, Cham, 2018. Springer International Publishing. 2

[8] Jiankang Deng, Jia Guo, Y Zhou, J Yu, I Kotsia, and S Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. arxiv 2019. *arXiv preprint arXiv:1905.00641*, 1905. 5

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[10] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5202–5211, 2020. 5

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 3, 5

[12] Denis Dresvyanskiy, Elena Ryumina, Heysem Kaya, Maxim Markitantov, Alexey Karpov, and Wolfgang Minker. End-to-end modeling and transfer learning for audiovisual emotion recognition in-the-wild. *Multimodal Technologies and Interaction*, 6(2), 2022. 4, 5

[13] Denis Dresvyanskiy, Maxim Markitantov, Jiawei Yu, Peitong Li, Heysem Kaya, and Alexey Karpov. Sun team's contribution to abaw 2024 competition: Audio-visual valence-arousal estimation and expression recognition. *arXiv preprint arXiv:2403.12609*, 2024. 2, 8

[14] Dmitrii Fedotov, Heysem Kaya, and Alexey Karpov. Context modeling for cross-corpus dimensional acoustic emotion recognition: Challenges and mixup. In *Speech and Computer*, pages 155–165, Cham, 2018. Springer International Publishing. 8

[15] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013. 2, 5

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021. 1, 2

[17] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. 5

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 2

[19] Guang-Bin Huang, Hongming Zhou, Xiaojian Ding, and Rui Zhang. Extreme Learning Machine for Regression and Multiclass Classification. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(2):513–529, 2012. 4

[20] Min Huang, Huimin Qian, Yi Han, and Wenbo Xiang. R(2+1)D-based Two-stream CNN for Human Activities Recognition in Videos. In *2021 40th Chinese Control Conference (CCC)*, pages 7932–7937, 2021. 2

[21] Philip Jackson and SJUoSG Haq. Surrey audio-visual expressed emotion (SAVEE) database. *University of Surrey: Guildford, UK*, 2014. 5

[22] Heysem Kaya, Furkan Gurpinar, and Albert Ali Salah. Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 4

[23] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 65:66–75, 2017. Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing. 4

[24] Junhwa Kim, Namho Kim, Minsoo Hong, and Cheesun Won. CCA-Transformer: Cascaded cross-attention based transformer for facial analysis in multi-modal data. *Available Online*, 2024. 2, 8

[25] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 2328–2336, 2022. 1

[26] Dimitrios Kollias. Abaw: learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2023. 1

[27] Dimitrios Kollias. Multi-label compound expression recognition: C-expr database & network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2023. 1

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

[29] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 1

[30] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 1

[31] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 1

[32] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019. 1

[33] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800, 2020. 1

[34] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 1

[35] 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. 1

[36] 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. 1, 4, 8

[37] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017. 5

[38] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller, et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1022–1040, 2019. 5

[39] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. EMOTIC: Emotions in context dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 5

[40] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. Two-stream aural-visual affect analysis in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 600–605. IEEE, 2020. 2

[41] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989. 6

[42] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[43] Zheng Lian, Bin Liu, and Jianhua Tao. Ctnet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:985–1000, 2021. 4

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 3

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 6

[46] 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. 5

[47] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2011. 5

[48] Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2022. 2

[49] Seongjae Min, Junseok Yang, Sangjun Lim, Junyong Lee, Sangwon Lee, and Sejoon Lim. Emotion recognition using transformers with masked learning. *arXiv preprint arXiv:2403.13731*, 2024. 2, 8

[50] 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. 5

[51] R Gnana Praveen and Jahangir Alam. Recursive cross-modal attention for multimodal fusion in dimensional emotion recognition. *arXiv preprint arXiv:2403.13659*, 2024. 2, 8

[52] Andreas Psaroudakis and Dimitrios Kollias. Mixaugment & mixup: Augmentation methods for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2367–2375, 2022. 1

[53] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2013. 5

[54] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 3

[55] Elena Ryumina, Maxim Markitantov, Dmitry Ryumin, Heysem Kaya, and Alexey Karpov. Zero-shot audio-visual compound expression recognition method based on emotion probability fusion. *CVPRW*, page in print, 2024. 1

[56] Resve A. Saleh and A. K. Md. Ehsanes Saleh. Statistical properties of the log-cosh loss function used in machine learning, 2024. 7

[57] Andrey V. Savchenko. Hsemotion team at the 6th abaw competition: Facial expressions, valence-arousal and emotion intensity prediction. *arXiv preprint arXiv:2403.11590*, 2024. 2, 8

[58] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5696, 2019. 3, 5

[59] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 6105–6114. PMLR, 2019. 3, 5

[60] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2019:6558–6569, 2019. 7

[61] John W Tukey. *The measurement of power spectra: from the point of view of communications engineering*. Dover, 1958. 6

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 4, 6

[63] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10745–10759, 2023. 4

[64] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[65] 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. 2, 8

[66] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: a simple framework for masked image modeling. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9643–9653, 2022. 2

[67] Zhiwei Yang, Peng Wu, Jing Liu, and Xiaotao Liu. Dynamic local aggregation network with adaptive clusterer for anomaly detection. In *European Conference on Computer Vision*, pages 404–421. Springer, 2022. 2

[68] Jun Yu, Gongpeng Zhao, Yongqi Wang, Zhihong Wei, Yang Zheng, Zerui Zhang, Zhongpeng Cai, Guochen Xie, Jichao Zhu, and Wangyuan Zhu. Multimodal fusion method with spatiotemporal sequences and relationship learning for valence-arousal estimation. *arXiv preprint arXiv:2403.12425*, 2024. 2, 8

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

[70] 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), 2023. 2

[71] 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. 1, 2, 8

[72] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126:550–569, 2018. 5

[73] 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. 2, 8