Audio-Visual Gated-Sequenced Neural Networks for Affect Recognition

Decky Aspandi, Federico Sukno, Björn W. Schuller, *Fellow, IEEE*, and Xavier Binefa

Abstract—The interest in automatic emotion recognition and the larger field of Affective Computing has recently gained momentum. The current emergence of large, video-based affect datasets offering rich multi-modal inputs facilitates the development of deep learning-based models for automatic affect analysis that currently holds the state of the art. However, recent approaches to process these modalities cannot fully exploit them due to the use of oversimplified fusion schemes. Furthermore, the efficient use of temporal information inherent to these huge data are also largely unexplored hindering their potential progress. In this work, we propose a multi-modal, sequence-based neural network with gating mechanisms for Valence and Arousal based affect recognition. Our model consists of three major networks: Firstly, a latent-feature generator that extracts compact representations from both modalities that have been artificially degraded to add robustness. Secondly, a multi-task discriminator that estimates both input identity and a first step emotion quadrant estimation. Thirdly, a sequence-based predictor with attention and gating mechanisms that effectively merges both modalities and uses this information through sequence modelling. In our experiments on the SEMAINE and SEWA affect datasets, we observe the impact of both proposed methods with progressive increase in accuracy. We further show in our ablation studies how the internal attention weight and gating coefficient impact our models' estimates quality. Finally, we demonstrate state of the art accuracy through comparisons with current alternatives on both datasets.

Index Terms—Affective computing, deep learning, multi-modal fusion, sequence modelling

1 Introduction

Emotions of humans made accessible and 'readable' to computing devices keeps trending, as we near a robust automatic recognition which opens up real world usage such as in education [1], healthcare [2], [3], [4], human computer interaction [5] among others. Given its wide range of application potentials, the generalised affect recognition task is rapidly growing as reflected by recent availability of affect-related datasets, such as SEMAINE [6] as was used in the first Audio/Visual Emotion Challenge (AVEC). This is further extended by the recently introduced SEWA [7] database enabling the rapid development of general, automatic visual-based emotion recognition up to in the wild settings. While the field started from the use of handcrafted methods, it currently heavily relies on deep learning-based approaches due to the higher potential accuracy achieved [8], [9]. The use of other modalities such as sound and text has also improved the current systems in other emotional aspects that the visual modality lacks or situations, where it is not accessible or disturbed. This in turn also encourages the combination of these modalities, typically by direct concatenation approaches [10], [11], [12]. However, such a straightforward approach may produce sub-optimal results given the difference characteristics of each modality [13].

Another aspect to consider is the need to deal with *bigger-data*, given the emergence of video-based datasets that enrich the widely used modality features with the inclusion of temporal information. To this end, several authors have explored the use of deep-learning based sequence modelling of long-short term memory (LSTM) recurrent neural networks (RNNs) [14], [15], endowed also with attention mechanisms [16], [17], [18] to exploit these sequence based data inputs. However, such spatio-temporal modelling often results in very high-dimensional feature spaces and large volumes of data, making training difficult and time consuming.

This work addresses the current lack of efficient temporal modelling and effective multi-fusion approaches to general affect analysis, by proposing the use of latent sequence networks combined with gating mechanisms to effectively fuse multi-modal inputs. We do so by incorporating three major networks, coined Generator (G), Discriminator (D), and Predictor (P), which are trained in an adversarial setting to estimate the affect domains of Valence (V) and Arousal (A). Furthermore, we capitalise on these latent features to enable temporal modelling using internal LSTMs that are trained progressively using curriculum learning enhanced with adaptive attention. Finally, we combine the input modalities through gating

- *Decky Aspandi is with the Department of Information and Communication Technology, Universitat Pompeu Fabra, Barcelona 08026, Spain, and also with the Institute for Parallel and Distributed Systems, University of Stuttgart, Stuttgart 70569, Germany. E-mail: decky.aspandilatif@gmail.com.*
- *Federico Sukno and Xavier Binefa are with the Department of Information and Communication Technology, Universitat Pompeu Fabra, Barcelona 08026, Spain. E-mail: {federico.sukno, xavier.binefa}@upf.edu.*
- *Björn W. Schuller is with the Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Augsburg 86159, Germany. E-mail: schuller@informatik.uni-augsburg.de.*

This work was supported in part by the Spanish Ministry of Science and Innovation under Grant PID2020-114083GB-I00, in part by donation bahi2018-19 to the CMTech at UPF, in part by European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 826506 (sustAGE), and in part by German BMBF-KMU Innovativ Program through UDeco project.
(Corresponding author: Decky Aspandi.)
Digital Object Identifier no. 10.1109/TAFFC.2022.3156026

mechanisms for more effective modality fusion, leading to our state of the art accuracy. Specifically, the contributions of this paper are as follows:

1) We upgrade the standard adversarial setting with a third network that fuses features from the Generator and Discriminator. This produces features that combine the latent space from the autoencoder-based Generator and a V/A Quadrant estimate produced by the modified Discriminator, resulting in a compact, but meaningful representation that helps reduce the training complexity.
2) We propose the use of sequential modelling with attention to enhance our model estimates, and also quantify the relative impact of these adaptive attention mechanism by calculating the respective internal weight activation differences.
3) We extend our temporal modelling with gating networks for more effective fusion of both, audio and visual modalities. We further evaluate its effectiveness in our ablation study using thresholding analysis.
4) We report state of the art accuracy of our models on both the SEMAINE and SEWA datasets and compare our results to other alternatives.

Preliminary results of our modified adversarial training and sequential attention can be found in [19] and [20] respectively. Specifically, in [19], we initially introduced latent based neural network for multi-modal inputs to perform V/A prediction using adversarial learning. Then in [20], we applied sequential modelling with attention to the visual input modality. The findings in these two previous papers lead to our current work that focuses in improving the models accuracy using effective fusion of both audio and visual modality inputs through combined gating mechanisms and sequential attention. Subsequently, we provide the quantitative analysis of the improvements made using each of our proposed mechanisms (including sequential attention).

The rest of the paper is organised as follows: Section 2 describes the related work in the context of facial-based emotion recognition and the use of different modalities and other relevant temporal modelling; in Section 3, we explain our Audio-Visual Gated-Sequenced Neural Networks consisting of two major networks combined using our gated-sequence modelling. In Section 4, we report our results on both, the SEMAINE and SEWA datasets in relation to each of our methods, and further compare our results with current state of the arts models. Finally, Section 5 provides the conclusion.

2 Related Work

Multi-modal emotion recognition started by the use of classical machine learning techniques, applied to visual and audio features to enable automatic affect estimation. Examples of early approaches include partial least squares regression [21], support vector machines [22] and low-level audio descriptor [23]. These techniques had been applied largely to their specific applications, such as Health [24], Physchology [25] and Education [26] with limited success. This is mainly because of the accuracy limit of such approaches and they are highly conditioned to work in their specialized task, thus preventing their potential use onto broader application spectrums [27]. Thus, to further progress the investigations in this field, the development of larger, generalized affect datasets was initiated, with the SEMAINE [6] dataset as one popular instance. This audio-visual dataset facilitates direct, and general affect analysis for human and agent interactions, and has been used by many authors. One of the early works is Gunes *et al.* [28], who used global head motions consisting of nod and shake to be fed to individual Hidden Markov Models (HMM) to construct the baseline features, which are then utilised by Support Vector Regression (SVR) to estimate the final affect dimension. Using a person independent scheme in their experiment, they proved that automatic affect recognition is indeed possible to a certain degree. Progressing ahead, Kossaifi *et al.* [8] introduced a hybrid system that used deep learning alongside classical geometrical and texture features for affect recognition. Specifically, they proposed to include the use of features extracted from Scale-Invariant Feature Transform (SIFT), Local Binary Pattern (LBP), and facial Landmarks combined with several classifiers, such as Bag of Words (BOW) and conditional Random Field (RF). In addition, they performed transfer learning to several Convolutional Network-based models to investigate the effectiveness of these deep learning models. Using the SEMAINE database, alongside the other related affect dataset of AFEW-VA [8], they found that deep learning-based methods constantly outperformed other classical approaches and provides new baselines for each dataset.

More recently, the SEWA dataset [7] was published to allow more extensive deep learning-based modelling under unconstrained settings and conditions (in-the-wild) and offer multiple languages and cultures at the same time. Such deep learning-based approaches can be seen in the recent works of [29] and [9]. Mitenkova *et al.* [29] introduced tensor modelling for affect estimations applied to visual inputs. Specifically, they utilised tucker tensor regression optimised by deep gradient techniques, which permits the preservation of the structure of the data and reduction of the number of parameters. Similarly, Kosaifi *et al.* [9] introduced the use of tensor decomposition to enable their multi-dimensional convolutional approach for visual-based emotion recognition. In their work, they applied a generalised factorised higher-order framework to several convolutional models, such as ResNet, Inception, and Mobile net. Furthermore, they proposed to perform a more efficient tensor decomposition on Convolutional Operations by the introduction of weight vector coefficients with non-linearities affecting the magnitude of the decomposed factors. Then, they also added higher-order transduction and automatic rank selections in their pipelines to further optimise the necessary calculation operations. Using this approach, they arrived at state of the art results.

Visual-based approaches have been considerably gaining attention lately, since facial expressions are considered one of the dominant channels to display affective information [30], [31]. Some early examples of facial based emotion recognition is the work of [32] and [33] which used hand-crafted based features for general human computer interactions and healthcare applications respectively. Furthermore, recent works of [34] shows that utilising deep learning based models allows for more accurate estimates to deal with in the wild affect recognition (for recent compilation of facial based emotion recognition, the reader could see in this recent review [35]). However, facial expressions does

not always provide the full emotional information [7]. Indeed, it has been shown that modalities such as Electrocardiogram (ECG) and audio can complement and enhance the performance obtained from visual-features [36]. Specifically, the audio modality has been highlighted for its accurate Arousal estimates [7], [37]. Similar to visual based affect recognition, the audio based emotion recognition started by the use of handcrafted features [23], [38], and proceeded by the use of Deep Learning based models applied to bigger, and generalized affect recognition. One example of the latter is the work of Yang *et al.* [39] that exploited both the soundwave and its spectrogram derivatives as main features for affect recognition applied to both SEMAINE and Recola [40] dataset. In their work, they used a 3D Convolutional Neural Network (3DCNN) to extract the individual waveform and spectral features. Then, these features were combined using basic concatenation approaches and passed to a Bidirectional Long Short-term Memory (BLSTM) network to estimate final Valence and Arousal values. The main drawback of this mechanisms however is the model complexity that involves multiple level of temporal information. This limitation further motivated our previous work [19] where we introduce latent features modelling for more efficient use of visual features that are later combined with audio features. The visual latent features are formed using a Generator network that is trained with a Discriminator through the adversarial training. These visual features are then used to condition the Discriminator that receives both raw images and reduced sound features (eGeMAPS [41]) resulting to higher and more balanced accuracy in both Valence and Arousal predictions. Another recent approach is the Dialogue-RNN [37] that tries to incorporate the notion of dialogue for affect predictions, thus expanding the potential of multi-modal affect recognition approach. In their work, the authors also used the text modality alongside visual and sound information as main feature. They chose LSTM and Gated Recurrent Units (GRU) to explicitly model the interaction between the user (global, speaker, listener) through sequential learning, thus benefiting from this additional knowledge. They found that adding attention modules improved the accuracy of their models, suggesting the importance of modelling the interaction between modalities. Finally, current advances on multimodal affect recognition could be found on AVEC challenges which evaluate the models across different applications (including health and in the wild settings) [42], [43], [44]. It has been observed that using fusion of several modalities often leads to more robust and accurate predictions, with the concatenation operations are commonly used to fuse each modality [7], [44]. However, even though these concatenation strategies could work to certain extent, they can potentially neglect important relationships between different modalities as shown in other machine vision fields [45].

The recent availability of video-based datasets has stimulated the use of temporal modelling that has been shown to enhance models' training [46], [47]. Some related examples in Affective Computing include the works by Tellamekala *et al.* [14] and Ma *et al.* [15]. Specifically, Tellamekala *et al.* [14] enforced temporal coherency combined with smoothness priors during feature representation by constraining the differences between adjacent frames. On the other hand, Ma *et al.* [15] utilised LSTMs with residual connections to process sequences of multi-modal data inputs. Furthermore, the use of attention methods has also been recently explored in the works of Xiaohua *et al.* [18] and Li *et al.* [17]. Xiahoua *et al.* incorporated multi-stage attention consisting of both spatial and temporal attention in their facial-based affect estimation pipeline. Meanwhile, Li *et al.* used deep networks that capitalise an attention mechanism [16] on top of their LSTM networks to process a spectrogram representation of audio input, allowing them to perform the respective affective states prediction. In our previous work [20], we further increased the efficiency of sequential attention to perform affect estimations using visual inputs. We did this by using an auxiliary network (Combiner) that is trained in tandem with a set of Generator and Discriminator. As such, reducing their training complexity and allowing us to perform sequential attention more effectively using learned latent features from the Generator. We have shown that the use of sequence modelling leads to more accurate and stable affect estimates. Furthermore, we observed that the length of sequences involved also impact the models behaviour and accuracy produced.

In summary, recent developments of large scale, generalized affect datasets such as SEMAINE [6] and SEWA [7] have facilitated the development of automatic affect recognition with a broad application and high accuracy potentials. The starting point was the use of handcrafted features and classifiers applied to visual features, typically the facial area. The field then progressed toward the use of Deep Learning approaches, given that it allows for more accurate affect estimations. Furthermore, the use of other modalities has emerged due to the limitation of the visual features in regards to the accuracy obtained, and some works also tried to combine modalities by the use of simple concatenation. However, these straight-forward approaches have the limitation of their tendency to give equal weights to the different modalities [45], [48]. This can be problematic, in the situation where the importance of one modality may be considered higher than that of the others [7]. This problem could be mitigated by the use of Gating Mechanisms [13] that permit the adaptive weighting for the considered modalities. Furthermore, we also propose to combine it with our temporal modelling including attention, to allow for more accurate results as recently shown in other related studies [46], [47], including our preliminary results in [19], [20]. These approach enable our proposed model to make more efficient use of both Audio and Visual modalities through respective latent features [19], and adaptively fuse them together through gating mechanisms, modulating their relevant importance. Then applying the dynamic attention over temporal modelling through this combined features will allow our proposed system to benefit on the inherent sequential property of both modalities, resulting to a higher potential accuracy (which we will show in our ablation study). To the best of our knowledge, we are the first to explore such adaptive combinations of multiple modalities with attention for temporal modelling within Valence/Arousal based affective predictions.

3 Audio-Visual Gated-Sequenced Neural Network (AViGaS-NET) for Affect Recognition

Fig. 1 shows the structure of direct latent-based V/A estimator (DiLaST) that consists of a coupled Generator and

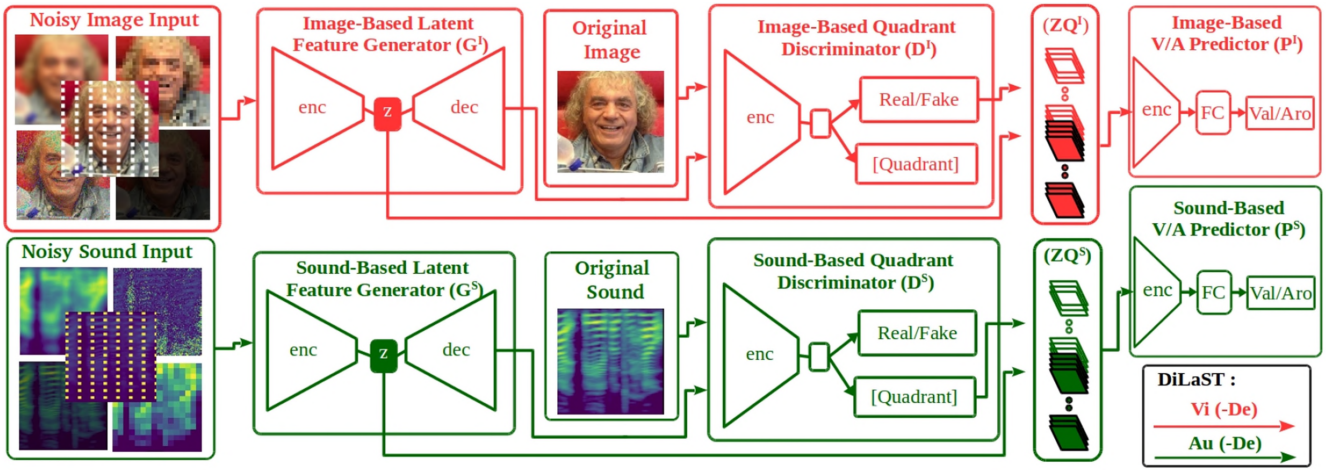

Fig. 1. Schematic representation of the pipelines of the individual modality versions of the DILaST that consists of AU-De Net and VI-De Net.

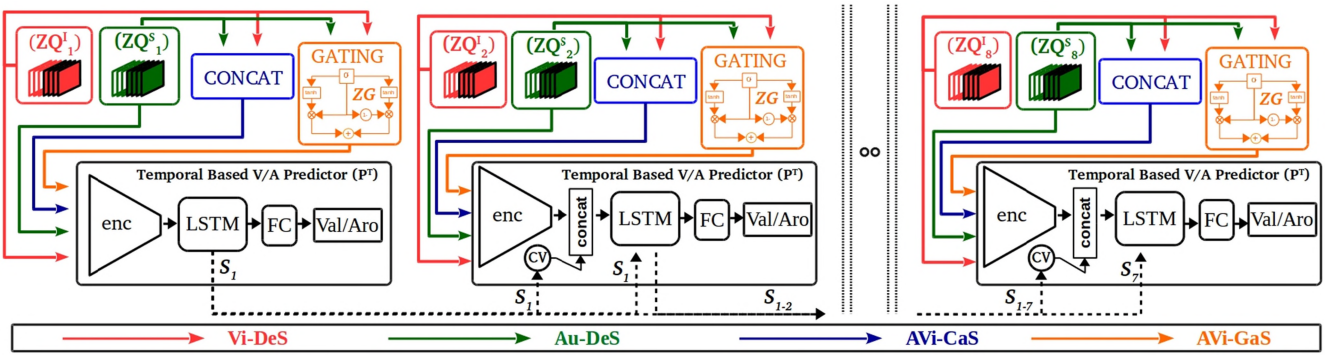

Fig. 2. Visualization of the process of our sequence-based models with attention (DILAST-SA) that constitutes AU-DeS Net, Vi-DeS Net, AVi-CaS Net, and our complete model of AVi-GaS Net (AViGaS).

Discriminators that are trained under adversarial settings to denoise input images, and subsequently create latent features. These latent features are then combined with the first step V/A quadrants and later used for final V/A estimates using the V/A estimator parts (C). In this work, we consider two modality inputs: visual and audio features. The second part of our models pipeline consists of the inclusion of two combining mechanisms that are later used as input for a sequence-based V/A estimator with attention [16]. These process produce our Sequence-based models variants (DiLaST-SA) and visualized on the Figure 2. In the end, several combinations of the two modalities, temporal modelling and combination strategies lead to possible sub-models, as listed below, which we will include in our experiments to motivate the need for our full system:

1) Au-De: direct latent-based V/A estimator using only the audio stream (without temporal modelling).
2) Vi-De: direct latent-based V/A estimator using only the video stream (without temporal modelling).
3) Au-DeS: sequence-based V/A estimator using only the audio stream.
4) Vi-DeS: sequence-based V/A estimator using only the video stream.
5) AVi-CaS: sequence-based V/A estimator using concatenated audio and video streams.
6) AVi-GaS: sequence-based V/A estimator using the gating mechanism to fuse the audio and video streams.

3.1 The Direct Latent-Based V/A Estimator (DiLaST)

The first part of of our approach uses latent features extracted through adversarial learning combined with first step V/A quadrant estimates given the noisy image inputs. We use similar approach to our previous model to perform this initial estimation [19], however, we further consider to use two different modalities separately to asses their respective impact. We use RGB images as visual input, while a spectrogram of corresponding sounds is used for the audio modality. The details of the data pre-processing of both modalities can be seen in Section 3.4. The use of similar network structures that independently process both modalities ensures that each network is able to absorb and benefit from distinct characteristics of each modality, and further facilitates our gating mechanism to weight their respective importances. These latent features are then used for the consecutive prediction task iteration that aims to infer actual Valence/Arousal values.

Specifically, the pipeline of **DiLaST** models starts with the use of Latent Feature Generator Network (G) that takes either the original inputs F, consisting of images I or sounds S, and also its distorted counterpart, $\tilde{F} \in \{\tilde{I}, \tilde{S}\}$, as described in [49], [50]. Given the noisy versions of these modalities, G estimates the cleaned reconstruction of both input modalities $\hat{F}$, along with a 2D latent representation that is utilised as features ($\mathbb{Z}$) in subsequent operations:

$$G(F)_{\Phi^G} = dec_{\Phi^G}(enc_{\Phi^G}(F)) \; with \; \mathbb{Z}^F \approx enc_{\Phi^G}(F), \quad (1)$$

where Φ are the respective networks' parameters, enc and dec constitute the encoder and decoder networks, and G consists of both G^I and G^S for the image and sound input modalities, respectively. Subsequently, the Quadrant Discriminator Network (D) receives $\hat{F}$ and predicts whether the sample was obtained from a true or fake examples (i. e., an original or distorted version of both modalities), as well as a rough estimate of the affective state in the form of a Circumplex Quadrant ($\mathbb{Q}$) [51] that discretises emotion states along the Valence and Arousal dimensions (thus into four quadrants) [20]. This multi-task setting helps the Discriminator to reach convergence during training [52] resulting in more accurate [53], [54] label predictions (namely Real and Fake identity), along with quadrant values that, together with the subsequent refinement of emotion label predictions, results in a coarse-to-find arrangement that can often benefit accuracy [55], [56]). Thus, with FC as fully connected layer:

$$D(F)_{\Phi^D} = FC_{\Phi^D}(enc_{\Phi^D}(F)) = (\mathbb{Q}^F\,,\,\Upsilon^F), \tag{2}$$

where Υ^F is a binary variable indicating whether the sample is classified as real (1) or fake (0), and D consists of D^I and D^S for each image and sound inputs. Then, we condition the extracted latent features $\mathbb{Z}$ with the estimated quadrant number ($\mathbb{Q}$) by means of layer-wise concatenation operations, which we call as $\mathbb{ZQ}$ [57], [58]. Capitalising on these conditioned latent features (which hold the extracts from previous coarse prediction task), the Valence/Arousal (V/A) Predictor Network (P) then performs the final stage of affect estimation to produce more refined (thus more precise [59]) affect predictions (Valence and Arousal values). Hence, letting $\hat{\theta}$ to denote the predicted V and A values:

$$\begin{aligned} DiLaST(F) &= P_{\Phi^P}([G_{\Phi^G}(F); D_{\Phi^D}(G_{\Phi^G}(F))]) \\ &= FC_{\Phi^P}(enc_{\Phi^P}([G_{\Phi^G}(F); D_{\Phi^D}(G_{\Phi^G}(F))])) \\ &= FC_{\Phi^P}(enc_{\Phi^P}([\mathbb{Z}^F; \mathbb{Q}^F])) = \hat{\theta}^F_{DiLaST}. \end{aligned} \tag{3}$$

The coarse-to-fine iterative prediction task used in our methods pipeline, which starts by predicting coarser emotion labels (discrete V/A quadrant number) by a Quadrant Estimator, and uses them to condition the following refinement prediction step by a Predictor Network (continuous Valence/Arousal inference), potentially increases the accuracy produced by our proposed approach [60], [61]. Given that this arrangement eases models' training through a gradual increment on task complexity [55], along with effective conditioning (through intermediate features) allows the flow of richer gradient signals, which translates into an effective learning process [56]. Finally, depending on the modality inputs, we call the DiLaST as Vi-De and Au-De Net when it uses Visual and Audio input, respectively.

3.2 Multi-Modal Fusion With Attention Enhanced Sequence Modelling (DiLaST-SA)

The compact size of ZQ extracted from the previous pipeline allows us to perform more complex processing to reach a higher accuracy. Motivated by our previous findings in [20] about the importance of sequence modelling, we propose to use such approach on both available latent features. This is reached by employing LSTM combined with attention mechanisms [16] and training with Curriculum Learning [59], [62], [63]. Our sequence modelling (**DiLaST-SA**) uses the extracted ZQ as the primary input for processing. Furthermore, alongside the use of individual ZQ to the sequence pipelines, we also propose to investigate the impact of two different fusion strategies to merge the sound ZQ^S and image ZQ^I features:

1) By direct concatenation, which has been the most popular in the field, and consists of simply concatenating both inputs ZQ as new features. Thus,

$$ZQ^{CAT} = [ZQ^I; ZQ^S]. \tag{4}$$

2) By gating mechanisms, where we use a gated multi-unit approach [13] that relates these two distinct modalities. It is calculated as follows:

$$\begin{aligned} GMU(ZQ^I, ZQ^S) &= ZQ^{GATE} \\ &= ZG(ZQ^I, ZQ^S) \odot h_v \\ &\quad + (1 - ZG(ZQ^I, ZQ^S)) \odot h_s), \end{aligned} \tag{5}$$

with ZG, h_v, and h_s calculated as:

$$ZG(ZQ^I, ZQ^S) = \sigma(W_{ZG}[ZQ^I; ZQ^S]) \tag{6}$$

$$h_v = tanh(W_v \odot (ZQ^I)^T) \tag{7}$$

$$h_s = tanh(W_s \odot (ZQ^S)^T). \tag{8}$$

Thus, the ZG coefficient controls the importance of each modality as input.

Subsequently, these features of ZQ (either ZQ^I or ZQ^S), ZQ^{CAT}, and ZQ^{GATE} will be individually fed to our LSTM modelling that is based on the P network, but with attention modules. Thus, given the ZQ as example of the input and P^T is the Temporal Based V/A Predictor, the final results of our models are:

$$\begin{aligned} &\forall n \in \mathrm{N}\,, DiLaST - SA, h_n = \\ &\qquad FC_{\Phi^{P^T}}(LSTM_{\Phi^{P^T}}([\mathbb{S}_n; \mathbb{ZQ}_n], h_{n-1})), \end{aligned} \tag{9}$$

where LSTM is the Long Short Term Memory network [64], h_n the set of LSTM states (h) after n successive frames and ($\mathbb{S}$) consists of both the LSTM inner state (c) and outgoing states (h) [65] to provide the full previous information. Here, we also adapt these techniques to consider sequences of up 8 previous states (n= 8) following our curriculum learning approach [20]. Afterwards, we utilise the context vector [16] of ($\mathbb{CV}$) that allows adaptive weighting during model inferences by summarising the importance (or relevance) of each previous state h. This is done by first calculating the alignment score that involves the combined LSTM states at frame t, denoted ($\mathbb{S}_t = [c_t, h_t]$), and n previous states ($\bar{S}$) following the formula below:

$$\begin{aligned} a_n(t) &= \mathrm{align}(\mathbb{S}_t, \bar{\mathbb{S}}_t)\,,\, with\, S_x = [h_x; c_x] \\ &= \frac{\exp\big(W_a[\mathbb{S}_t^\top; \bar{\mathbb{S}}_n]\big)}{\sum_{N'} \exp\big(W_a[\mathbb{S}_t^\top; \bar{\mathbb{S}}_{n'}]\big)}. \end{aligned} \tag{10}$$

Subsequently, we use the location-based function below to calculate the alignment scores from the previous states ($\bar{S}$):

$$a_n = \text{softmax}(W_a \bar{\mathbb{S}}). \tag{11}$$

The alignment vector is then used to quantify the context vector $\mathbb{CV}_t$ as the weighted average across the considered n preceding hidden states:

$$\mathbb{CV}_t = \frac{\sum_n a_n \odot \mathbb{S}_n}{n}. \tag{12}$$

Depending on the configurations, the above pipelines will yield three different models: i) a sequence-based single modality affect estimator (Au/Vi-Des Net), when the direct ZQ^I and ZQ^S are used as input; ii) a concatenated-based affect estimator (AVi-CaS Net), when the concatenated latent features from both the visual and sound modality (ZQ^{CAT}) are used as input; and iii) our full model of a gated affect estimator (AVi-GaS Net), when the gating mechanism is used to fuse both modalities (ZQ^{GATE}).

3.3 Training Losses

To train our G and D networks, we adopt the modified adversarial training from [19], [20] and feed their extracted features to the P network on the fly to permit simultaneous training of the latter. With this arrangement, the P network will further benefit from the improved quality of the features extracted by G and D as the training progresses. Thereby, the equations for the modified adversarial training of these networks are:

$$\begin{aligned}\mathcal{L}_{adv} =& \lambda_D \, \mathbb{E}_F[\log D(F)] \\ &+ \lambda_G \, \mathbb{E}_F[\log (1 - D(G(\tilde{F})))] + \lambda_P \, \mathbb{E}_{va}[P(F), \theta_F].\end{aligned} \tag{13}$$

The first term evaluates the discriminator (D) predictions given the real feature (F); the second terms measures the discriminator predictions given generated input from Generators; and the third terms assess the quality of affect predictions from P networks given both input features F and affect ground truth (θ_F). These three terms are controlled by their respective regularization coefficients (λ_D, λ_G and λ_F). Furthermore, the $\mathbb{E}_{va}$ is maximized by minimizing the L_{afc} losses as used in [19], [20] that integrates multiple affect metrics consisting of Rooted Mean Square Error (RMSE) (Eq. 15), Correlation (COR) (Eq. 16), Concordance Correlation Coefficients (CCC) (Eq. 17) [8] along with the Intra-class Correlation Coefficient (ICC) [7]. We chose to incorporate these combined loss because it has been shown to work better than solely relying only on RMSE [66], [67], and has been used as standard in other recent related works [68], [69]. Hence, letting $\{\hat{\theta}, \theta\}$ as the predicted and the ground truth V/A values respectively, we define the $\mathcal{L}_{afc}$ as follow:

$$\mathcal{L}_{afc} = \sum_{i=1}^{K} \frac{k_i}{K} (\mathcal{L}_{RMSE} + \mathcal{L}_{COR} + \mathcal{L}_{CCC} + \mathcal{L}_{ICC}) \tag{14}$$

$$\mathcal{L}_{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2}, \tag{15}$$

$$\mathcal{L}_{COR} = \frac{\mathbb{E}[(\hat{\theta} - \hat{\mu_\theta}) - (\theta - \mu_\theta)]}{\sigma_{\hat{\theta}} \sigma_\theta}, \tag{16}$$

$$\mathcal{L}_{CCC} = 2x \frac{\mathbb{E}[(\hat{\theta} - \hat{\mu_\theta}) - (\theta - \mu_\theta)]}{\sigma_{\hat{\theta}}^2 + \sigma_\theta^2 + (\mu_{\hat{\theta}} - \mu_\theta)^2}, \tag{17}$$

$$\mathcal{L}_{ICC} = 2x \frac{\mathbb{E}[(\hat{\theta} - \hat{\mu_\theta}) - (\theta - \mu_\theta)]}{\sigma_{\hat{\theta}}^2 + \sigma_\theta^2}, \tag{18}$$

with K as a normalisation factor [50] for the total V/A classes (it is discretised by a value of 10) and k_i is the total number of instances of discrete V/A classes i. The normalisation factor is essential in reducing the impact from large imbalance in the number of instances per class in the dataset.

3.4 Data Pre-Processing and Model Training

We use both the SEMAINE [6] and SEWA [7] datasets to train all of our proposed models by following subject-independent protocol (i.e., five fold cross validation). Using these datasets, we obtain the facial area by running a state of the art facial tracker [59]. To extract the sound features, we first calculate the whole Mel-spectrogram of the respective sound files of the video inputs. We extract the Mel-scaled spectrogram using Librosa library [70] with the parameters of a fast Fourier transform at a sample rate of 22 kHz, and with number of Mel dimensions of 128 to match the input image dimension of 128 x 128. Subsequently, we convert the obtained power spectrogram to decibel units using its maximum amplitude. We then crop the parts of spectrogram centred with the time-stamp of the input frames, with the left and right pad of half of the input image size, i. e., 64. Finally, we replicate these spectrograms into 3 channels, to allow them to be processed with a similar network structure as the one used for the visual input.

We start the training process with the DiLaST network, which involves the respective G, D, and P networks simultaneously using an adversarial loss as indicated in (13). To quantify the impact of the denoising, we also choose to train the standard DiLaST without any noisy image inputs. This stage produces our baseline consisting of individual results for each modality with and without noise modelling (Au/Vi and Au-De/Vi-De Nets, respectively) and the conditional latent features $\mathbb{ZQ}$ of each modality to be used on the sequence modelling (DiLaST-SA). This is done by the use of multi-stage transfer learning from 2, 4, and 8 [59] with attention enabled. We use the individual ZQ directly to the sequence modelling pipelines to produce the sequence variants of the original DiLaST of both modalities, i.e., Au-DeS net and Vi-DeS nets. Furthermore, we combine both ZQ altogether to be used as the input to our sequence modelling to produce the $AVi - CaS$ network results. We further use the gating mechanism to perform selective merging as the input to create our final $AVi - GaS$ networks. Lastly, we optimise the hyperparameters of G, D and P networks by using equal regularization values for G and P (i.e., λ_G, and λ_D is 0.25) but with twice the values for P networks (i.e., λ_P is 0.5). Since we found that the P networks to be slower to train than the other two, especially during the optimizations involving sequential with attention operations.

We need to mention that the training process requires extensive computation power. Thus, the use of latent features from both modalities (that is known as an effective method for reducing the dimensionality representations) is

TABLE 1
Quantitative Comparisons of Our Models Utilising Each Modality (DiLaST) on the SEMAINE Dataset

Model	RMSE ↓		COR ↑		CCC ↑		ICC ↑	
	VAL	ARO	VAL	ARO	VAL	ARO	VAL	ARO
Vi	0.268	0.315	0.364	0.238	0.350	0.233	0.368	0.235
Vi-De	0.247	0.297	0.391	0.246	0.373	0.234	0.399	0.238
Vi-DeS	**0.232**	0.289	**0.441**	0.250	**0.412**	0.234	**0.455**	0.238
Au	0.302	0.228	0.261	0.495	0.238	0.476	0.249	0.484
Au-De	0.290	0.217	0.266	0.506	0.240	0.486	0.249	0.489
Au-DeS	0.306	**0.226**	0.272	**0.509**	0.250	**0.495**	0.262	**0.500**

TABLE 2
Quantitative Comparisons of Our Models Utilising Each Modality (DiLaST) on the SEWA Dataset

Model	RMSE ↓		COR ↑		CCC ↑		ICC ↑	
	VAL	ARO	VAL	ARO	VAL	ARO	VAL	ARO
Vi	0.340	0.345	0.444	0.375	0.434	0.375	0.466	0.381
Vi-De	0.335	0.343	0.463	0.383	0.447	0.377	0.484	0.388
Vi-DeS	**0.328**	0.333	**0.501**	0.400	**0.476**	0.398	**0.520**	0.405
Au	0.363	0.341	0.391	0.483	0.379	0.467	0.386	0.480
Au-De	0.343	0.332	0.411	0.530	0.397	0.512	0.405	0.523
Au-DeS	0.342	**0.327**	0.430	**0.534**	0.420	**0.520**	0.424	**0.528**

critical to accelerate our training process, making our experiments feasible. By the use of the extracted latent features, we observe a reduction of up to a quarter of the original times required for the training each of our models, i. e., using a single NVIDIA Titan X GPU and latent features, it took us around 12 hours to fully train our models as opposed to 2 days when using the original inputs size. The training computation complexity alongside the models inference speed can further be seen in Section 4.4, where we present the respective complexity comparison between our models' variants.

4 Experimental Results

In this section, we first describe the datasets used in our experiments with the respective metrics to quantify the performance of each model in (Section 4.1). Secondly, we perform an ablation study to highlight the importance of each element of our model: we first analyse the results produced by each of our modality approaches and its correlations with the sequence modelling with attention (Section 4.2). Then, we focus on the importance of our gating mechanisms to aggregate both modalities to reach our best results (Section 4.3), followed by analysis of models complexity (Section 4.4). Lastly, Section 4.5 compares our best results with other alternatives on both of SEMAINE and SEWA datasets.

4.1 Dataset and Experiment Settings

We utilise two relevant affective datasets to provide a comprehensive analysis and comparison of our models' results: the SEMAINE [6] and SEWA [7] datasets.

- The SEMAINE dataset [6] is a large audio-visual database built from the interactions between an agent and users from stimulated settings. It consists of recordings from 150 participants with a total of 959 conversations. Alongside the emotion labels, It also includes other annotations such as race, gender, and the fully transcribed conversation scripts to allow rich data analysis.
- The SEWA dataset [7] is one of in-the-wild (captured in unconstrained settings) affect datasets that consists of both video and audio recordings involving 398 subjects across multiple cultures. It is divided into 538 sequences that include respective meta-data (e. g., subject id, culture, etc.) along with actual affect ground truth of Valence/Arousal and liking/disliking.

In each experiment, we provide the results from the variants of our models to highlight the importance of each approach. All results are reported by following the original subject-independent protocol (5-fold cross validation) for both datasets. To facilitate the quantitative comparison to other results reported in the literature, we first calculate RMSE values for both datasets as baseline estimates of models accuracy [7], [44]. However, an important drawback of the RMSE metric is that it overlooks structural information of both predicted and ground truth label throughout the input sequences (i.e., it disregards their correlation) [22], [71]. Thus, to account for this aspect, we also calculate the COR metrics along with its derivations, i.e., ICC for the SEMAINE and and CCC for the SEWA dataset, respectively.

We use all variants of our models given continuous streams of processed video and audio inputs for each dataset (as explained in Section 3.4) to obtain our models' results. Specific to sequential modelling, we perform the initialization of LSTM hidden states at the beginning of the inference process and use up to previous eight [19] inputs (including current observation) for attention modellings. This allows our models to run on arbitrary sequence range regardless of the original length used during training [59], [72], [73].

4.2 Single-Modality and Sequence Modelling Analysis

Tables 1 and 2 provide the comparisons of each of our single-modality approaches with denoising and sequential modelling. In these tables, we can see that the result of the Vi-DeS network that utilises visual input, produces higher accuracy in the Valence domain, while the Au-DeS network attains higher accuracy in the Arousal domain. These results confirm the previously reported studies [7], [8] that these modalities are more relevant to each of these domains due to the very nature of each modality.

In these results, we further notice an increase in accuracy for both of our baseline Vi and Au networks when we add the Denoiser operations. This finding is in agreement with our previous work [19], where we found that the inclusion of the denoiser improves the robustness of the learnt latent features, leading to higher accuracy. In regard to this, examples of the denoising results for both image and audio input of our models can be seen in Figure 4. Notice that our models can clean both input modalities quite well, which is remarkable considering the different characteristics of these modalities.

We also found that the activation of temporal modelling provides further improvement of the accuracy, which confirms the benefit of including such sequential inputs [20].

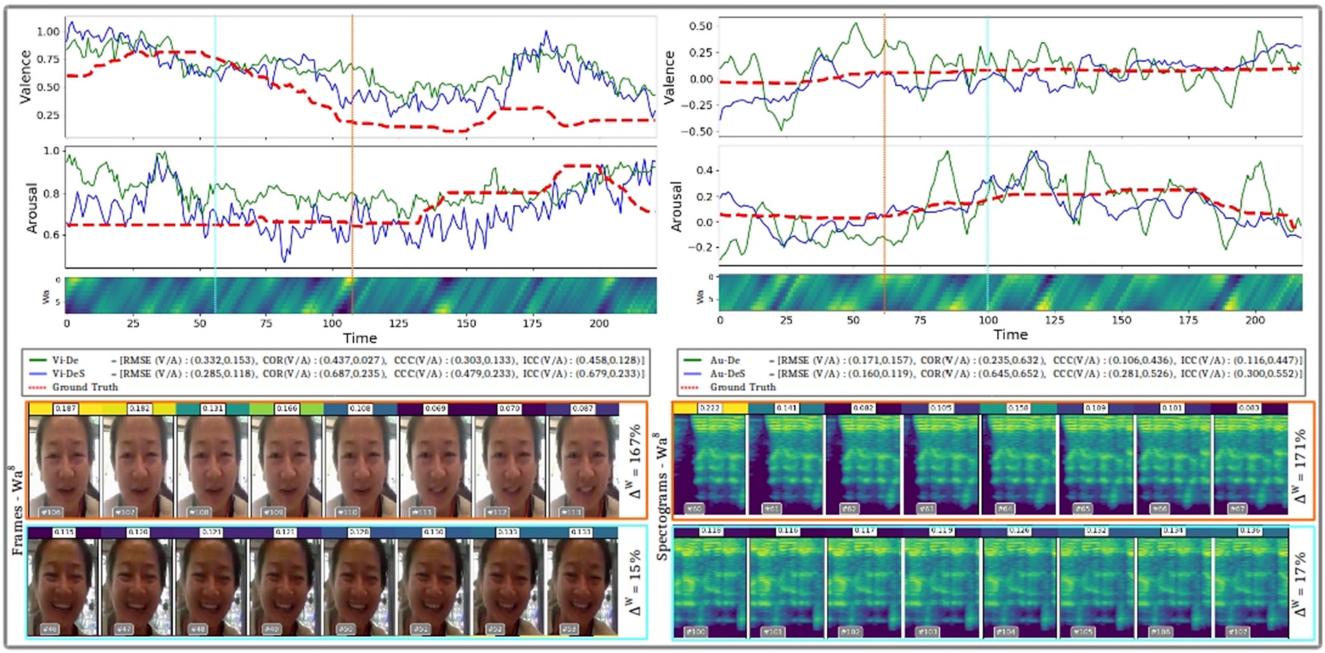

Fig. 3. The impact of attention on both, the image and sound modalities as input to our model. The left part shows examples of sequence modelling with attention improving our model estimates with regards to the change captured on the visual input. The right part shows the changes captured with our attention modelling using sound inputs. From the top, the first two horizontal graphs show the predictions of the evaluated models on Valence and Arousal respectively. The third graph shows the learnt Wa (attention weight). The fourth graphs show the legend followed by specific visualizations examples of learnt attention alongside the respective sequences of input modality.

Examples of the models predictions alongside of the learnt sequential attention can be seen in Figure 3. There, we can see the predictions examples of models with attention enabled (Vi-DeS and Au-DeS) and disabled (Vi-De and Au-De) outlined with the ground truth. We can further see that the activation of sequence modelling with attention yields more accurate predictions in both modalities (as shown in the bottom legends). This is also visually confirmed by comparing to the ground truth curve, which is also provided. As a note, we scale the graph to the min and maximum values of the evaluated models to highlight the accuracy difference between compared models (the margin would be hard to observe when using full scale of each dataset, i. e., -1 to 1 for SEMAINE and 0 to 1 for SEWA).

The observed accuracy improvements can be attributed to the sequential attention mechanisms that allow the network to focus on the 'important' parts within sequences [16], [20] through the learnt Wa (attention weights). This weight activation of each sub-sequence(i. e., eight input frames/audio, Cf Section 3.2) is shown on the middle part of Figure 3. Notice the different weight activation patterns throughout the overall inference durations with rapid changes of intensity within sub-sequences indicating that the network is allocating its focus on the salient parts from particular subsequence (thus potentially benefiting from the attention mechanisms). In contrast, while the Wa activations is rather uniform (i. e., difference in weight activations are small), the network evenly distributes the attention weights throughout all of sub-sequence inputs (thus in this case, the role of attention to models prediction will be modest). To visualize these two conditions, we introduce the coefficient Δ^{Wa} that is calculated from the percentage of the disparity between the minimum and maximum w values for each sequence. Thus resulting in the first case with high Δ^{Wa} and the latter with low value of Δ^{Wa}. Subsequently, we show them on the last two pictures for both visual (left) and audio modality (right) inputs. The first rows (in orange) show the first case (high Δ^{Wa}), and the bottom rows (in blue) show the second case (low Δ^{Wa}). In these two contrasting examples, we can observe the correlations between higher attention intensity with higher changes observed in the input. For instance, in the first row, we see that both, the facial input and the spectrogram changed slightly compared to the second row (blue), and this is reflected in the respective attention intensities on

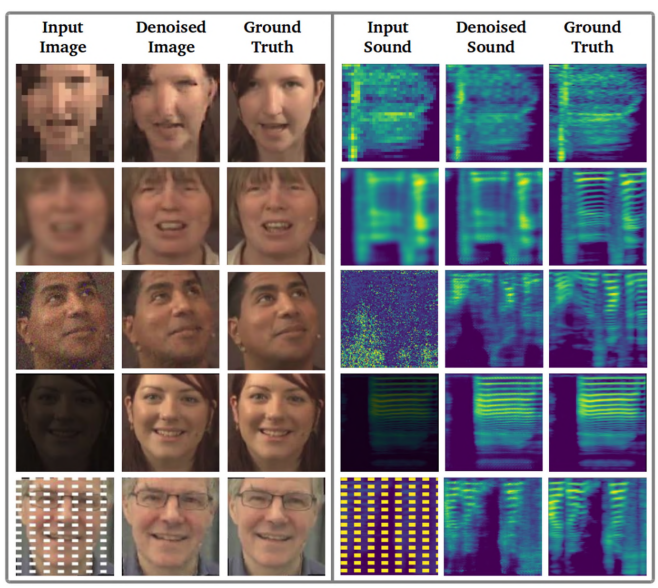

Fig. 4. Visual examples of our denoised input of both modalities. Columns 1 and 4 show the noisy inputs. Next, columns 2 and 5 show the corresponding denoised examples of our models. Finally, columns 3 and 6 show the ground-truth, e. g., the clean versions.

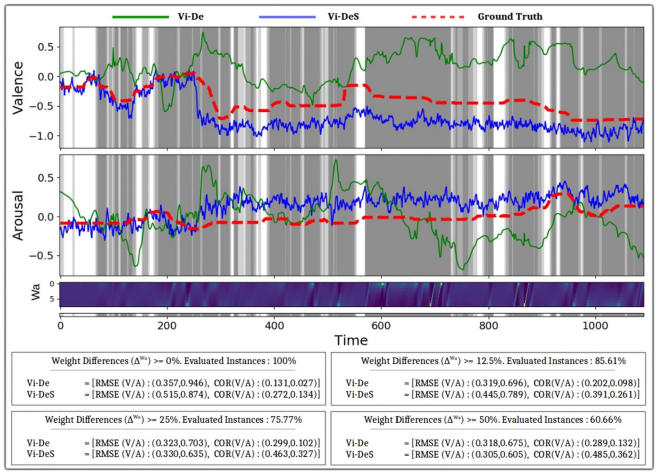

Fig. 5. The examples of the results from Vi-DeS using several W value differences (Δ^{Wa}).

top of each figure. This ability to capture such changes within the sequence is in line with our previous findings [20]; however in this case, we further found that the associated accuracy improvement is more pronounced (the accuracy of Vi/Au-DeS is more accurate compared to Vi-AuD) in the first example as opposed to the latter, as visualized in the graph. This suggests that the benefit of the attention mechanism is higher when it is activated strongly within sub-sequences (i. e., the higher Δ^{Wa}, the higher the improvements made).

To quantitatively confirm the above analysis, we consider three levels of differences for Δ^{Wa}: 12.5%, 25%, and 50%, and evaluate the portion of the datasets with Δ^{Wa} activation weights above each level. Visual examples, as well as the portion size and accuracy computed for each level Δ^{Wa} can be seen in Figure 5. We see that in general, there is a decrease in the size of the frames included as the Δ^{Wa} accompanied by a raise in accuracy. Tables 3 and 4 further show the accuracy of our single-based models without (Au-De and Vi-De) and with the sequences (Au-DeS and Vi-DeS) for different Δ^{Wa} values. Based on these results, we observe that the accuracy gain (in terms of RMSE and COR) increases, as we raise Δ^{Wa} for both datasets. For instance, the highest accuracy gain at $\Delta^{W} = 12.5$ is 16% compared to more than 119% when the threshold is higher ($\Delta^{W} = 50\%$). This indicates that attention impacts the results more, when the variation of the attention weights are (relatively) high. Furthermore, we also see that the gain is pretty balanced across modalities, suggesting its compatibility to both types of input modalities.

4.3 The Impact of the Multi-Modal Approach With Concatenation and Internal Gating Modelling

Tables 5 and 6 present the results of our multi-modal approaches by means of concatenation (AVi-CaS) and Gating (AVi-GaS) together with the comparison with the best preforming results from previous sections of each modality. In this comparison, we can see that the combination of these modalities yields an increase in accuracy for both approaches with more balanced results in both domains. This suggests the benefit of aggregating these modalities. However, comparing the results of AVi-GaS with AVi-CaS, we find that in general, the results from our gating mechanisms are better than the basic concatenation approach. This supports the need of more sophisticated approaches to combine these modalities.

Examples of the effectiveness of our gating approach compared to the standard concatenation counterparts are visualised in Figure 6, where we can see more accurate predictions of our gating mechanisms over the other compared models. The respective bottom sections provide two different examples of learnt ZG coefficients that are able to 'control' the importance of each modality. That is, in the the first column, we can see examples where the higher values of ZG indicate changes detected in the visual features. This is also synchronously detected by the sequence attention. We also see the other instances, where ZG is able to detect changes in the sound features, thus giving a higher priority to this modality. We also see, again that these coefficients correlate well with the sequence activation, in line with the perceived activation of ZG. All of this explains the quite substantial improvement on accuracy of the AVi-GaS network compared to the marginal improvement achieved by the AVi-CaS network with respect to the single-modality variants.

Analogously to the analysis of high level of attention weight activations in the previous section, we evaluate the importance of ZG using different thresholds T_{ZG}. Specifically, we chose three different T_{ZG} thresholds (0.25, 0.125, and 0.0625) that affect the range of T_{ZG}. However, because

TABLE 3
The Relative Impact of Attention on the Level of Relative Differences (Δ^{Wa}) on the Involved Sequences of the SEMAINE Dataset

Model	Δ^{Wa}>=12.5%				Δ^{Wa}>=25%				Δ^{Wa}>=50%			
	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)
Vi-De	(0.47,0.48)	—	(0.31,0.15)	—	(0.42,0.49)	—	(0.35,0.24)	—	(0.55,0.59)	—	(0.12,0.09)	—
Vi-DeS	(0.45,0.46)	⇓ (3.54%,2.63%)	(0.32,0.16)	⇑ (2.80%,6.49%)	(0.38,0.45)	⇓ (8.33%,8.52%)	(0.38,0.28)	⇑ (7.71%,16.7%)	(0.49,0.52)	⇓ (10.5%,11.9%)	(0.18,0.15)	⇑ (62.2%,46.1%)
Au-De	(0.40,0.32)	—	(0.19,0.45)	—	(0.35,0.25)	—	(0.15,0.21)	—	(0.29,0.29)	—	(0.07,0.11)	—
Au-DeS	(**0.39**,**0.30**)	⇓ (3.6%,7.1%)	(**0.2**,**0.53**)	⇑ (10.8%,16.9%)	(**0.30**,**0.20**)	⇓ (15.2%,18.9%)	(**0.20**,**0.35**)	⇑ (34.5%,61.3%)	(**0.14**,**0.14**)	⇓ (51.2%,49.4%)	(**0.12**,**0.25**)	⇑ (65.6%,119%)

TABLE 4
The Relative Impact of Attention on the Level of Relative Differences (Δ^{Wa}) on the Involved Sequences of the SEWA Dataset

Model	Δ^{Wa}>=12.5%				Δ^{Wa}>=25%				Δ^{Wa}>=50%			
	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)
Vi-De	(0.36,0.37)	—	(0.46,0.36)	—	(0.35,0.36)	—	(0.43,0.32)	—	(0.44,0.40)	—	(0.22,0.18)	—
Vi-DeS	(0.34,0.37)	⇓ (0.68%,4.54%)	(0.47,0.38)	⇑ (0.95%,3.44%)	(0.31,0.34)	⇓ (6.83%,10.9%)	(0.45,0.36)	⇑ (6.08%,11.1%)	(0.38,0.33)	⇓ (13.5%,18.4%)	(0.40,0.32)	⇑ (78.8%,81.4%)
Au-De	(0.46,0.45)	—	(0.19,0.29)	—	(0.40,0.47)	—	(0.25,0.34)	—	(0.55,0.59)	—	(0.09,0.12)	—
Au-DeS	(**0.45**,**0.44**)	⇓ (2.73%,3.84%)	(**0.20**,**0.31**)	⇑ (7.24%,5.35%)	(**0.37**,**0.44**)	⇓ (6.19%,8.32%)	(**0.29**,**0.40**)	⇑ (18.2%,15.5%)	(**0.47**,**0.50**)	⇓ (15.7%,15.6%)	(**0.13**,**0.16**)	⇑ (40.7%,30.6%)

TABLE 5
The Results of Our Multi-Modal Approach of Employing Concatenation and Gating Mechanisms Compared to the Single-Modality-Based Approaches on the SEMAINE Dataset

Model	RMSE ↓			COR ↑			CCC ↑			ICC ↑		
	VAL	ARO	AVG	VAL	ARO	AVG	VAL	ARO	AVG	VAL	ARO	AVG
Vi-DeS	0.232	0.289	0.260	0.441	0.250	0.346	0.412	0.234	0.323	0.455	0.238	0.347
Au-DeS	0.306	0.226	0.266	0.272	0.509	0.390	0.250	0.495	0.372	0.262	0.500	0.381
AVi-CaS	0.239	0.219	0.229	0.459	0.516	0.487	0.405	0.507	0.456	0.466	0.512	0.489
AVi-GaS	**0.224**	**0.180**	**0.202**	**0.618**	**0.656**	**0.637**	**0.587**	**0.642**	**0.615**	**0.600**	**0.650**	**0.625**

TABLE 6
The Results of Our Multi-Modal Approach of Employing Concatenation and Gating Mechanisms Compared to the Single-Modality-Based Approaches on the SEWA Dataset

Model	RMSE ↓			COR ↑			CCC ↑			ICC ↑		
	VAL	ARO	AVG	VAL	ARO	AVG	VAL	ARO	AVG	VAL	ARO	AVG
Vi-DeS	0.328	0.333	0.331	0.501	0.400	0.450	0.476	0.398	0.437	0.520	0.405	0.462
Au-DeS	0.342	0.327	0.334	0.430	0.534	0.482	0.420	0.520	0.470	0.424	0.528	0.476
AVi-CaS	0.321	0.312	0.317	0.541	0.510	0.525	0.525	0.502	0.513	0.535	0.506	0.521
AVi-GaS	**0.282**	**0.282**	**0.282**	**0.697**	**0.604**	**0.651**	**0.686**	**0.583**	**0.634**	**0.693**	**0.589**	**0.641**

we are now evaluating a gating block, we are interested in deviations around the central value (i. e., 0.5); for instance, with $T_{ZG} = 0.125$, we evaluate $ZG < 0.125$ and $ZG > 0.875$. Examples of the considered segments using these different thresholds with their respective accuracy can be seen in Figure 7. Similar to the analysis in the previous section, we can see that increasing the threshold reduces the amount of data that is considered, but also raises the associated accuracy.

Tables 9 and 10 show the quantitative results of our $AVi - GaS$ with respect to the different threshold T_{ZG} for both, SEMAINE and SEWA, while the results of different Δ^{Wa} can be seen in the Tables 7 and 8, respectively. Notice that the results of the AVi-GaS networks are consistently better than those of the other models, including the concatenation-based approach AVi-CaS. An analysis of the ZG values reveals that the accuracy improvement grows as the threshold values decrease (which implies using only the most activated ZG values), showing the benefit of the gating approach in successfully controlling each modality to boost performance. For example, the highest gains with the threshold at 0.25 are 13.4% and 17.5%, but they grow to 35% and 20% for the SEMAINE and SEWA datasets using the narrowest threshold value of 0.0625.

4.4 Analysis of Models Complexity and Running Time

Table 11 shows the number of network parameters, size and the running speed in both Training and Test for all of our models variants. We obtain these information by evaluating the models using a single workstation under UNIX environment, using NVIDIA 1080-TI GPU, with 32 GB of RAM and Intel Core i7-4770 CPU. During inference, we observe that all of our models require less resources (less number of parameters and smaller size) with faster running time compared to training (half of number of parameters, quarters of memory space and tenth of running time). This is because we only need partial parts of networks to perform the inference (for instance we do not need Discriminator and only half part of Generator to create latent features for P networks) and with no requirements to calculate the gradients required for optimizations.

Specifically, our baseline models (Vi-De/Au-De) are the least complex requiring the smallest amount of parameters (i.e., also the smallest size and faster running time) with our complete model AVi-GaS in the opposite end. Here we observe a slight increase in the number of parameter as we progressively add the sequential modelling (Vi-DeS/Au-

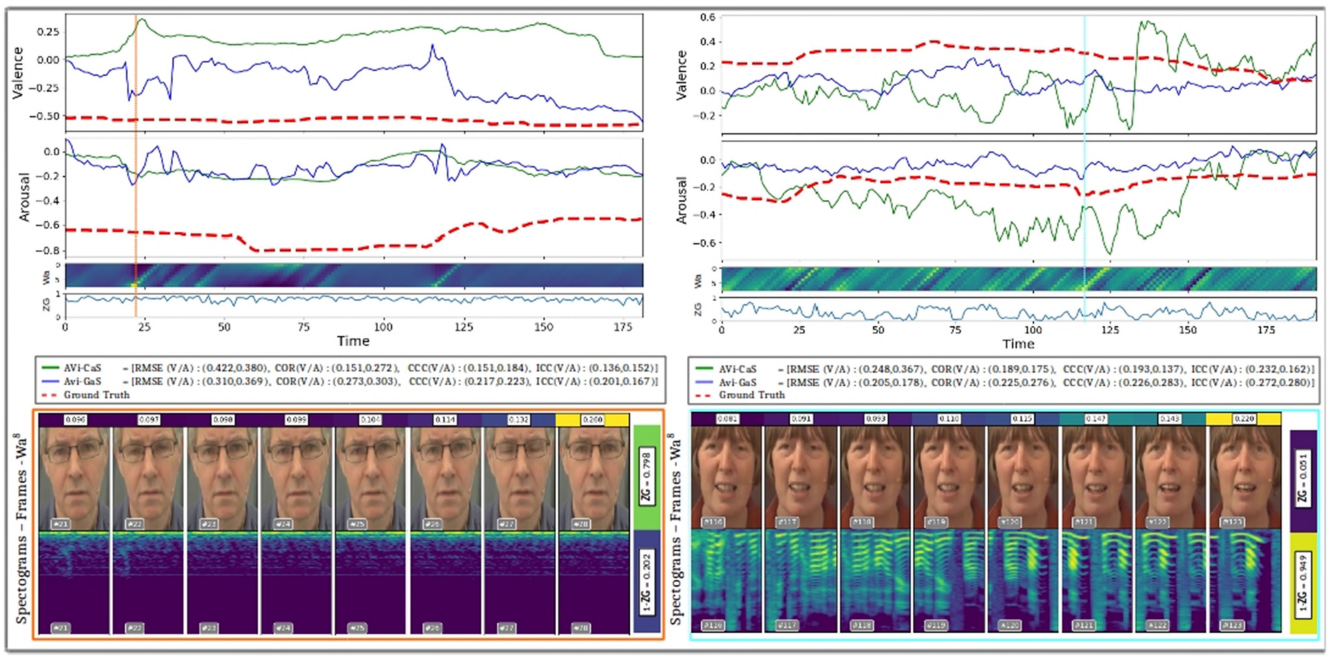

Fig. 6. Example of the results of our model with the concatenation (AViCaS) and gating (AviGaS) approaches. The bottom left and right show examples of how the internal ZG of the AViGaS network detected the change that happened on the visual and audio input, respectively. The first two graphs show the predictions of the evaluated models on Valence and Arousal respectively. The middle graphs show the learnt ZG (gating coefficients) during the predictions. The fourth graphs show the legend followed by the visualizations examples of learn attention and gating coefficients on the specific sequences for each input modality.

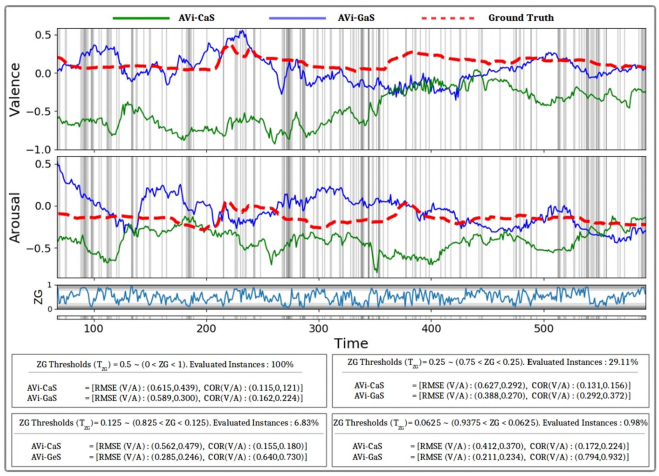

Fig. 7. Visualisation of the results on AVi-GaS using several threshold (T_{ZG}) values: 0.5, 0.25, 0.125, and 0.0625. The first row shows the example of the area of each respective T_{ZG} value. The second row provides the associated accuracy.

DeS), using both modalities (Avi-CaS), and finally incorporating gating mechanisms (Avi-GaS). However, we notice that in the end, the complexity differences between our models are rather small. For example, there is only less than 12% margin (4.63 to 5.21 ms) of inference time between AVi-GaS and our baselines, but with substantially higher accuracy for the former (in some cases more than double, see Tables 5 and 6). These differences are even less pronounced during training, since we need to optimise all network parts (less than 5% margin on running time, 54.63 to 56.35 ms). Thus we argue that these small sacrifices in complexity are justified since it allows our models to perform more effective data processing, leading to our state of the art accuracy.

Lastly, we can see that in overall our models demand modest computation resources with relatively fast inference speed. Using our full model of AVi-GaS, for instance, we only need to allocate about 15 MB of GPU VRAM to perform inference with more than 190 fps, which is quite accessible given current computation standards.

4.5 Comparison to the State of the Art

In this section, we present the comparison of our best performing models from the previous ablation analysis (AVi-GaS network), including the results of our single-modality-based models (the Au-DeS, and Vi-DeS networks) against other alternative approaches on both, the SEMAINE and SEWA datasets. Specifically, we compare our model to the following ones:

1) FT-DCNN [8], a hybrid deep learning-based system that uses handcrafted visual and geometrical facial features.

TABLE 7
The Results of Our Model's Variant Compared to Our Full Model of AVi-GaS by Their Level of Relative Differences (Δ^{Wa}) on the SEWA Dataset

Model	Δ^{Wa}>=12.5%				Δ^{Wa}>=25%				Δ^{Wa}>=50%			
	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)
Vi-DeS	(0.42,0.45)	⇓ (7.50%,21.9%)	(0.17,0.13)	⇑ (147%,136%)	(0.40,0.44)	⇓ (17.9%,21.4%)	(0.33,0.24)	⇑ (70.1%,56.2%)	(0.41,0.41)	⇓ (22.6%,19.8%)	(0.41,0.33)	⇑ (25.6%,44.6%)
Au-DeS	(0.49,0.42)	⇓ (24.8%,13.1%)	(0.18,0.19)	⇑ (65.6%,91.4%)	(0.43,0.40)	⇓ (24.2%,11.1%)	(0.26,0.32)	⇑ (80.1%,28.2%)	(0.40,0.41)	⇓ (19.2%,20.9%)	(0.36,0.40)	⇑ (62.0%,27.6%)
AVi-CaS	(0.45,0.42)	⇓ (20.0%,17.3%)	(0.25,0.21)	⇑ (27.7%,44.9%)	(0.40,0.41)	⇓ (22.2%,17.0%)	(0.36,0.34)	⇑ (35.1%,17.8%)	(0.39,0.39)	⇓ (21.7%,17.6%)	(0.44,0.39)	⇑ (30.1%,24.1%)
AVi-GaS	(**0.37**,**0.36**)	—	(**0.35**,**0.39**)	—	(**0.33**,**0.35**)	—	(**0.55**,**0.42**)	—	(**0.32**,**0.33**)	—	(**0.63**,**0.51**)	—

TABLE 8
The Results of Our Model's Variant Compared to Our Full Model of AVi-GaS by Their Level of Relative Differences (Δ^{Wa}) on the SEMAINE Dataset

Model	Δ^{Wa}>=12.5%				Δ^{Wa}>=25%				Δ^{Wa}>=50%			
	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)
Vi-DeS	(0.43,0.46)	⇓ (5.87%,14.6%)	(0.37,0.34)	⇑ (2.70%,9.17%)	(0.42,0.43)	⇓ (21.52%,20.6%)	(0.40,0.35)	⇑ (16.01%,26.2%)	(0.30,0.36)	⇓ (7.25%,27.82%)	(0.41,0.30)	⇑ (22.5%,82.7%)
Au-DeS	(0.42,0.40)	⇓ (4.14%,2.49%)	(0.29,0.30)	⇑ (31.5%,21.9%)	(0.46,0.43)	⇓ (28.52%,20.9%)	(0.33,0.34)	⇑ (41.92%,28.7%)	(0.38,0.30)	⇓ (26.7%,13.96%)	(0.32,0.35)	⇑ (57.1%,54.8%)
AVi-CaS	(0.43,0.40)	⇓ (4.50%,0.89%)	(0.33,0.36)	⇑ (16.8%,2.37%)	(0.41,0.36)	⇓ (19.60%,6.56%)	(0.40,0.42)	⇑ (15.69%,4.43%)	(0.33,0.29)	⇓ (15.0%,11.49%)	(0.48,0.42)	⇑ (4.24%,29.6%)
AVi-GaS	(**0.41**,**0.39**)	—	(**0.38**,**0.37**)	—	(**0.33**,**0.34**)	—	(**0.46**,**0.44**)	—	(**0.28**,**0.26**)	—	(**0.50**,**0.54**)	—

TABLE 9
The Results of Our Model's Variant With Respect to a Different Threshold T_{ZG} of Learnt ZG on the SEMAINE Dataset

Model	T_{ZG} = 0.25,($ZG < 0.25, ZG > 0.75$)				T_{ZG} = 0.125, ($ZG < 0.125, ZG > 0.875$)				T_{ZG} = 0.0625, ($ZG < 0.0625, ZG > 0.9375$)			
	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)
Au-DeS	(0.32,0.29)	⇓ (6.70%,6.29%)	(0.43,0.46)	⇑ (4.71%,1.39%)	(0.34,0.31)	⇓ (11.9%,14.9%)	(0.44,0.48)	⇑ (14.5%,9.42%)	(0.31,0.30)	⇓ (4.79%,15.3%)	(0.46,0.50)	⇑ (22.1%,16.4%)
Vi-DeS	(0.30,0.31)	⇓ (1.90%,12.1%)	(0.44,0.41)	⇑ (2.81%,13.4%)	(0.30,0.33)	⇓ (2.03%,19.9%)	(0.46,0.41)	⇑ (9.90%,27.4%)	(0.31,0.34)	⇓ (4.61%,24.3%)	(0.45,0.43)	⇑ (24.8%,35.8%)
AVi-CaS	(0.31,0.28)	⇓ (4.68%,1.88%)	(0.44,0.45)	⇑ (3.39%,3.34%)	(0.38,0.37)	⇓ (21.3%,28.6%)	(0.44,0.47)	⇑ (15.4%,12.1%)	(0.33,0.31)	⇓ (9.45%,18.3%)	(0.47,0.49)	⇑ (21.63%,17.8%)
AVi-GaS	(**0.30**,**0.27**)	—	(**0.45**,**0.46**)	—	(**0.30**,**0.27**)	—	(**0.50**,**0.53**)	—	(**0.30**,**0.26**)	—	(**0.57**,**0.58**)	—

TABLE 10
The Results of Our Model's Variant With Respect to a Different Threshold T_{ZG} of Learnt ZG on the SEWA Dataset

Model	T_{ZG} = 0.25,($ZG < 0.25, ZG > 0.75$)				T_{ZG} = 0.125, ($ZG < 0.125, ZG > 0.875$)				T_{ZG} = 0.0625, ($ZG < 0.0625, ZG > 0.9375$)			
	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)	RMSE (V,A) ↓	GAIN (V,A)	COR (V,A) ↑	GAIN (V,A)
Au-DeS	(0.41,0.40)	⇓ (10.6%,7.01%)	(0.35,0.40)	⇑ (17.3%,6.57%)	(0.43,0.41)	⇓ (13.3%,15.1%)	(0.43,0.45)	⇑ (4.48%,1.62%)	(0.43,0.40)	⇓ (18.4%,6.27%)	(0.39,0.41)	⇑ (22.1%,4.11%)
Vi-DeS	(0.39,0.38)	⇓ (5.75%,2.03%)	(0.41,0.38)	⇑ (0.50%,12.5%)	(0.38,0.41)	⇓ (1.86%,13.7%)	(0.42,0.36)	⇑ (7.32%,25.3%)	(0.38,0.41)	⇓ (8.15%,8.89%)	(0.41,0.36)	⇑ (19.3%,20.9%)
AVi-CaS	(0.38,0.38)	⇓ (2.55%,3.87%)	(0.41,0.40)	⇑ (2.44%,7.97%)	(0.40,0.38)	⇓ (6.52%,8.36%)	(0.41,0.41)	⇑ (7.79%,11.6%)	(0.40,0.39)	⇓ (13.8%,5.68%)	(0.40,0.39)	⇑ (20.9%,10.4%)
AVi-GaS	(**0.37**,**0.37**)	—	(**0.42**,**0.43**)	—	(**0.38**,**0.35**)	—	(**0.45**, **0.45**)	—	(**0.35**,**0.37**)	—	(**0.48**,**0.43**)	

TABLE 11
The Parameters, Size and Running Time of Our Models Variants

Model	No. of Parameters		Size (in MB)		Running time (in ms)	
	Training	Inference	Training	Inference	Training	Inference
Vi-De / Au-De	5.287.308	2.483.346	59.35	11.15	54.63	4.63
Vi-DeS / Au-DeS	5.550.989	2.747.027	60.41	11.61	55.17	4.81
AVi-CaS	6.000.508	3.196.546	62.26	13.41	55.88	5.17
AVi-GaS	6.264.189	3.460.227	63.33	14.46	56.35	5.21

2) BLSTM-WS [39], a sequence-based neural network that utilises both, direct sound wave input and its spectrogram derivatives as input to LSTM networks.
3) DialogueRNN [37], a deep learning model that uses multiple modalities such as visual, sound, and text features that are aggregated using GRU networks for modelling. Additionally, they also include the interaction properties on their modelling.
4) ResNet-18 [7] is a CNN-based model (with residual connection) that operates directly on the video frames to produce emotion estimates.
5) Tensor [29], a tensor-based neural network that processes visual input and is optimised using the tucker tensor regression.
6) Factorised [9], a deep learning model that uses a similar approach to [29], but uses generalised factorisations to allow for more efficient decomposition.
7) AEG-CD-SZ [19], our previous latent-based approach that also applies adversarial training, using both, visual and sound modalities.
8) ANCLaF-SA [20] the precursor of our sequential modelling with attention, relying only on the visual input.

Tables 12 and 13 provide the comparisons of the evaluated models on the SEMAINE and SEWA dataset, respectively. In general, we can see that the results of our models compare favourably with respect to other methods on both datasets. Specifically, our single modality-based models (Vi-DeS and Au-DeS) perform quite well and are able outperform some of the other alternatives, especially in terms of COR and its variants (ICC and CCC) which detail the models ability to produce correlated results among sequence of prediction and actual emotion label. For example, our single-modality based models' results are higher in terms of both COR and CCC values against FT-DCNN [8] on the SEMAINE dataset, likewise, in comparison to ResNet-18 [7] both in terms of COR and ICC values on the SEWA datasets. This highlights the effectiveness of our base models in exploiting individual modality input to produce accurate results with structurally sound (correlated) predictions in relations to actual ground-truth labels.

TABLE 12
Quantitative Comparisons on the SEMAINE Dataset

Model	Modalities	RMSE ↓			COR ↑			ICC ↑		
		VAL	ARO	AVG	VAL	ARO	AVG	VAL	ARO	AVG
FT-DCNN [8]	VIS	**0.161**	0.217	**0.189**	0.283	0.296	0.290	0.331	0.277	0.304
BLSTM-WS [39]	VIS	-	-	-	**0.680**	0.506	0.593	-	-	-
DialogueRNN [37]	VIS+AUD+TXT	-	-	-	0.350	0.590	0.470	-	-	
AEG-CD-SZ [19]	VIS+AUD	0.303	0.262	0.283	0.175	0.301	0.238	0.173	0.291	0.232
ANCLaF-SA [20]	VIS	0.258	0.297	0.278	0.410	0.279	0.345	0.423	0.298	0.360
Vi-DeS	VIS	0.232	0.289	0.260	0.441	0.250	0.346	0.455	0.238	0.347
Au-DeS	AUD	0.306	0.226	0.266	0.272	0.509	0.390	0.262	0.500	0.381
AVi-GaS	VIS+AUD	0.224	**0.180**	0.202	0.618	**0.656**	**0.637**	**0.600**	**0.650**	**0.625**

TABLE 13
Quantitative Comparisons on the SEWA Dataset

Model	Modalities	RMSE ↓			COR ↑			CCC ↑		
		VAL	ARO	AVG	VAL	ARO	AVG	VAL	ARO	AVG
ResNet-18 [7]	VIS	-	-	-	0.350	0.350	0.350	0.350	0.290	0.320
Tensor [29]	VIS	0.334	0.391	0.363	0.503	0.439	0.471	0.469	0.392	0.431
Factorised [9]	VIS	**0.240**	0.320	**0.280**	**0.840**	0.600	**0.720**	**0.750**	0.520	**0.635**
AEG-CD-SZ [19]	VIS+AUD	0.323	0.350	0.337	0.442	0.478	0.460	0.405	0.430	0.418
ANCLaF-SA [20]	VIS	0.336	0.328	0.332	0.558	0.332	0.445	0.405	0.529	0.467
Vi-DeS	VIS	0.328	0.333	0.331	0.501	0.400	0.450	0.476	0.398	0.437
Au-DeS	AUD	0.342	0.327	0.334	0.430	0.534	0.482	0.420	0.520	0.470
AVi-GaS	VIS+AUD	0.282	**0.282**	0.282	0.697	**0.604**	0.651	0.686	**0.583**	0.634

Subsequently, in comparison to our previous approaches (AEG-CD-SZ [19] and ANCLaF-SA [20]), we first observe that our models' average results (of both the Valence and Arousal domains) are better compared to AEG-CD-SZ with healthy accuracy gain, albeit their individual accuracy for the Valence and Arousal dimension are relatively less balanced (which could be attributed to their lack of multi-modal input, which AEG-CD-SZ uses). Secondly, in comparison with ANCLaF-SA [20], we see that our results are slightly better, especially in comparison with Vi-DeS, where both are achieving very high accuracy on the Valence domain due to their similar single-modal based architectural designs.

The utilisation of our complete pipeline (AVi-GaS) results in large accuracy improvements (cf. Section 4.3) allowing to substantially outperform both of our previous methods (AEG-CD-SZ and ANCLaF-SA) with state of the art results against other approaches. Here we can see that AVi-GaS achieves comparable accuracy in terms of RMSE metrics (that provide a rough overview of models' prediction accuracy) relative to the current top performer result on both datasets, with relatively low margin of differences. As an example, the average RMSE margin between our model against FT-DCNN [37] is less than 0.14 on SEMAINE and around 0.03 against Factorised [9] on SEWA, thus, it is fairly negligible. These results demonstrate general competitiveness of our models results across datasets. We also see similar outcomes (with our single-modality results) in terms of models' results correlation capacity (judged by the COR, CCC, and ICC metrics) but overall with slightly higher accuracy. That is, our AVi-GaS results are better compared to other alternatives on certain datasets (e. g., SEMAINE dataset where our results are better overall) and emotion dimensions (especially for Arousal). For instance, our models' accuracies surpass the results of all alternatives on the SEMAINE dataset for both the COR and ICC values, while on SEWA, it attains the highest COR and CCC values for Arousal, and rank only slightly lower than the Factorised model [9] on the Valence emotion label, which is highly concentrated in processing visual input (it thus explains its high accuracy on the Valence dimension). In this respect, we need to note, however that the Factorised model [9] involves far larger CNN models than we do by evaluating three different sub-networks, including ResNet18 [74] as baseline with more than 11 million parameters, almost doubling ours during training, and quadrupling during inference (cf. Section 4.4) to process the visual input. This fairly large size network will thereby hinder its real world application to accommodate such size in on the fields operation compared to our relatively lighter size.

In addition, the AVi-GaS (that utilises multi-modal input) further manages to produce quite balanced accuracy

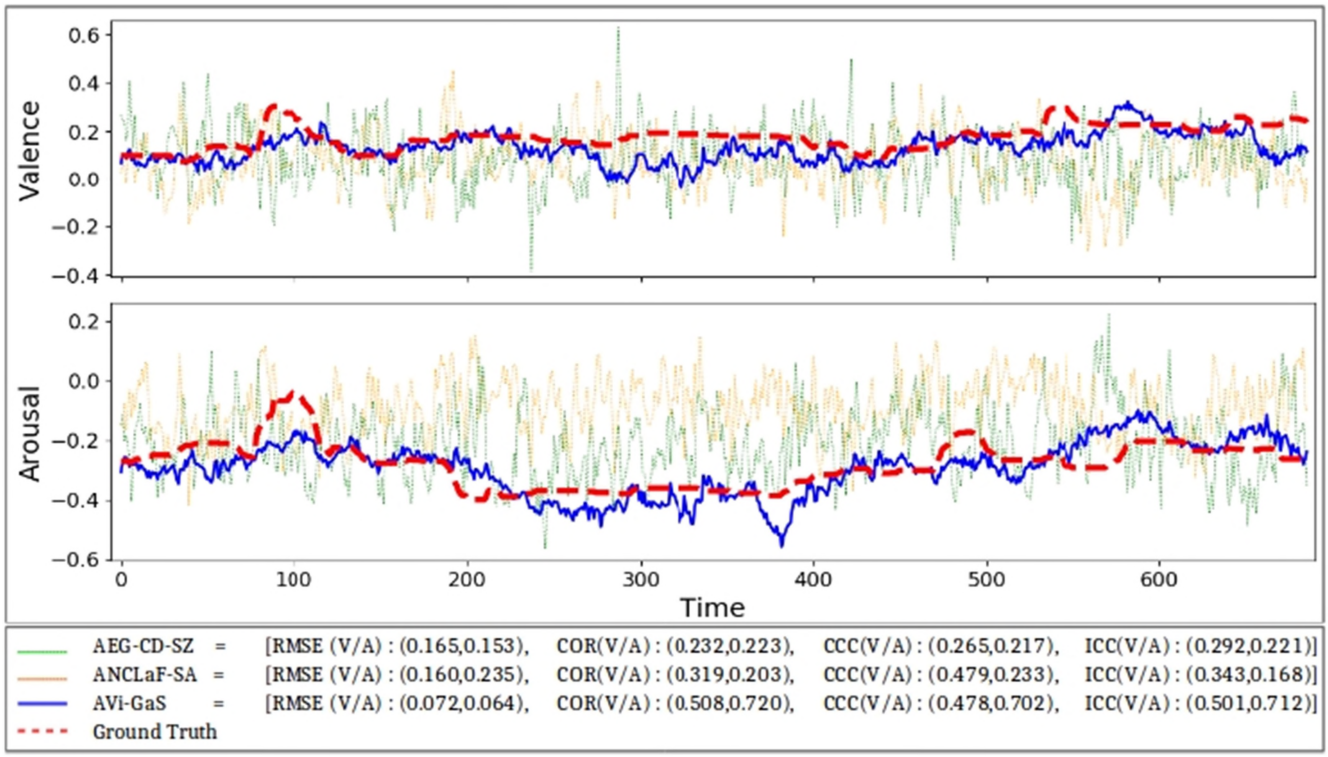

Fig. 8. The comparison of our gated multi-modal sequence model approach (AviGaS) versus our previous approaches (AEG-CD-SZ and ANCLaF-SA). Notice that the current proposed approach is able to produce accurate predictions on both emotion dimensions and outperforms our previous results.

on both emotional dimensions (Valence / Arousal) when compared to the current state of the art, and also appears more accurate than our previous approach (AEG-CD-SZ, where it utilises similar modalities). For instance, the DialogueRNN [37] (that achieves lowest RMSE scores) produces less accuracy in Valence on COR metrics on the SEMAINE dataset, whereas the Factorised [9] method (one of the top methods on the SEWA dataset) lacks the accuracy in the Arousal domain. In contrast, our models consistently attain high level of accuracy in both, the Valence and Arousal domains across metrics and datasets with higher accuracy on average against AEG-CD-SZ. This highlights the effectiveness of our approach to aggregate [19] these modalities in support of each other allowing them to arrive at such accuracy balance (cf. Section 6). This is desirable, because precise estimations of both Valence and Arousal are deemed necessary to properly pinpoint the exact emotion values [51]. All of these results demonstrate the effectiveness of our approach in achieving competitive results against the state of the art, while simultaneously permitting them to maintain a high level of efficiency and consistency to accurately predict both emotion labels (Valence/Arousal) that others currently lack.

Finally, to further show the improvement of our approach (AVi-GaS) in comparison to our previous results (AEG-CD-SZ and ANCLaF-SA), we plot their prediction examples in the Figure 8. Notice that our current results are more accurate than our previous results for both datasets. Furthermore, we can also see that the predictions of AVi-GaS are also more stable, which could be attributed to our internal sequence modellings that benefit from the inherent temporal information. This is especially noticeable when observing the results of AEG-CD-SZ which does not include any temporal modelling (as evidently shown with its lowest COR and ICC values). ANCLaF-SA on other hands, performs better in Valence domain with higher affect metrics values, although it is not as stable as our AVi-GaS. This could be due to the lack of additional modality which helps the models in performing their predictions (cf. Section 4.3) in terms of raw accuracy gained (versus its individual modality versions) and its balance across affect domain (both Valence and Arousal), but in this case, it also improves the models stability. Overall, all of these findings highlight the importance of our proposed methods to yield more accurate, stable, and balanced affect predictions that ultimately translate to more reliable predictions.

5 Conclusions

In this work, we presented multi-modal affect networks that are capable to efficiently process bi-modal inputs using our combined latent-based representations with sequence modelling and attention mechanisms. We then equipped our networks with gating mechanisms to allow for more effective multi-modal fusion. We trained our models using adversarial learning to extract more representative latent features given the noisy inputs of both visual and sound modality. We then used these latent features from both modalities fused together with our gating mechanisms, and fed the result to our sequential modelling, which was trained through curriculum learning to allow for progressive training.

We demonstrated the effectiveness of our approach as a whole as well as for each of its components, on the two most widely used and accessible affective datasets: SEMAINE and

SEWA. In our ablation studies, we firstly showed the impact of denoising to improve the base results of our single-modal input models, and we also observed the consistently cleaner results from noisy modality inputs; secondly, we find that sequence modelling with attention further improved the results in terms of accuracy and provided a detailed quantitative analysis to support this conclusion by thresholding on the relative learnt attention differences; thirdly, we observed a noticeable gain in accuracy when both modalities were merged, either by concatenation or by our gating mechanism, the latter producing the highest improvement; lastly, we showed that in general our models require modest computation resources with relatively fast inference speed to current computational standards.

We further compare our best performing models against current state of the art alternatives on both datasets, including our two previous approaches. In the comparison, we show that our model is able to consistently produce high accuracy in majority of quantitative metrics, comparable to the top results from the state of the art, with an outstanding balance in the performance obtained for Valence and Arousal emotion dimension estimates with respect to alternative approaches.

Future efforts may consider other types of attention and an extension to other modalities such as physiology or textual cues. It will be exciting to see the benefits of the proposed architecture in a self-learning potentially cross-modal context.

References

[1] S. Duo and L. Song, "An e-learning system based on affective computing," *Phys. Procedia*, vol. 24, pp. 1893–1898, 2010.

[2] C. Liu, K. Conn, N. Sarkar, and W. Stone, "Online affect detection and robot behavior adaptation for intervention of children with autism," *IEEE Trans. Robot.*, vol. 24, pp. 883–896, Aug. 2008.

[3] J. Comas, D. Aspandi, and X. Binefa, "End-to-end facial and physiological model for affective computing and applications," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Los Alamitos, CA, USA, May 2020, pp. 1–8. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/FG47880.2020.00001

[4] C. Joaquim, D. Aspandi, M. Ballester, F. Carreas, L. Ballester, and X. Binefa, "Short-term impact of polarity therapy on physiological signals in chronic anxiety patients," in *Proc. IEEE 9th Int. Conf. Bioinf. Comput. Biol.*, 2021, pp. 180–186.

[5] D. Aspandi, S. Doosdal, V. Ülger, L. Gillich, and S. Staab, "User interaction analysis through contrasting websites experience," 2022, *arXiv:2201.03638*.

[6] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Proc. IEEE Int. Conf. Multimedia*, 2010, pp. 1079–1084.

[7] J. Kossaifi *et al.*, "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," 2019, *arXiv:1901.02839*.

[8] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "Afew-va database for valence and arousal estimation in-the-wild," *Image Vis. Comput.*, vol. 65, pp. 23–36, 2017.

[9] J. Kossaifi, A. Toisoul, A. Bulat, Y. Panagakis, T. M. Hospedales, and M. Pantic, "Factorized higher-order CNNs with an application to spatio-temporal emotion estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6060–6069.

[10] S. Chen and Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proc. 5th Int. Workshop Audio/Visual Emotion Challenge*, 2015, pp. 49–56.

[11] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, and Y. Zong, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, vol. 309, pp. 27–35, 2018.

[12] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Trans. Multimedia*, vol. 14, pp. 597–607, 2012.

[13] J. Arevalo, T. Solorio, M. Montes-y Gomez, and F. A. González, "Gated multimodal networks," *Neural Comput. Appl.*, vol. 32, no. 14, pp. 10209–10228, 2020.

[14] M. K. Tellamekala and M. Valstar, "Temporally coherent visual representations for dimensional affect recognition," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact.*, 2019, pp. 1–7.

[15] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual LSTM network," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 176–183.

[16] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.

[17] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition," *Inf. Process. Manage.*, vol. 57, no. 3, 2020, Art. no. 102185.

[18] W. Xiaohua, P. Muzi, P. Lijuan, H. Min, J. Chunhua, and R. Fuji, "Two-level attention with two-stage multi-task learning for facial emotion recognition," *J. Vis. Commun. Image Representation*, vol. 62, pp. 217–225, 2019.

[19] D. Aspandi, A. Mallol-Ragolta, B. Schuller, and X. Binefa, "Latent-based adversarial neural networks for facial affect estimations," in *Proc. 15th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, Los Alamitos, CA, USA, May 2020, pp. 348–352. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/FG47880.2020.00053

[20] D. Aspandi, F. Sukno, B. Schuller, and X. Binefa, "An enhanced adversarial network with combined latent features for spatio-temporal facial affect estimation in the wild," in *Proc. 16th Int. Conf. Comput. Vis. Theory Appl.*, 2020, pp. 172–181.

[21] F. Povolny *et al.*, "Multimodal emotion recognition for AVEC 2016 challenge," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, New York, NY, USA, 2016, p. 75–82. [Online]. Available: https://doi.org/10.1145/2988257.2988268

[22] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 92–105, Apr.–Jun. 2011.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "Openear-introducing the munich open-source emotion and affect recognition toolkit," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, 2009, pp. 1–6.

[24] B. Pogorelc, Z. Bosnić, and M. Gams, "Automatic recognition of gait-related health problems in the elderly using machine learning," *Multimedia Tools Appl.*, vol. 58, no. 2, pp. 333–354, 2012.

[25] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *Proc. ISCA Tut. Res. Workshop Speech Emotion*, 2000, pp. 19–24.

[26] P. Ruvolo and J. Movellan, "Automatic cry detection in early childhood education settings," in *Proc. 7th IEEE Int. Conf. Develop. Learn.*, 2008, pp. 204–208.

[27] P. V. Rouast, M. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Trans. Affective Comput.*, vol. 12, no. 2, pp. 524–543, Apr.–Jun. 2021.

[28] H. Gunes and M. Pantic, "Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners," in *Proc. Int. Conf. Intell. Virtual Agents*, Berlin, Heidelberg: Springer, 2010, pp. 371–377.

[29] A. Mitenkova, J. Kossaifi, Y. Panagakis, and M. Pantic, "Valence and arousal estimation in-the-wild with tensor methods," in *Proc. 14th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2019, pp. 1–7.

[30] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb.2001.

[31] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *IEEE Trans. Multimedia*, vol. 8, pp. 500–508, 2006.

[32] M. H. Bindu, P. Gupta, and U. Tiwary, "Cognitive model-based emotion recognition from facial expressions for live human computer interaction," in *Proc. IEEE Symp. Comput. Intell. Image Signal Process.*, 2007, pp. 351–356.

[33] S. Tivatansakul, M. Ohkura, S. Puangpontip, and T. Achalakul, "Emotional healthcare system: Emotion detection by facial expressions using japanese database" in *Proc. 6th Comput. Sci. Electron. Eng. Conf.*, 2014, pp. 41–46.

[34] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou, "Analysing affective behavior in the first abaw 2020 competition," 2020, *arXiv:2001.11409*.

[35] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, pp. 1–20, 2018.

[36] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, Apr.–Jun. 2021.

[37] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[38] J. Zhou, G. Wang, Y. Yang, and P. Chen, "Speech emotion recognition based on rough set and SVM," in *Proc. 5th IEEE Int. Conf. Cogn. Informat.*, 2006, vol. 1, pp. 53–61.

[39] Z. Yang and J. Hirschberg, "Predicting arousal and valence from waveforms and spectrograms using deep neural networks," in *Proc. Interspeech*, 2018, pp. 3092–3096.

[40] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.

[41] F. Eyben *et al.*, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affective Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016.

[42] M. Valstar *et al.*, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 3–10.

[43] F. Ringeval *et al.*, "AVEC 2017: Real-life depression, and affect recognition workshop and challenge," in *Proc. 7th Annu. Workshop Audio/Visual Emotion Challenge*, 2017, pp. 3–9.

[44] F. Ringeval *et al.*, "AVEC 2019 workshop and challenge: State-of-mind, detecting depression with Ai, and cross-cultural affect recognition," in *Proc. 9th Int. Audio/Visual Emotion Challenge Workshop*, 2019, pp. 3–12.

[45] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3, 2019, Art. no. 100004.

[46] J. Xie, R. Girshick, and A. Farhadi, "Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, Cham: Springer, 2016, pp. 842–857.

[47] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Computer Vision - ECCV'98*, H. Burkhardt and B. Neumann, Eds. Berlin, Heidelberg:Springer Berlin Heidelberg, 1998, pp. 484–498.

[48] W. Zhang, H. Huang, M. Schmitz, X. Sun, H. Wang, and H. Mayer, "Effective fusion of multi-modal remote sensing data in a fully convolutional network for semantic labeling," *Remote Sens.*, vol. 10, no. 1, pp. 1–14, 2018.

[49] D. Aspandi, O. Martinez, F. Sukno, and X. Binefa, "Robust facial alignment with internal denoising auto-encoder," in *Proc. 16th Conf. Comput. Robot Vis.*, 2019, pp. 143–150.

[50] D. Aspandi, O. Martinez, and X. Binefa, "Heatmap-guided balanced deep convolution networks for family classification in the wild," in *Proc. 14th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, May 2019, pp. 1–5.

[51] J. A. Russell, "A circumplex model of affect," *J. Pers. Social Psychol.*, vol. 39, no. 6, 1980, Art. no. 1161.

[52] W. Ding and L. He, "Mtgan: Speaker verification through multitasking triplet generative adversarial networks," *Proc. Interspeech*, 2018, pp. 3633–3637.

[53] Q. Lin, L. Liang, Y. Huang, and L. Jin, "Learning to generate realistic scene Chinese character images by multitask coupled GAN," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, Cham: Springer, 2018, pp. 41–51.

[54] Y. Liu, Z. Wang, H. Jin, and I. Wassell, "Multi-task adversarial network for disentangled feature learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3743–3751.

[55] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1263–1272.

[56] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[57] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional GAN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2989–2998.

[58] H. Ye, G. Y. Li, B.-H. F. Juang, and K. Sivanesan, "Channel agnostic end-to-end learning based communication systems with conditional GAN," in *Proc. IEEE Globecom Workshops*, 2018, pp. 1–5.

[59] D. Aspandi, O. Martinez, F. Sukno, and X. Binefa, "Fully end-to-end composite recurrent convolution network for deformable facial tracking in the wild," in *Proc. 14th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2019, pp. 1–8.

[60] J.-J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3691–3700.

[61] D. Triantafyllidou and A. Tefas, "Face detection based on deep convolutional neural networks exploiting incremental facial part learning," in *Proc. 23rd Int. Conf. Pattern Recognit.*, 2016, pp. 3560–3565.

[62] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, New York, NY, USA, 2009, pp. 41–48. [Online]. Available: https://doi.org/10.1145/1553374.1553380

[63] D. G. A. Farhadi and D. Fox, "Re 3: Real-time recurrent regression networks for visual tracking of generic objects," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 788–795, Apr. 2018.

[64] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[65] C. Kim, F. Li, and J. M. Rehg, "Multi-object tracking with neural gating using bilinear LSTM," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 200–215.

[66] D. Le, Z. Aldeneh, and E. M. Provost, "Discretized continuous speech emotion recognition with multi-task deep recurrent neural network," in *Proc. Interspeech*, 2017, pp. 1108–1112.

[67] G. Trigeorgis *et al.*, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 5200–5204.

[68] D. Kollias, M. A. Nicolaou, I. Kotsia, G. Zhao, and S. Zafeiriou, "Recognition of affect in the wild using deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1972–1979.

[69] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," 2019, *arXiv:1910.04855*.

[70] B. McFee *et al.*, "librosa/librosa: 0.8.0," Zenodo, Feb. 2020, doi: 10.5281/zenodo.6097378.

[71] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "Deep learning vs. kernel methods: Performance for emotion prediction in videos," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2015, pp. 77–83.

[72] D. Gordon, A. Farhadi, and D. Fox, "Re 3: Real-time re current regression networks for object tracking," *IEEE Robot. Automat. Lett.*, vol. 3, no. 2, pp. 788–795, 2018.

[73] D. Aspandi, O. Martinez, F. Sukno, and X. Binefa, "Composite recurrent network with internal denoising for facial alignment in still and video images in the wild," *Image Vis. Comput.*, vol. 111, 2021, Art. no. 104189.

[74] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6450–6459.

Decky Aspandi received the bachelor's degree in computer science from Mulawarman University, Indonesia, in 2012, the MSc degree in computer engineering from the King Mongkut's University of Technology Thonburi, Thailand, in 2014, and the PhD degree in information and communication technologies from the Universitat Pompeu Fabra, Barcelona, in 2021. He is currently a postdoctoral researcher with the University of Stuttgart. His research interests include machine vision and deep learning topics along with their real life applications.

Federico Sukno received the degree in electrical engineering from La Plata National University, Argentina, in 2000, and the PhD degree in biomedical engineering from Zaragoza University, Spain, in 2008. In 2020, he was an associate professor with the Department of Information and Communication Technologies, Universitat Pompeu Fabra. His research interests include image analysis with statistical models of shape and appearance, targeting diverse applications within facial biometrics, affective computing, and medical imaging. He was awarded the Marie Curie Fellowship in 2012 and Ramon & Cajal Fellowship in 2015.

Björn W. Schuller (Fellow, IEEE) received the Diploma, Doctoral, Habilitation degrees in electrical engineering and information technology from the Technical University of Munich (TUM), Munich, Germany. He is currently an adjunct teaching professor of machine intelligence and signal processing with TUM. He is also a full professor of artificial intelligence and the head of GLAM with Imperial College London, U.K., a full professor and the chair of embedded intelligence for health care and well-being with the University of Augsburg, Germany, co-founding CEO and current CSO of audEERING – an Audio Intelligence company-based near Munich and in Berlin, Germany, a guest professor with Southeast University, Nanjing, China, and a permanent visiting professor with HIT, China, amongst other professorships and affiliations.

Xavier Binefa received the degree in mathematics from the Universitat de Barcelona in 1976, the degree in computer engineering from the Universitat Autònoma de Barcelona in 1988, and the PhD degree in computer vision from the Universitat Autònoma de Barcelona in 1996. He is currently an associate professor with the Department of Information and Communication Technologies, Universitat Pompeu Fabra.