Article

End-to-End Modeling and Transfer Learning for Audiovisual Emotion Recognition in-the-Wild

Denis Dresvyanskiy [1,2,*,†], Elena Ryumina [3,†] , Heysem Kaya [4] , Maxim Markitantov [3], Alexey Karpov [3]
and Wolfgang Minker [1]

1 Dialogue Group, Institute of Communications Engineering, Ulm University, 89081 Ulm, Germany;
  wolfgang.minker@uni-ulm.de
2 Information Technologies and Programming Faculty, ITMO University, 197101 St. Petersburg, Russia
3 St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences,
  St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS),
  199178 St. Petersburg, Russia; ryumina_ev@mail.ru (E.R.); m.markitantov@yandex.ru (M.M.);
  karpov@iias.spb.su (A.K.)
4 Department of Information and Computing Sciences, Utrecht University, 3584 CC Utrecht, The Netherlands;
  h.kaya@uu.nl
* Correspondence: denis.dresvyanskiy@uni-ulm.de
† These authors contributed equally to this work.

Abstract: As emotions play a central role in human communication, automatic emotion recognition has attracted increasing attention in the last two decades. While multimodal systems enjoy high performances on lab-controlled data, they are still far from providing ecological validity on non-lab-controlled, namely "in-the-wild" data. This work investigates audiovisual deep learning approaches to emotion recognition in in-the-wild problem. Inspired by the outstanding performance of end-to-end and transfer learning techniques, we explored the effectiveness of architectures in which a modality-specific Convolutional Neural Network (CNN) is followed by a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) using the AffWild2 dataset under the Affective Behavior Analysis in-the-Wild (ABAW) challenge protocol. We deployed unimodal end-to-end and transfer learning approaches within a multimodal fusion system, which generated final predictions using a weighted score fusion scheme. Exploiting the proposed deep-learning-based multimodal system, we reached a test set challenge performance measure of 48.1% on the ABAW 2020 Facial Expressions challenge, which advances the first-runner-up performance.

1. Introduction

Emotions play a vital role in daily human–human interactions [1]. Automated recognition of emotions from multimodal signals has attracted increasing attention in the last two decades with applications in domains ranging from intelligent call centers [2,3] to intelligent tutoring systems [4–6]. Emotion recognition is studied in the broader affective computing field, where the research of natural emotions is the focal point. Research in this domain is shifting to "in-the-wild" conditions, namely away from lab-controlled studies. This is due to the availability of new and challenging datasets collected and introduced in competitions such as Affective Facial Expressions in-the-Wild (AFEW) [7,8] and Affective Behavior Analysis in-the-Wild (ABAW) [9–13]. Considering the challenging nature of the data, e.g., background noise in audio, cluttered background, and pose variations in video, benefiting from multiple modalities including, but not limited to acoustics, vision (face and body pose), physiological signals, and linguistics is essential [14].

Maturing over the last decade, earlier approaches to audiovisual emotion recognition included hand-crafted acoustic and visual features, which are then fed to classifiers that can handle high-dimensional feature vectors such as Support Vector Machines (SVMs). In acoustic emotion recognition, extracting Low-Level Descriptors (LLD) and summarizing them over short (1–4 s) chunks of audio using statistical functionals have been proven to be successful in a range of paralinguistics tasks [15,16] and have been popularly used in the INTERSPEECH Computational Paralinguistics Challenge series since 2009 [17–19]. This scheme was later followed by clustering-based LLD summarization approaches such as Bag-of-Audio-Words (BoAW) [20,21] and Fisher-Vectors (FVs) [22–25]. The BoAW approach was inspired by its counterpart in the linguistics domain, where the set of all words in the training set after preprocessing (e.g., stemming) is first used to form a "bag" of N-words, and then, each document is represented as an N-dimensional histogram of word occurrences. In audio and video BoW representation, however, the LLDs from the training set are first clustered using K-means or Gaussian Mixture Models (GMMs), and then, the LLDs are assigned to the nearest cluster for a fixed-length, suprasegmental representation. While the BoAW approach computes the zeroth-order statistics, the FV approach, which was originally introduced in the vision domain [26], calculates the change in the underlying model (usually the GMM) parameters with respect to new coming data, thus also including the first- and the second-order statistics. With the popularity of Deep Learning (DL) and transfer learning, state-of-the-art systems in speech emotion recognition and paralinguistics benefit from extracting features from pretrained models [27,28] and deploying end-to-end models [29–32].

In vision-based emotion recognition, hand-crafted features included Local Binary Patterns (LBPs) [33], Histograms of Oriented Gradients (HOGs) [34], which are still being popularly used [35–37]. For video (spatio-temporal) emotion recognition, extensions of these visual descriptors, such as Local Gabor Binary Patterns from Tree Orthogonal Planes (LGBPs-TOPs), were proposed [38], and recently have been successfully employed in combination with deep representations [39–42]. Given sufficient data and/or pretrained models in a relevant task, state-of-the-art systems in both unimodal and multimodal emotion recognition heavily deploy DL, in particular Convolutional Neural Networks (CNNs) and/or Recurrent Neural Networks (RNNs) [43–47].

Motivated by these developments in multimodal emotion recognition and the recent outstanding performance of deep learning in the audio [48,49] and video [50,51] domains, as well as the performance of deep transfer learning to alleviate data scarcity in the target problem [40,52,53], in this study, we employed both deep end-to-end learning and deep transfer learning for both audio and video modalities, fusing the scores of the uni-modal subsystems for multimodal affect recognition in out-of-lab conditions. We experimented with and used the official challenge protocol for the ABAW challenge, the Facial Expressions Sub-challenge (ABAW-FER Challenge), originally run for Face and Gesture 2020, but later extended until October 2020. This sub-challenge includes Ekman's six basic emotions plus neutral, thus featuring a seven-class classification task.

We conducted extensive experiments with our system (and its components) on the AffWild2 dataset, adhering to the challenge protocol [54]. The contributions of this paper include: (1) a novel multimodal framework that leverages deep and transfer learning in the audio and video modalities; (2) analysis of the components of our proposed system; (3) comparative results on the official ABAW-FER Challenge. Here, we present, extend, and advance our contribution to the challenge, which has been ranked third in the competition and was published in arXiv [55].

The remainder of this article is organized as follows: We analyze the current state of the multimodal emotion recognition domain and one of the challenges devoted to it in Section 2. Section 3 presents a developed multimodal emotion recognition system and describes its parts. Next, in Section 4, we provide the setting used in this work, including the data utilized, the training hyperparameters, and the preprocessing features. Section 5 presents the results obtained with the proposed multimodal emotion recognition system

and its unimodal parts. In Section 6, we discuss the features of the developed system and present interesting findings during the research. Lastly, Section 7 summarizes the performed work and considers the directions of future research in multimodal emotion recognition in-the-wild.

2. Related Work

Having exhibited outstanding performance in a range of audio and image recognition tasks, such as speech and object recognition, deep learning has recently become the most dominant approach also in affect recognition. Training deep models has become increasingly popular and relatively easier, as the hunger of the deep models for massive data is met by the production of more and larger datasets related to affect recognition tasks. Moreover, deep learning models are developing not only in "depth", but also in "breadth": models process different modalities (e.g., audio, visual) at lower layers with modality-specific filters and at different sampling frequencies, which are subsequently combined at a higher abstraction layer via different fusion techniques.

Initial research studies about the application of different machine learning techniques to the FER started to appear at the beginning of the second millennium [56,57]. The second explosion of emotion research started with the investigation of deep learning, thanks to the availability of pretrained deep CNNs such as VGG16 [58] and ResNet50 [59]. While new developments in the machine learning and deep learning fields have led to significant improvements, still many research problems remain open including, but not limited to the alignment of heterogeneous signals (audio, video), handling small and imbalanced datasets, ensuring the reliability of subjective annotations, and handling data recorded in naturalistic conditions [60,61].

2.1. Multimodal Emotion Recognition

Multimodal emotion recognition involves taking advantage of multiple modalities (audio, video, text, physiological, and others) that, through fusion techniques, provide a single, final prediction. The primary fusion strategies from previous studies for multimodal emotion recognition can be classified into feature-level (early) fusion, decision-level (late) fusion, and model-level fusion [62]. In addition, there is a hybrid fusion that involves the combining of the feature- and decision-level fusions [60]. In a recent review paper on multimodal affect recognition [60], the authors mentioned that the increase in the feature set may decrease the classification accuracy if the training set is not large enough, pointing to the curse of dimensionality problem. Another problem is that the features from different modalities are collected in varied time scales and hence need to be synchronized appropriately.

In the last decade, researchers have focused more on the deep neural network in the multimodal emotion recognition domain. In [63], the authors applied supervised and unsupervised techniques for feature selection, investigated early fusion, and experimented with Convolutional Deep Belief Networks (CDBNs). In [64], LSTM was used to capture the correlation between different modalities and within the modalities. Finally, the results of LSTM were used as the input of the classifier LIBSVM to make the final prediction. A similar approach based on two LSTMs instead of LIBSVM was proposed by Tzirakis P. et al. [65]. Another approach is to replace the softmax layer of each unimodal classifier with a new layer that will combine the deep embeddings of all modalities [66].

There have also been new investigations in the direction of new fusion techniques for modality merging. In [67], the authors discussed the role of speaker-exclusive models, the importance of different modalities (audio, video, text), and the generalizability (between datasets). In terms of ternary modalities, Majumder et al. [68] presented a novel feature fusion strategy that proceeds in a hierarchical fashion, first fusing the modalities one by one and only then fusing all three modalities. In [69], novel attention-based methods with LSTMs were proposed to fuse the modality-specific features.

Multimodal emotion recognition is still a developing area, and the task itself is far from maturity. Although there have been insights that deep learning models will be efficient for this domain, they require more diversified data for training. Challenges are organized to advance the state-of-the-art in multimodal emotion recognition in real-world environments by providing novel data and a comparable protocol. The ABAW-FER Challenge [9–13] is one such competition, whose publicly available data and common evaluation protocol were used in this paper.

2.2. ABAW-FER Challenge

In the context of the ABAW-FER Challenge, Kuhnke et al. [70] proposed a multimodal system that exploits Mel-spectrogram, appearance-based visual and facial-landmarks-based features. The last two are combined and used by a pretrained 3-D CNN, while Mel-spectrograms are fed into a slightly modified ResNet-18 Deep Neural Network (DNN). Fusing the output from the models by an additional fully connected layer, the authors reached a 0.509 Challenge Performance Measure (CPM) on the test set, taking first place in the Facial Expressions Sub-challenge in the ABAW challenge [70].

Gera, Darshan, and S. Balasubramanian implemented an attention-based framework [71], which fuses local and global attentive features and complementary context information to obtain features robust to non-trivial face positions and occlusions. The fusion process was performed on both the feature and decision levels (the loss function was calculated based on predictions from different "branches" of the framework). Thus, the authors obtained a 0.441 CPM and took second place in the ABAW-FER Challenge.

Liu et al. [72] combined ResNet and a Bidirectional Long Short-Term Memory Network (BLSTM), achieving a CPM of 0.408 on the test set. In [73], the authors explored data balancing techniques and their applications to multitask emotion recognition and found that data balancing is beneficial for classification, but not necessarily for regression. The proposed approach yielded a 0.405 CPM. Nhu-Tai Do et al. [74] proposed a temporal and statistical module to exploit and then fine-tune the face feature extraction model, obtaining a CPM of 0.389. Sachihiro Youoku et al. [75] used multiple optimized time windows (short-, middle-, and long-term) for extracting features from video data. In addition, the authors applied data balancing and fused the single-task models to further improve the prediction accuracy (CPM = 0.369).

3. Materials and Methods

Since emotion recognition is a complex multimodal paralinguistics task, we used both audio and video modalities. The final fusion of all systems was implemented as model- and class-based weighted decision-level fusion, which weights the class probabilities from each model separately. The pipeline of the fusion system is presented in Figure 1.

We hypothesized that the multimodal system for in-the-wild emotion recognition can benefit from alternative feature representations and pipelines for visual processing. The constituents of the fusion system and their optimal hyperparameters may vary depending on the target dataset. The design settings of the respective components are elaborated in Sections 3 and 5.
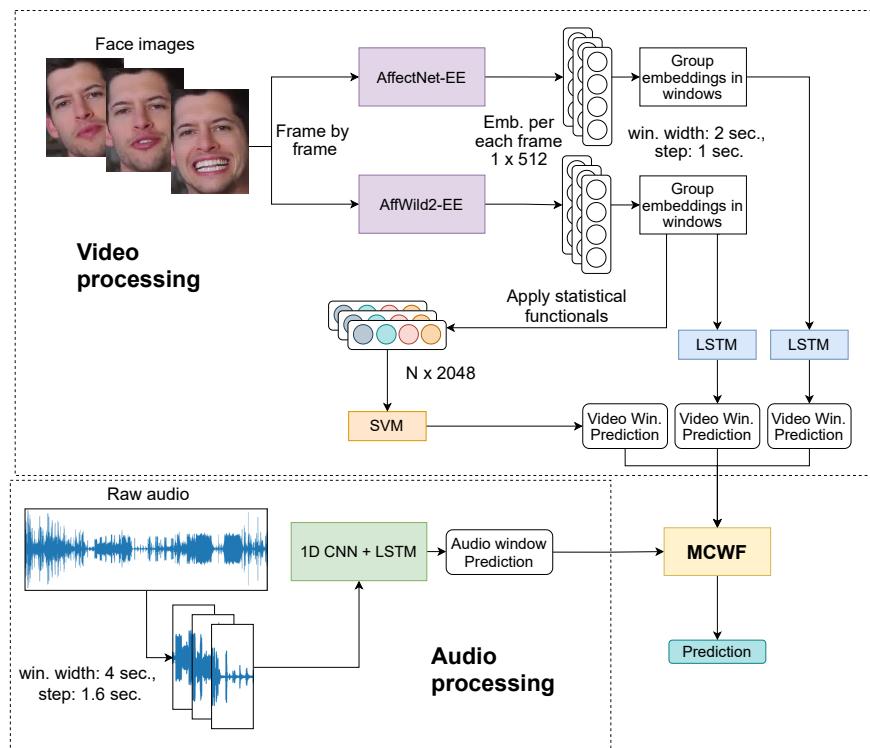
Figure 1. The proposed pipeline for the multimodal emotion recognition system: Emb.—Embeddings, Win.—Window, MCWF—Model- and Class-based Weighted Fusion.

3.1. Video-Based Deep Networks

3.1.1. Transfer Learning with VGGFace2-Based CNN

In the visual emotion recognition domain, one of the most informative body parts is the human face. Multimodal emotion recognition systems often contain a facial expression recognition model, which catches much information reflected via changes in mimics ofhuman faces. Moreover, while the audio modality can be absent (the user is silent), the face of the user is generally available for facial analysis. Therefore, the development of a robust and efficient facial expression recognition model is one of the important tasks in the construction of multimodal emotion recognition systems.

A typical choice for facial emotion recognition systems is a deep convolutional neural network, such as ResNet50. In the emotion recognition domain, it is common to use transfer learning, namely training the deep neural network on a relevant task/domain, where the annotated data are sufficiently diverse and rich. While the main idea is using the knowledge distilled in the pretrained deep network, a popular trend is to use model embeddings (deep features), which can be extracted from different layers of the deep neural network. Usually, embeddings from the last convolutional layer are exploited.

To take advantage of transfer learning, as a base model, we used the VGGFace2-model, which is the ResNet50 model pretrained on the VGGFace2 dataset [76]. The VGGFace2 dataset is intended for training models to recognize identities by faces. It contains 8631 identities in the training set, each with 363 images on average.

Based on the VGGFace2-model, we trained three facial expression recognition models, which differed in terms of data on which they were fine-tuned. We should note that for the fine-tuning, we took the VGGFace2-model, removed the last layer that provides class probabilities over the identities, and stacked two dense layers with 1024 and 7 neurons above it accordingly (hereinafter, we call this model the VGGFace2-model). The last layer with the softmax activation function allows the model to predict the probability of every emotional category. Thus, we obtained the following models:

- *VGGFace2-EE.* The VGGFace2-model fine-tuned on the AffWild2 dataset;
- *AffectNet-EE.* The VGGFace2-model fine-tuned on the AffectNet [77] and FER2013 [78] datasets (static frames);
- *AffWild2-EE.* AffectNet-EE, which was further fine-tuned on the AffWild2 dataset (dynamic frames).

3.1.2. Temporal Modeling

The expression of emotions is inherently dynamic. Some emotions cannot be successfully recognized without exploiting the dynamics [40]. Therefore, current emotion recognition systems effectively exploit different techniques to take into consideration the temporal context. These techniques include calculating the statistics of frame-level features over a time window and employing algorithms specifically developed for working with time series, such as recurrent neural networks. We implemented both techniques to investigate their recognition performance for the facial expression recognition task:

- *EE + SVM system.* First of all, embeddings for every frame were extracted by the Embedding Extractor (EE), which is one of the fine-tuned CNNs presented earlier. Next, we grouped the embeddings according to the chosen window in a sequence and applied statistical functionals, namely the mean and Standard Deviation (STD), on each sequence. This process summarizes the LLD matrix having dimensions of $N \times M$ (where N denotes the number of frames in a sequence and M is the size of the embeddings vector obtained from the EE for one image frame) into a suprasegmental feature vector with dimensions of $1 \times M * 2$ (since this vector contains the M mean and M STD statistics). Finally, an SVM classifier was trained on the obtained vectors to predict one emotion category for the whole window (sequence). The target emotion to train the SVM model was calculated using the mode (i.e., voting) of the emotion annotations in the sequence. This approach works best when the window size is close to the average duration of emotions (2–4 s based on previous research);
- *LSTM-based systems.* These systems are based on recurrent neural networks, namely on LSTM networks. As an EE subsystem, AffectNet-EE was used. For the LSTM network's training, we also grouped embeddings using a time window.
  Here, we experimented with two alternative training schemes:
  - EE + LSTM system. The extraction of the embeddings from the EE subsystem was separated from the LSTM network. Thus, the EE subsystem did not take part in the fine-tuning (training) process, and we exploited it solely for the embedding extraction;
  - E2E system. The EE subsystem was combined with the LSTM network during the fine-tuning (training) process. Thus, the system was trained as a whole, making it an End-to-End (E2E) deep neural network system. Moreover, training the EE subsystem as a part of the E2E system allowed us to generalize the EE subsystem more since it had "seen" more data including the AffWild2 dataset.

The description of the datasets used for training, training the hyperparameters, and the preprocessing procedures are elaborated in Section 4.

3.2. Audio-Based Deep Networks

3.2.1. Audio Separation

It is well known that extraneous sounds in audio recordings (e.g., background noises, music) may significantly decrease the effectiveness of the training process, which results in poorer emotion recognition performance [79]. To alleviate this, we applied Blind Source Separation (BSS) using the open-source library Spleeter [80], which contains a set of pre-trained DNN models. We used a model that allowed us to separate audio into vocals and accompaniment (all other sounds including music). Spleeter models were trained with data having 11 kHz and 16 kHz sampling rates. Therefore, we downsampled audio files to 16 kHz and applied BSS. The obtained audio (vocals) was used in all further experiments.

3.2.2. Synchronization of Labels

Usually, videos from the datasets are annotated per frame, while various videos can differ in terms of the frame rate. Thus, the annotations also had different sampling rates. To align them, we downsampled all labels to a sampling rate of 5 Hz. We think that this frequency is sufficient to detect changes in emotions, because the emotional category switches rarely.

3.2.3. 1D CNN + LSTM-Based Deep Network

Since there is no publicly available pretrained 1D CNN on raw audio, we constructed our own. To grasp the temporal information from 1D CNN embeddings more efficiently, we stacked two LSTM layers on top of it. The final layer had seven softmax neurons to match the number of classes. To represent window temporal modeling, the sequence-to-one modeling scheme was implemented. It maps one portion of the input acoustic raw data (for example, 4 s of audio) into emotional category probabilities. The number of model parameters was around 4.5 M. The architecture of the developed sequence-to-one model is presented in Figure 2. We would like to note that such a model is also an E2E model since it directly maps the raw audio to one of the emotion categories (or class probabilities).
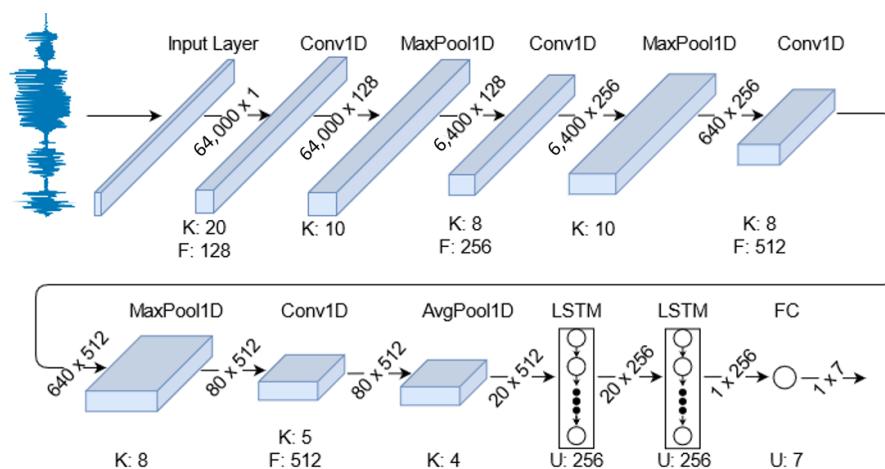

Figure 2. The architecture of the proposed 1D CNN + LSTM-based deep neural network.

3.3. Fusion Techniques

We employed "Model- and Class-based Weighed Fusion" (MCWF), where we had a fusion matrix of $L \times K$, where L and K denote the number of models and classes, respectively. That is, we had an importance weight for each class of each model, separately. The fusion weights can be optimized for any measure of interest from a pool of matrices that are generated randomly using a Dirichlet distribution for each class, such that the weights for each class over models sum up to unity. This approach has been successfully applied in former video-based affect recognition in-the-wild challenges [40].

4. Experimental Setting

In this section, we provide the experimental setting. This includes the data description, the features of dataset preprocessing, the set of hyperparameters used, and the way they were optimized during the training of the models. The code for reproducing the results is available at https://github.com/DresvyanskiyDenis/ABAW-SIU, accessed on 30 November 2021.

4.1. Datasets Used in the Study

AffectNet [77] is a large dataset of human emotional expressions, collected by web querying in six different languages and annotated in terms of seven emotional categories, namely Ekman's six basic emotions [81] plus neutral. Overall, the database contains around

450 K images of facial expressions, partially in an in-the-wild manner. The AffectNet class distribution is strongly shifted to the neutral and happiness classes (these two classes comprise 73.7% of all dataset instances), which introduces a class imbalance issue [82] that needs to be solved.

To relax the AffectNet class imbalance, we mixed it (excluding major neutral and happiness classes) with the FER2013 [78] dataset, increasing the examples of minority classes to around 15% on average. The FER2013 database contains approximately 34 K grayscale face images, annotated in terms of 7 emotional categories, and has a class imbalance problem as well.

The AffWild2 [11] dataset was used to evaluate the effectiveness of the proposed approaches. The database was collected from YouTube and consists of about three million annotated frames with different qualities. The annotation process for seven basic emotions was performed in a frame-by-frame manner, eliminating the widely known problem of evaluator lagging during time-continuous annotation. The class distribution of the AffWild2 dataset is presented in Table 1. Thus, it has an even higher class imbalance (the top two classes account for 79.7% of all instances and the top three for 90.6%) in comparison with AffectNet. Moreover, the challenging conditions of the data include different occlusions, pose variations, the absence of a person/face, and in some cases, the existence of multiple people in a frame, while each frame is annotated and should be predicted properly.

Table 1. The class distribution of the AffWild2 dataset.

Emotion	Training	Validation
Neutral	589,215 (63.39%)	183,636 (56.76%)
Anger	24,080 (2.59%)	8002 (2.47%)
Disgust	12,704 (1.37%)	5825 (1.80%)
Fear	11,155 (1.20%)	9754 (3.01%)
Happiness	152,010 (16.35%)	53,702 (16.60%)
Sadness	101,295 (10.90%)	39,486 (12.21%)
Surprised	39,035 (4.20%)	23,113 (7.14%)

The typical video sequence from the AffWild2 database is presented in Figure 3. For example, the first row of images represents the arising occlusions and unusual colors, while the second one the arising second human in the frame (in one frame, the face detection algorithm erroneously detected him as the main person) and a high amount of movements from the participant (wiggling, nodding, occluding the head with the arms). Moreover, the AffWild2 database contains several videos with two annotated persons occurring in different timings. Such difficulties make this database essentially in-the-wild and difficult to process, yet interesting and challenging to the research community.
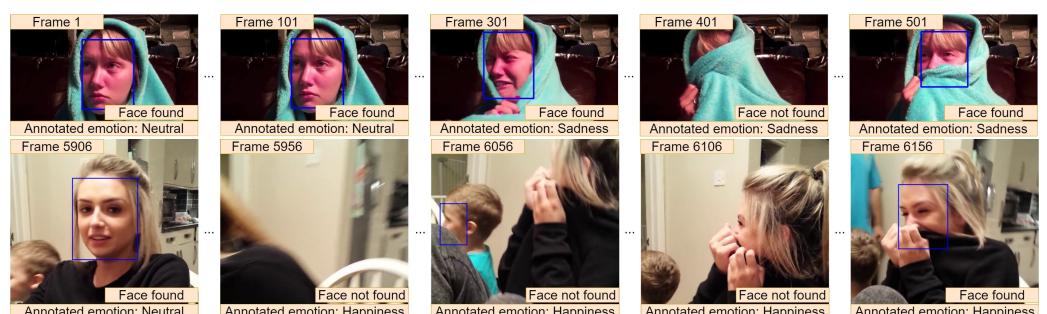

Figure 3. The example frames of two videos in the AffWild2 dataset, which demonstrate such challenges as pose variations and various occlusions.

4.2. Preprocessing of the AffWild2 Dataset

As a helpful supplement, the authors of the AffWild2 dataset provided the cropped faces for all frames in the video files. They localized them utilizing the HeadHunter

detector [83]. However, we found several problems while working with these localized faces: (1) the target face is confused with secondary faces (see Figures 3 and 4); (2) in case the face is covered by hands or other objects (obstacles), the face detector does not work properly, as exemplified in Figure 4.
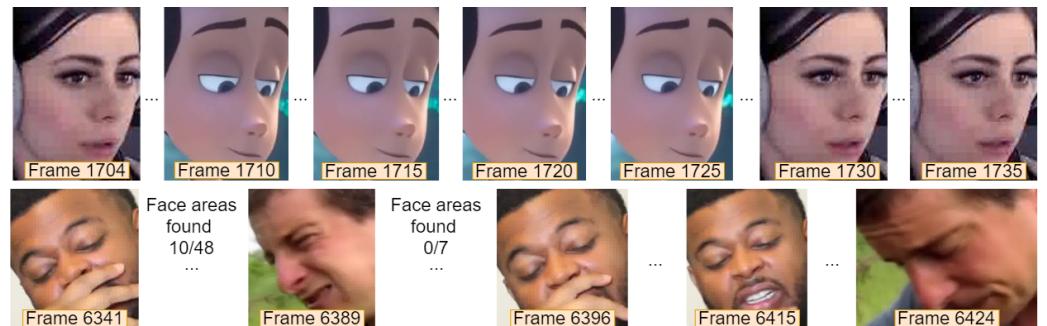

Figure 4. The examples of two video sequences, on which the HeadHunter face detector had many failures.

The noticed problems highly contributed to the effectiveness of the model training due to the loss of important information, consequently affecting the efficacy of the temporal aggregation and recognition. To eliminate the second problem, we exploited the RetinaFace detector [84], which works more accurately, including the cases when the face is covered by obstacles by more than 50% [85]. We tackled the first problem as well by utilizing the pure VGGFace2-model: we extracted the facial embeddings from the preceding and the current frames and compared the embeddings, using the cosine similarity. In case the cosine similarity is more than 0.5, the current face is considered as a correctly detected area. We can observe the higher reliability and accuracy of the obtained face detection approach in Figure 5. In comparison with the HeadHunter face detector, no faces were missed, and all of them were identified correctly.

Figure 5. The face detection results on the same two video sequences presented in Figure 4 using RetinaFace.

4.3. Models Setup

4.3.1. Audio Emotion Recognition Models

To obtain more data during the training, we set the shift of the window with a step equal to 2/5 the size of the window so that every audio chunk had an overlap of 3/5 with the respective former chunk. Thus, we approximately doubled the amount of training data. Since the audio after the noise cleaning (see Section 3.2.1) had a sampling rate of 16 kHz, the temporal context of 4 s in this case, for example, equals 48,000 samples in the waveform and 20 labels. As we applied the sequence-to-one modeling, we should have retained only one label per window. This was performed by selecting the mode of the whole label sequence, which is the most frequent emotion category.

The training process was conducted using the Adam optimizer with *learning rate =* 0.00025 and the *categorical cross-entropy* loss function. To regularize the model, a dropout with a rate of 0.3 after every convolutional layer was applied as well.

4.3.2. Visual Emotion Recognition Models

VGGFace2-EE. To construct an embedding extractor, we used the VGGFace2-model. To smooth the class imbalance, while training the VGGFace2-model on the AffWild2 database, we downsampled every class category in the training set: for the *neutral* class, we considered every tenth frame in every video, for the *anger, disgust, fear*, and *surprise* classes every second frame, and for the *happiness* and *sadness* classes every fifth frame. Subsequently, to make the model more robust to interference, we used data augmentation techniques such as rotation, horizontal flip, and brightness variation. Moreover, we applied logarithmic weighting to the loss function. According to the logarithmic weighting, the weight of class *i* is calculated as:

$$w_i = ln(\frac{rM}{n_i}), \quad \hat{w}_i = \left\{ \begin{array}{ll} 1, & if \ w_i < 1, \\ w_i & otherwise, \end{array} \right. \tag{1}$$

where *ln()* denotes the natural logarithm, *r* is a regularization parameter, *M* is the number of samples in the training set, and n_i is the number of samples in class *i*. We experimented with various values of the hyperparameter *r* in the [0, 1] range and optimized it as 0.47.

AffectNet-EE. For the AffectNet-EE training, we removed the VGGFace2-model's last layer and added above it Gaussian noise and one fully connected layer with 512 neurons and l2-regularization. Lastly, a softmax layer with seven neurons was stacked on top of the obtained model. The main difference of the AffectNet-EE model is that it was trained on the AffectNet dataset (and not on AffWild2). Moreover, to increase the proportion of the minority classes, the samples from the FER2013 dataset were added to the overall training set.

In the training process, such augmentation techniques as rotation, shear, horizontal flip, shifting, and changing the image contrast were used. In addition, to "dilute" the majority classes, we used the mixup [86] approach, which mixes two images and their labels, applying the weights generated by the Beta-distribution. Moreover, inversely proportional class weighting was applied to the categorical loss function.

During training, we exploited the cosine annealing with cold restart [87] as follows: the initial learning rate was set to 0.0001, which steadily descended to a minimum learning rate of 0.00001 within 6 epochs (=1 annealing cycle) and then sharply recovered to the initial learning rate. This procedure was repeated 5 times, resulting in the model training within 30 epochs.

AffWild2-EE. Structurally, Affwild2-EE is the same as AffectNet-EE; however, it was further fine-tuned on the AffWild2 corpus. To train the embedding extractor, we downsampled AffWild2 as follows: in every video, for category *neutral*, we took every tenth frame, for categories *anger, disgust, fear*, and *surprise* every second frame, and for *happiness* and *sadness* every fifth frame. All the other parameters, as well as the augmentation techniques, were chosen the same as for the AffectNet-EE training.

4.3.3. Temporal Modeling Techniques

As mentioned earlier, to handle the varying Frames Per Second (FPS) over the videos and correctly the implement subsystems using LSTM, we aligned all videos to have 5 FPS. Then, using AffectNet-EE, CNN embeddings were extracted and formed in a sequence of different lengths according to the window length and fed to the LSTM input.

As we describe below, we applied two different techniques of modeling temporal dependencies: calculating statistics from features within one window or using specially designed methods such as RNNs. Both methods have their pros and cons, and therefore, we employed both in our fusion framework. In addition, we should note that in most cases (except the fully end-to-end approach described in Section 5.1.2), we prepared the

features for temporal modeling in advance for computational efficiency. For every video frame, we extracted deep embeddings exploiting either the VGGFace2-EE, AffectNet-EE, or AffWild2-EE models. Thus, every frame was represented as 512 deep features obtained from the penultimate fully connected layer.

EE + SVM system. For statistical functional-based summarization of the frame-level features, we used two approaches:

- *SVM.* Means and STDs were calculated over each fixed-sized window, resulting in a vector with a length of 1024. Next, a meta-parameter search for the SVM using calculated suprasegmental features was carried out. We conducted extensive experiments with various kernels including *linear*, *polynomial*, *RBF* (γ optimized in [0.001, 0.1]), and regularization parameter C (in [1, 25]). The best result was obtained with following settings: kernel - *polynomial*, $\gamma = 0.1$, $C = 3$;

- *L-SVM.* We calculated the means, STDs, and leading coefficients for polynomials of the first and the second orders over each fixed-size window, resulting in a suprasegmental feature vector of dimensionality 2048. Next, we consistently applied cascaded normalization, in the form of z-, power-, and l2-normalization, respectively, and trained the Linear Support Vector Machine (L-SVM) on the obtained normalized vectors, as suggested in [88].

EE + LSTM system. As the second approach, we used "raw" deep embeddings as the input in the constructed LSTM neural network. It consisted of two layers with 512 and 256 neurons with l2-regularization and dropout between them (with a dropout rate of 0.5). On top of them, the fully connected softmax layer with seven neurons was placed.

In addition, for the training of the LSTM systems, we decimated the FPS rate in all AffWild2 dataset videos to 5. This had to be done, since all videos had different FPS (the lowest FPS was 7.5; the maximum FPS was 30), while the similarity in the rate of temporal context is a crucial point for LSTM networks. Then, using AffectNet-EE or AffWild2-EE (depending on the model), CNN embeddings were extracted and formed in a sequence of different lengths according to the window length and fed to the LSTM input.

Thus, both temporal modeling approaches performed sequence-to-one modeling, mapping frame-level deep embeddings within a fixed window into seven class probabilities. However, originally, one window contained several emotional labels, the number of which depended on the video frame rate. Therefore, after prediction, we expanded the vector of predicted probabilities with size 1×7 to $n \times 7$, where n is the original number of emotional labels in the particular fixed window. Since we had an intersection between windows, to obtain the final decision on each concrete video frame, the intersected class probabilities were averaged.

4.4. Performance Measure

During our research, we considered two measures for model evaluation: Unweighted Average Recall (UAR) and the official ABAW-FER CPM, which is defined as [54]:

$$CPM = 0.67 * F1 + 0.33 * Accuracy, \qquad (2)$$

where $F1$ is the weighted average of the Recall and the precision (also known as the *F*-measure) and the *Accuracy* is the fraction of predictions that are correctly classified.

We utilized UAR mostly during choosing the models (see the experiments described in Section 5.1) because of its ability to elicit and neglect overfitted models, while the CPM was used in order to be able to compare our models with other participants of the ABAW-FER Challenge.

5. Results

In this section, we describe the results obtained during the extensive experiments we conducted on the AffWild2 dataset. Moreover, the adoption of a variant of the proposed multimodal system within the ABAW-FER Challenge is presented.

5.1. Visual Emotion Recognition SubSystem

5.1.1. EE + SVM Subsystem

As we noted before, different techniques for temporal aggregation may be used. In this work, we focused on such methods as SVM and LSTM applied to the CNN embeddings. However, the first question to address is: What is the optimal context length? To find this out, we conducted experiments, fixing all the parameters except the temporal window length. For simplicity, we fixed the following parameters: (1) the face detector: HeadHunter; (2) AffWild2 as a training dataset (that is, VGGFace2-EE); (3) the temporal aggregation technique: EE + SVM system; (4) the logarithmic class weighting technique. We varied the length of the window from 2–8 with exponential steps, with an additional probe of 3 s. The experimental results are presented in Table 2. We chose two performance measures for the models' evaluation: Unweighted Average Recall (UAR) was chosen since it is known to be an unbiased measure of the reliability of the model (how much attention the model pays to every class), while the CPM was also chosen for comparison in terms of the ABAW-FER Challenge. As we can see from Table 2, a temporal context of 4 s was the best option for functional-based temporal aggregation (EE + SVM system), and therefore, all the subsequent experiments with this technique were performed using a temporal window of 4 s.

Table 2. The experimental results on the temporal window length selection on the AffWild2 validation set. VGGFace-EE was used as the embeddings extractor; SVM was used as a classifier on suprasegmental features. The best result within the column is highlighted in bold.

SysID	Window Length	UAR (%)	CPM (%)
1	2	41.9	54.5
2	3	42.8	55.3
3	4	**43.3**	**55.6**
4	8	37.5	47.8

Fixing the length of the temporal context, we continued optimizing the model via experiments on the variation of the face detector, embeddings extractors, and class weighting schemes. The results of the experiments are presented in Table 3. Here, we can highlight the models with the SysID 1 and 2: they had the highest UAR and CPM over all the other models. We discuss the results of the ABAW-FER Challenge in Section 5.3; however, we should note that one of these two models (namely SysID 2 in Table 3) was submitted during the ABAW competition.

Table 3. The performance results (%) for different combinations of face detectors, embeddings extractors, and class weighting schemes. SVM was used as a classifier on the suprasegmental features. The best results within the column are highlighted in bold.

SysID	Face Detector	Embedding Extractor	Class Weighting	UAR	CPM
1	HeadHunter	VGGFace2-EE	Logarithmic	**43.3**	**55.6**
2	HeadHunter	VGGFace2-EE	Balanced	**43.2**	**55.6**
3	RetinaFace	VGGFace2-EE	Logarithmic	39.0	52.6
4	HeadHunter	AffectNet-EE	Logarithmic	42.3	55.3
5	RetinaFace	AffectNet-EE	Logarithmic	39.1	51.7
6	RetinaFace	AffectNet-EE	Balanced	38.2	50.9
7	HeadHunter	Affwild2-EE	Logarithmic	42.1	54.7
8	RetinaFace	Affwild2-EE	Logarithmic	42.4	55.3

5.1.2. EE + LSTM Subsystem

Although the systems based on SVM with suprasegmental features reached a decent performance on the validation set, relatively low results were obtained on the test set. This may be caused by an insufficient ability to generalize data expressed via overfitting to the training set. Therefore, we decided to utilize a more sophisticated temporal aggregation method such as LSTM-NN, which was originally developed to process time series and has a memory mechanism to capture long temporal dependencies. Fixing the choice of the face detector and training data (pretraining the model on the AffectNet dataset, that is AffectNet-EE), we conducted an experiment, varying the weighting scheme and length of the temporal window, as was done in the previous subsection. We used only context window lengths of 2 s and 4 s due to the computational complexity faced during the investigation. The results of the experiments are presented in Table 4. First of all, we should note that we chose the best model according to the UAR score, since it reflects the model's generalization ability more adequately than the CPM (this was observed during overfitting to the development set in the previous subsection). Thus, the model with the balanced class weighting technique and a temporal window size of 2 s was chosen for the further experiments.

Apart from that, we decided to train the AffWild2-EE + LSTM model using the best parameters found during the experiments with the AffectNet-EE + LSTM model. Despite the relatively low performance (UAR = 43.85% and CPM = 52.84%) obtained on the AffWild2 validation data, we found that AffWild2-EE + LSTM achieved a Recall performance gain of 22.6% on the neutral class and 3.23% on the happiness class. This makes the AffWild2-EE + LSTM model a good contributor in the fusion systems (comprising several models); therefore, we used it in our further experiments.

Table 4. The experimental results on a selection of temporal window length and class weighting scheme on the AffWild2 validation set. AffectNet-EE was used as an embeddings extractor; LSTM was used as a temporal aggregation technique. The best result within the column is highlighted in bold.

SysID	Window Length	Class Weighting	UAR (%)	CPM (%)
1	2	Logarithmic	43.3	**55.1**
2	4	Logarithmic	43.2	52.8
3	2	Balanced	**52.8**	49.0
4	4	Balanced	50.8	49.0

Additionally, we carried out the experiment with the number of frozen layers while training the model. To do so, we fixed the meta-parameters of training as in the previous experiment and tried to conduct experiments with different options: training the whole CNN + LSTM model (AffectNet-E2E) simultaneously and training only the LSTM part of the model, while the CNN part was frozen (AffectNet-EE + LSTM). The results are presented in Table 5. We can clearly see that E2E training demonstrated higher performance in terms of the CPM, while using the main convolutional layers as embedding extractors, the AffectNet-EE + LSTM approach gave a more balanced Recall over classes expressed via a higher UAR. This means that E2E models sacrifice the Recall of minority classes in favor of higher accuracy. In addition, we can observe that our previous findings (see Table 4) were reinforced with the E2E learning: the temporal window of 2 s turned out to be the best one in terms of both performance measures. Thus, since we believe that UAR is more informative and adequate in comparison with the CPM, we chose the AffectNet-EE + LSTM model with a window size of 2 s for future fusion experiments.

Table 5. The classification performance comparison of the fully end-to-end (AffectNet-E2E) and separate model training (AffectNet-EE + LSTM) approaches on the AffWild2 validation set. The best result within the system is highlighted in bold.

Window Length	AffectNet-EE + LSTM		AffectNet-E2E	
	UAR (%)	CPM (%)	UAR (%)	CPM (%)
2	**52.8**	49.0	45.9	**53.6**
4	50.8	49.0	42.2	51.1

5.2. Audio Emotion Recognition Sub-System

For the audio model (1D CNN + LSTM), we also optimized the length of the temporal context (window size). We conducted experiments using the AffWild2 validation set with different lengths of windows (from 2–12 s with steps of 2 s). The best performance was observed with a temporal context of 4 s, which is partially in line with our findings from the video emotion recognition subsystem.

We also tried to use Pretrained Audio Neural Networks (PANNs) [89] that have demonstrated state-of-the-art performance in audio pattern recognition. These models extract features from raw waveforms, then process those data and return predictions in real time. In this study, we used the CNN-14 model, which consists of one layer for extracting features and six convolutional blocks, inspired by VGG-like CNNs [58]. The CNN-14 model was pretrained on the AudioSet dataset [90]. We fine-tuned the PANN model for the Expression Challenge. In this pipeline, we used 2D Mel-spectrograms as the features. However, our 1D CNN + LSTM model showed better results.

5.3. Multimodal Emotion Recognition System

As already mentioned, we took part in the ABAW competition. However, our results are not limited only by this challenge: we subsequently continued experimenting and improving our multimodal system after the challenge as well. Therefore, to be consistent and to have the possibility to compare our results with other works, we continued working on the AffWild2 dataset, using the same performance measure as before: the CPM. In this section, we present our results in chronological order to explain the research progress during and after the challenge.

During the ABAW-FER Challenge, first of all, we examined the efficacy of the unimodal subsystems individually. We chose as the simplest the following models: 1D CNN + LSTM, VGGFace2-EE, and VGGFace2-EE + SVM. Each of them represents one of the possible usages of a modality, namely audio, visual-static, and visual-temporal. Since multimodal systems usually perform better than unimodal models, we decided to submit a multimodal system as well. Exploiting the visual-temporal model turned out to be more accurate over using just the visual-static model (framewise prediction); therefore, we constructed the multimodal system from the 1D CNN + LSTM and VGGFace2-EE + SVM approaches. It should be noted that the fusion of these two unimodal systems was performed via a decision-level fusion. The best results obtained within the ABAW competition are presented in the first part of Table 6. Using the decision fusion of the audio- and video-based systems, we reached a CPM of 42.10% and ranked third place in the ABAW-FER Challenge (the results of the first five participants are shown in Table 7).

Upon the analysis of the state-of-the-art works in the multimodal emotion recognition domain, we extended our system with new subsystems that collectively can catch the emotional states more accurately. To adjust the VGGFace2-model more to the motion recognition task, we applied a multi-state fine-tuning of the VGGFace2-model, obtaining the AffectNet-EE model after the first stage and AffWild2-EE after the second stage, as described in Section 4.3.2. In addition, since the SVM hyperparameters need to be adjusted every time the training set is changed, we shifted towards an end-to-end approach such as combining the CNN and LSTM.

Table 6. The performances of the Audio (A) and Visual (V) systems on the ABAW-FER Challenge Validation (Val.) and the test sets. The first block (four lines) shows the results obtained in the scope of the competition, and the rest are the results of our extended systems obtained after the challenge. All results are in % **CPM**. Baselines: CPM-Val. = 36%, CPM-Test = 30%. The best result within the column is highlighted in bold.

SysID	Modality	System	Val.	Test
1	A	1D CNN + LSTM	35.09	-
2	V	VGGFace2-EE	50.23	40.60
3	V	VGGFace2-EE + SVM	55.66	42.00
4	A & V	Decision Fusion of SysID 1 & 3	**55.90**	**42.10**
5	V	AffWild2-EE + LSTM	54.73	44.21
6	V	SysID 5 & AffectNet-EE + LSTM	57.61	46.34
7	A & V	SysID 6 & 1D CNN + LSTM	54.69	47.58
8	A & V	SysID 7 & AffWild2-EE + L-SVM	**58.95**	**48.07**

The results of the proposed multimodal fusion system and its subsystems are presented in the second part of Table 6. We see that the fine-tuned AffWild2-EE + LSTM alone was able to outperform our best multimodal system in the challenge by 2.1% absolute. This partially confirms that LSTM can operate better with time series than just estimating statistics and making decisions based on them by SVM. Combining the two fine-tuned visual EE + LSTM subsystems (AffectNet-EE + LSTM and AffWild2-EE + LSTM) further improved the CPM by 2.1% absolute. The addition of the end-to-end audio model, namely the 1D CNN-LSTM model, to these two EE + LSTM visual subsystems in a weighted fusion framework further contributed, reaching a CPM of 47.6%. Finally, in the proposed system (System 8 in Table 6), we fused all our best performing subsystems and, therefore, further extended this fully end-to-end multimodal system with the use of a Linear SVM (L-SVM) on the functional features obtained from the AffWild2-EE. This final one advances the first runner-up performance on this challenge corpus (see Table 7).

Table 7. Top-5-performing systems at the ABAW-FER Challenge 2020 compared to the performance of our work.

Rank	Work	CPM (%)
1	Kuhnke et al. [70]	50.9
2	Gera and Balasubramanian [71]	43.4
3	Dresvyanskiy et al. [55]	42.1
4	Zhang et al. [91]	40.8
5	Deng et al. [73]	40.5
	This work	48.1

6. Analysis and Discussion

When analyzing the classwise F1-scores of our proposed multimodal system (see Figure 6), we observed that the visual modalities had both higher and more balanced F1-scores over classes compared to the audio model, which eventually contributed to the final fusion system (see Figure 7). From the weights figure, we observed that the most significant contribution for *neutral*, *anger*, and *surprised* classes was made by L-SVM, for *fear* and *sadness* by AffWild2-EE + LSTM, for happiness by AffectNet-EE + LSTM and L-SVM (almost equally). Interestingly, the highest contribution of the audio modality was observed on the *disgust* emotion, which weighed evenly with the contribution from AffectNet-EE + LSTM. Thanks to the MCWF, which combines the classwise strengths of each model, the fusion system's F1-scores were better than or on par with the best unimodal counterpart. Although the found optimal fusion weights and the corresponding F1-scores were highly correlated, the models with the highest F1-score on a class did not always obtain the highest

weight. Alternatively, the genetic-algorithm-based informed search can be investigated to improve the MCWF.

Figure 6. Classwise F1-scores for the proposed unimodal models and their fusion. Ne—Neutral, An—Anger, Di—Disgust, Fe—Fear, Ha—Happiness, Sad—Sadness, Sur—Surprised.

Figure 7. The values of the weights used in the proposed multimodal fusion system.

We further analyzed our best embedding extractor (CNN model), namely AffWild2-EE, using GradCAM [92] to check where it attended and why it failed. In Figure 8, we provide sample saliency maps overlayed on the images. For the images in the first row, the CNN attended to eye region and was observed not to be effected by the partial occlusion of the mouth area. The second raw embeddingshowed that the CNN performed well on pose and occlusion variations. The examples show that the subtlety of facial expressions is an important factor for misclassifications.
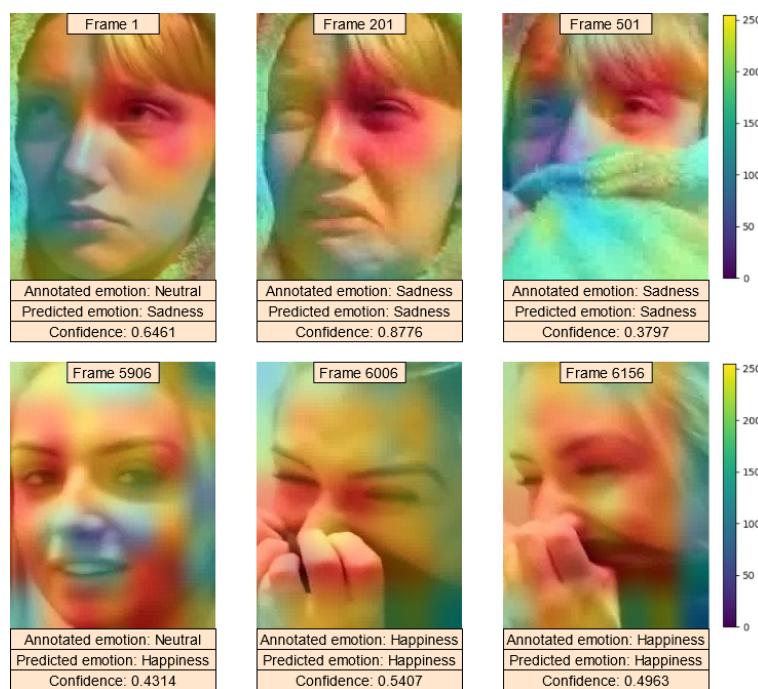
Figure 8. GradCAM-based saliency maps for predicted classes by AffWild2-EE.

We also investigated the performance of our system in the neighborhood of an emotion change point t with respect to the ground truth annotations. We provide a breakdown of the performance of the proposed system (Sys. 8) in the temporal neighborhood of the emotion change, namely in (A) $[t-4, t-2)$, (B) $[t-2, t)$, (C) $[t, t+2)$, and (D) $[t+2, t+4)$ s. The analysis window step of 2 s was based on our former window size optimization (2–4 s with different models). As shown in Table 8, we observed a drop in recognition performance around the emotion change point; however, even after the drop, the performance was around a 50% CPM and thus was at an acceptable level. The performance outside the ± 2 s relative to the emotion change points was almost the same as the overall development set CPM. We note that even though on average there were 11.6 emotion change points per video clip, considering the number of frames, this happened very rarely (the proportion of emotion change points to the total number of development set frames was 0.25%).

Table 8. The statistics of emotion change points and CPM performances around the emotion change points. # means number of; ECP means Emotional Change Point.

# Development Clips	# Frames	# ECPs	$[t-4, t-2)$	$[t-2, t)$	$[t, t+2)$	$[t+2, t+4)$
70	323,518	810	0.6191	0.4910	0.4835	0.5807

Sample sequences of frames taken from the temporal neighborhood of emotional change points are illustrated in Figure 9. Here, in each row, the frame in the center corresponds to the reference emotional change point t. From left to right, the relative time points (in seconds) are $t-4$, $t-2$, t, $t+2$, and $t+4$. Despite the challenges such as facial occlusions and pose variations, we observed a decent performance of capturing the emotions before and after the expression change points.
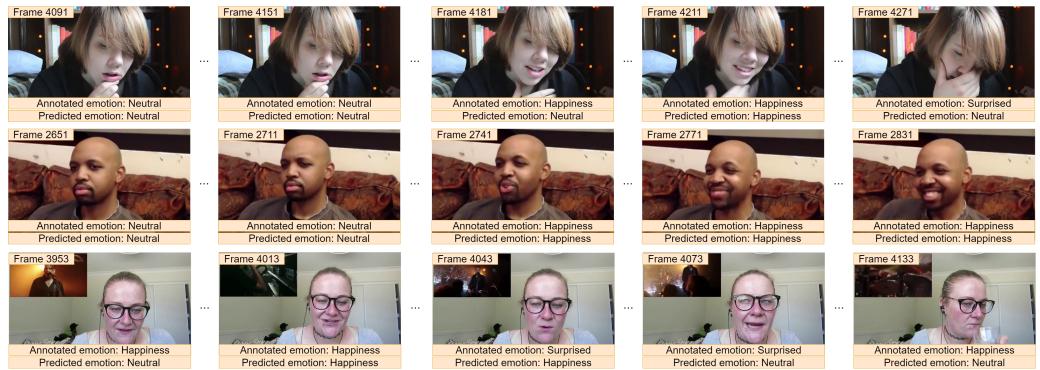
Figure 9. Example frames in the temporal vicinity of emotion change points in video clips with the corresponding ground truth and predicted emotion classes.

To sum up, our results are in line with the former works that reported higher unimodal predictive performance for and a contribution to the multimodal system by the visual modality over the acoustic modality [40,70,71,73,91]. Unlike former studies that reported only the visual modality contributing to the recognition of *disgust* [40], in our study, we observed the contribution of the acoustic system to this class. This contribution may stem from extralinguistic vocalizations, which deserves future investigation.

Inference Time Analysis

Additionally, we would like to note that the final system (SysID 8) presented in Table 6 (with a test set CPM of 48.07%) can be considered as a real-time method. To demonstrate this, we calculated the time needed for each processing step, namely from data preprocessing to the prediction generation step.

In Table 9, we provide the inference time based on the consecutive execution of all operations. However, some of them could be parallelized. For example, all the feature extraction models can be run in parallel on different CPUs. The same can be applied for the prediction generation process. Considering parallelization, the overall inference time would be 66.76 s for this video clip (0.56 SFI). This evidence allows stating that our method is a real-time inference method.

Table 9. The inference time for different sub-processes of the final fusion system. For the system's testing, the video file "134.avi" with a duration of 119 s was chosen. SFI means "Seconds For Inference", denoting the seconds needed to process one second of video.

Operation	Preprocessing	Feature Extraction	Prediction Generation
Sub-processes	Video frame decimation Face detection & Cropping Cosine similarity	AffectNet-EE (6.63 s, 0.06 SFI) AffWild2-EE (1.75 s, 0.01 SFI)	AffectNet-EE + LSTM (2.70 s, 0.02 SFI) AffWild2-EE + LSTM (2.40 s, 0.02 SFI) l-SVM (6.13 s) 1D CNN + LSTM (5.53 s, 0.05 SFI)
Inference time	54 s (0.45 SFI)	7.78 s (0.07 SFI)	16.76 s (0.14 SFI)
Total inference time: 78.54 s (0.66 SFI)			

7. Conclusions

This article investigated the efficacy of deep learning models in the in-the-wild audiovisual emotion recognition domain. We showed that the transfer learning performed via multi-stage fine-tuning of deep CNN model allows increasing the model performance significantly. We also observed that in both audio and video modalities, the deployed fully

end-to-end 1D CNN + LSTM and 2D CNN-EE + LSTM neural network architectures can be successfully exploited for catching the spatio-temporal patterns for the audiovisual emotion recognition task. Moreover, we demonstrated that the MCWF fusion of different deep neural networks with correctly fitted weights is able to enhance the predictive performance of the fused system over the unimodal systems.

One of the interesting findings in this research is that, despite its low individual performance, the audio-based model significantly contributed to the multimodal fusion process, especially for the emotion *disgust*. Analyzing the training dataset, we can note that the low performance of the audio model can be partly attributed to the long silence durations and background music or noise accompanying the participants' speech. Although the cleaning of the audio helped to significantly separate the human voices from other sounds, the ambiguity reflected by several different voices in one audio file was not illuminated. Nevertheless, even such a "confusing" audio channel contributed to the overall system performance. This underlines once more the necessity of using the multimodal emotion recognition system due to the possibility of combining the various advantages of unimodal subsystems.

In our future work, we plan to try more flexible information fusion techniques such as cross-modal attention-based fusion. Such an approach has shown promising results in other domains [93–95] and can significantly increase the performance of the considered system. To enhance the multimodal fusion via additional information, which can be extracted from already existing information channels, we plan to exploit the linguistic modality, which has shown promising results as an additional modality for multimodal emotion recognition systems [25].

References

1. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 2000.
2. Gupta, P.; Rajput, N. Two-stream emotion recognition for call center monitoring. In Proceedings of the Eighth Annual Conference of the International Speech Communication Association, Antwerp, Belgium, 27–31 August 2007; Citeseer: Princeton, NJ, USA, 2007.
3. Bojanić, M.; Delić, V.; Karpov, A. Call redistribution for a call center based on speech emotion recognition. *Appl. Sci.* **2020**, *10*, 4653. [CrossRef]
4. Zatarain-Cabada, R.; Barrón-Estrada, M.L.; Alor-Hernández, G.; Reyes-García, C.A. Emotion recognition in intelligent tutoring systems for android-based mobile devices. In Proceedings of the Mexican International Conference on Artificial Intelligence, Tuxtla Gutierrez, Mexico, 16–22 November 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 494–504.

5.   Yang, D.; Alsadoon, A.; Prasad, P.C.; Singh, A.K.; Elchouemi, A. An emotion recognition model based on facial recognition in virtual learning environment. *Procedia Comput. Sci.* **2018**, *125*, 2–10. [CrossRef]

6.   van der Haar, D. Student Emotion Recognition Using Computer Vision as an Assistive Technology for Education. In *Information Science and Applications*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 183–192.

7.   Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE Multimed.* **2012**, *19*, 34–41. [CrossRef]

8.   Dhall, A.; Goecke, R.; Ghosh, S.; Joshi, J.; Hoey, J.; Gedeon, T. From individual to group-level emotion recognition: Emotiw 5.0. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; pp. 524–528.

9.   Kollias, D.; Zafeiriou, S. Aff-Wild2: Extending the Aff-Wild Database for Affect Recognition. *arXiv* **2018**, arXiv:1811.07770.

10. Kollias, D.; Nicolaou, M.A.; Kotsia, I.; Zhao, G.; Zafeiriou, S. Recognition of affect in the wild using deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 1972–1979.

11. Kollias, D.; Zafeiriou, S. Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace. *arXiv* **2019**, arXiv:1910.04855.

12. Zafeiriou, S.; Kollias, D.; Nicolaou, M.A.; Papaioannou, A.; Zhao, G.; Kotsia, I. Aff-wild: Valence and arousal 'in-the-wild'challenge. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 1980–1987.

13. Kollias, D.; Zafeiriou, S. A multi-task learning & generation framework: Valence-arousal, action units & primary expressions. *arXiv* **2018**, arXiv:1811.07771.

14. Avots, E.; Sapiński, T.; Bachmann, M.; Kamińska, D. Audiovisual emotion recognition in wild. *Mach. Vis. Appl.* **2019**, *30*, 975–985. [CrossRef]

15. Eyben, F.; Weninger, F.; Gross, F.; Schuller, B. Recent developments in opensmile, the munich open-source multimedia feature extractor. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; pp. 835–838.

16. Eyben, F. *Real-Time Speech and Music Classification by Large Audio Feature Space Extraction*; Springer: Berlin/Heidelberg, Germany, 2015.

17. Schuller, B.; Steidl, S.; Batliner, A. The INTERSPEECH 2009 emotion challenge. In Proceedings of the 10th Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September 2009; pp. 312–315. [CrossRef]

18. Schuller, B.; Steidl, S.; Batliner, A.; Hantke, S.; Hönig, F.; Orozco-Arroyave, J.R.; Nöth, E.; Zhang, Y.; Weninger, F. The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition. In Proceedings of the 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015; pp. 478–482. [CrossRef]

19. Schuller, B.W.; Batliner, A.; Bergler, C.; Mascolo, C.; Han, J.; Lefter, I.; Kaya, H.; Amiriparian, S.; Baird, A.; Stappen, L.; et al. The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Brno, Czech Republic, 30 August–3 September 2021; pp. 431–435. [CrossRef]

20. Pancoast, S.; Akbacak, M. *Bag-of-Audio-Words Approach for Multimedia Event Classification*; Technical Report; SRI International Menlo Park United States: Menlo Park, CA, USA, 2012.

21. Schmitt, M.; Ringeval, F.; Schuller, B. At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech. In Proceedings of the 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; pp. 495–499. [CrossRef]

22. Kaya, H.; Karpov, A.A.; Salah, A.A. Fisher vectors with cascaded normalization for paralinguistic analysis. In Proceedings of the 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015; pp. 909–913. [CrossRef]

23. Kaya, H.; Karpov, A.A. Fusing Acoustic Feature Representations for Computational Paralinguistics Tasks. In Proceedings of the 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; pp. 2046–2050. [CrossRef]

24. Gosztolya, G. Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds. In Proceedings of the 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15–19 September 2019; pp. 2413–2417. [CrossRef]

25. Soğancıoğlu, G.; Verkholyak, O.; Kaya, H.; Fedotov, D.; Cadée, T.; Salah, A.A.; Karpov, A. Is Everything Fine, Grandma? Acoustic and Linguistic Modeling for Robust Elderly Speech Emotion Recognition. *Proc. Interspeech* **2020**, 2097–2101. [CrossRef]

26. Perronnin, F.; Dance, C. Fisher kernels on visual vocabularies for image categorization. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 18–23 June 2007; IEEE: Piscataway, NJ, USA, 2007; pp. 1–8.

27. Cummins, N.; Amiriparian, S.; Hagerer, G.; Batliner, A.; Steidl, S.; Schuller, B.W. An image-based deep spectrum feature representation for the recognition of emotional speech. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 478–484.

28. Keesing, A.; Koh, Y.S.; Witbrock, M. Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Brno, Czech Republic, 30 August–3 September 2021; pp. 3415–3419. [CrossRef]

29. Szep, J.; Hariri, S. Paralinguistic Classification of Mask Wearing by Image Classifiers and Fusion. In Proceedings of the 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25–29 October 2020; pp. 2087–2091. [CrossRef]

30. Lian, Z.; Tao, J.; Liu, B.; Huang, J.; Yang, Z.; Li, R. Context-Dependent Domain Adversarial Neural Network for Multimodal Emotion Recognition. In Proceedings of the 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25–29 October 2020; pp. 394–398. [CrossRef]

31. Markitantov, M.; Dresvyanskiy, D.; Mamontov, D.; Kaya, H.; Minker, W.; Karpov, A. Ensembling End-to-End Deep Models for Computational Paralinguistics Tasks: ComParE 2020 Mask and Breathing Sub-Challenges. In Proceedings of the 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25–29 October 2020; pp. 2072–2076. [CrossRef]

32. Dvoynikova, A.; Markitantov, M.; Ryumina, E.; Ryumin, D.; Karpov, A. Analytical Review of Audiovisual Systems for Determining Personal Protective Equipment on a Person's Face. *Inform. Autom.* **2021**, *20*, 1116–1152. [CrossRef]

33. Ahonen, T.; Hadid, A.; Pietikäinen, M. Face recognition with local binary patterns. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 469–481.

34. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 21–23 September 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 886–893.

35. Slimani, K.; Kas, M.; El Merabet, Y.; Messoussi, R.; Ruichek, Y. Facial Emotion Recognition: A Comparative Analysis Using 22 LBP Variants. In Proceedings of the 2nd Mediterranean Conference on Pattern Recognition and Artificial Intelligence, Rabat, Morocco, 27–28 March 2018; MedPRAI '18; Association for Computing Machinery: New York, NY, USA, 2018; pp. 88–94. [CrossRef]

36. Julina, J.K.J.; Sharmila, T.S. Facial Emotion Recognition in Videos using HOG and LBP. In Proceedings of the 2019 4th International on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), Bangalore, India, 17–18 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 56–60.

37. Lakshmi, D.; Ponnusamy, R. Facial emotion recognition using modified HOG and LBP features with deep stacked autoencoders. *Microprocess. Microsyst.* **2021**, *82*, 103834. [CrossRef]

38. Almaev, T.R.; Valstar, M.F. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2–5 September 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 356–361.

39. Gürpınar, F.; Kaya, H.; Salah, A.A. Combining Deep Facial and Ambient Features for First Impression Estimation. In *ECCV Workshop Proceedings*; Springer: Cham, Switzerland, 2016; pp. 372–385.

40. Kaya, H.; Gürpınar, F.; Salah, A.A. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image Vis. Comput.* **2017**, *65*, 66–75. [CrossRef]

41. Hu, C.; Jiang, D.; Zou, H.; Zuo, X.; Shu, Y. Multi-task micro-expression recognition combining deep and handcrafted features. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 946–951.

42. Escalante, H.J.; Kaya, H.; Salah, A.A.; Escalera, S.; Güçlütürk, Y.; Güçlü, U.; Baró, X.; Guyon, I.; Jacques, J.C.S.; Madadi, M.; et al. Modeling, Recognizing, and Explaining Apparent Personality from Videos. *IEEE Trans. Affect. Comput.* **2020**, 1. [CrossRef]

43. Fan, Y.; Lu, X.; Li, D.; Liu, Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 445–450.

44. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323.

45. Kaya, H.; Fedotov, D.; Dresvyanskiy, D.; Doyran, M.; Mamontov, D.; Markitantov, M.; Akdag Salah, A.A.; Kavcar, E.; Karpov, A.; Salah, A.A. Predicting Depression and Emotions in the Cross-Roads of Cultures, Para-Linguistics, and Non-Linguistics. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, AVEC '19, Nice, France, 21–25 October 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 27–35. [CrossRef]

46. Yu, D.; Sun, S. A systematic exploration of deep neural networks for EDA-based emotion recognition. *Information* **2020**, *11*, 212. [CrossRef]

47. Mou, W.; Shen, P.H.; Chu, C.Y.; Chiu, Y.C.; Yang, T.H.; Su, M.H. Speech Emotion Recognition Based on CNN+ LSTM Model. In Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021), Taoyuan, Taiwan, 15–16 October 2021; pp. 43–47.

48. Rizos, G.; Baird, A.; Elliott, M.; Schuller, B. Stargan for Emotional Speech Conversion: Validated by Data Augmentation of End-To-End Emotion Recognition. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 3502–3506.

49. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Epps, J.; Schuller, B.W. Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]

50. Pandit, V.; Schmitt, M.; Cummins, N.; Schuller, B. I see it in your eyes: Training the shallowest-possible CNN to recognise emotions and pain from muted web-assisted in-the-wild video-chats in real-time. *Inf. Process. Manag.* **2020**, *57*, 102347. [CrossRef]

51. Kapidis, G.; Poppe, R.; Veltkamp, R.C. Multi-Dataset, Multitask Learning of Egocentric Vision Tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, 1. [CrossRef]

52. Kollias, D.; Tzirakis, P.; Nicolaou, M.A.; Papaioannou, A.; Zhao, G.; Schuller, B.; Kotsia, I.; Zafeiriou, S. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *Int. J. Comput. Vis.* **2019**, *127*, 907–929. [CrossRef]

53. Verkholyak, O.; Fedotov, D.; Kaya, H.; Zhang, Y.; Karpov, A. Hierarchical Two-level Modelling of Emotional States in Spoken Dialog Systems. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6700–6704. [CrossRef]

54. Kollias, D.; Schulc, A.; Hajiyev, E.; Zafeiriou, S. Analysing Affective Behavior in the First ABAW 2020 Competition. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG), Buenos Aires, Argentina, 16–20 November 2020; IEEE Computer Society: Washington, DC, USA, 2020; pp. 794–800.

55. Dresvyanskiy, D.; Ryumina, E.; Kaya, H.; Markitantov, M.; Karpov, A.; Minker, W. An Audio-Video Deep and Transfer Learning Framework for Multimodal Emotion Recognition in the wild. *arXiv* **2020**, arXiv:2010.03692.

56. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [CrossRef]

57. Kwon, O.W.; Chan, K.; Hao, J.; Lee, T.W. Emotion recognition by speech signals. In Proceedings of the Eighth European Conference on Speech Communication and Technology, Geneva, Switzerland, 1–4 September 2003.

58. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

59. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

60. Al Osman, H.; Falk, T.H. Multimodal affect recognition: Current approaches and challenges. In *Emotion and Attention Recognition Based on Biological Signals and Images*; IntechOpen: London, UK, 2017; pp. 59–86.

61. Toisoul, A.; Kossaifi, J.; Bulat, A.; Tzimiropoulos, G.; Pantic, M. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nat. Mach. Intell.* **2021**, *3*, 42–50. [CrossRef]

62. Xie, B.; Sidulova, M.; Park, C.H. Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Cross-modality Fusion. *Sensors* **2021**, *21*, 4913. [CrossRef] [PubMed]

63. Ranganathan, H.; Chakraborty, S.; Panchanathan, S. Multimodal emotion recognition using deep learning architectures. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–9 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–9.

64. Liu, D.; Wang, Z.; Wang, L.; Chen, L. Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning. *Front. Neurorobotics* **2021**, *15*, 697634. [CrossRef] [PubMed]

65. Tzirakis, P.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.W.; Zafeiriou, S. End-to-end multimodal emotion recognition using deep neural networks. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1301–1309. [CrossRef]

66. Tripathi, S.; Tripathi, S.; Beigi, H. Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv* **2018**, arXiv:1804.05788.

67. Poria, S.; Majumder, N.; Hazarika, D.; Cambria, E.; Gelbukh, A.; Hussain, A. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intell. Syst.* **2018**, *33*, 17–25. [CrossRef]

68. Majumder, N.; Hazarika, D.; Gelbukh, A.; Cambria, E.; Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl.-Based Syst.* **2018**, *161*, 124–133. [CrossRef]

69. Tzirakis, P.; Chen, J.; Zafeiriou, S.; Schuller, B. End-to-end multimodal affect recognition in real-world environments. *Inf. Fusion* **2021**, *68*, 46–53. [CrossRef]

70. Kuhnke, F.; Rumberg, L.; Ostermann, J. Two-Stream Aural-Visual Affect Analysis in the Wild. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 366–371.

71. Gera, D.; Balasubramanian, S. Affect Expression Behaviour Analysis in the Wild using Spatio-Channel Attention and Complementary Context Information. *arXiv* **2020**, arXiv:2009.14440.

72. Liu, H.; Zeng, J.; Shan, S.; Chen, X. Emotion Recognition for In-the-wild Videos. *arXiv* **2020**, arXiv:2002.05447.

73. Deng, D.; Chen, Z.; Shi, B.E. Multitask Emotion Recognition with Incomplete Labels. *arXiv* **2020**, arXiv:2002.03557.

74. Do, N.T.; Nguyen-Quynh, T.T.; Kim, S.H. Affective Expression Analysis in-the-wild using Multi-Task Temporal Statistical Deep Learning Model. *arXiv* **2020**, arXiv:2002.09120.

75. Youoku, S.; Toyoda, Y.; Yamamoto, T.; Saito, J.; Kawamura, R.; Mi, X.; Murase, K. A Multi-term and Multi-task Analyzing Framework for Affective Analysis in-the-wild. *arXiv* **2020**, arXiv:2009.13885.

76. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A. VGGFace2: A dataset for recognising faces across pose and age. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 67–74.

77. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [CrossRef]

78. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 117–124.

79. Laugs, C.; Koops, H.V.; Odijk, D.; Kaya, H.; Volk, A. The Influence of Blind Source Separation on Mixed Audio Speech and Music Emotion Recognition. In Proceedings of the Companion Publication of the 2020 International Conference on Multimodal Interaction, Utrecht, The Netherlands, 25–29 October 2020; ICMI '20 Companion; Association for Computing Machinery: New York, NY, USA, 2020; pp. 67–71. [CrossRef]

80. Hennequin, R.; Khlif, A.; Voituret, F.; Moussallam, M. Spleeter: A fast and efficient music source separation tool with pretrained models. *J. Open Source Softw.* **2020**, *5*, 2154. [CrossRef]

81. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124. [CrossRef]

82. Ryumina, E.; Karpov, A. Comparative analysis of methods for imbalance elimination of emotion classes in video data of facial expressions. *J. Sci. Tech. J. Inf. Technol. Mech. Opt.* **2020**, *20*, 683–691. [CrossRef]

83. Mathias, M.; Benenson, R.; Pedersoli, M.; Van Gool, L. Face detection without bells and whistles. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 720–735.

84. Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; Zafeiriou, S. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5203–5212.

85. Ryumina, E.; Ryumin, D.; Ivanko, D.; Karpov, A. A Novel Method for Protective Face Mask Detection Using Convolutional Neural Networks and Image Histograms. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2021**, *XLIV-2/W1-2021*, 177–182. [CrossRef]

86. Zhang, H.; Cissé, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond Empirical Risk Minimization. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.

87. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings, Toulon, France, 24–26 April 2017.

88. Kaya, H.; Karpov, A.A.; Salah, A.A. Robust acoustic emotion recognition based on cascaded normalization and extreme learning machines. In *International Symposium on Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 115–123.

89. Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; Plumbley, M.D. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2880–2894. [CrossRef]

90. Gemmeke, J.F.; Ellis, D.P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 776–780.

91. Zhang, Y.; Huang, R.; Zeng, J.; Shan, S. M3F: Multi-Modal Continuous Valence-Arousal Estimation in the Wild. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 617–621.

92. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

93. Huan, R.H.; Shu, J.; Bao, S.L.; Liang, R.H.; Chen, P.; Chi, K.K. Video multimodal emotion recognition based on Bi-GRU and attention fusion. *Multimed. Tools Appl.* **2021**, *80*, 8213–8240. [CrossRef]

94. Hori, C.; Hori, T.; Lee, T.Y.; Zhang, Z.; Harsham, B.; Hershey, J.R.; Marks, T.K.; Sumi, K. Attention-based multimodal fusion for video description. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4193–4202.

95. Ye, L.; Rochan, M.; Liu, Z.; Wang, Y. Cross-modal self-attention network for referring image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–16 June 2019; pp. 10502–10511.