Are 3D Face Shapes Expressive Enough for Recognising Continuous Emotions and Action Unit Intensities?

Mani Kumar Tellamekala, Ömer Sümer, Björn W. Schuller, Elisabeth André, Timo Giesbrecht, Michel Valstar

Abstract—Recognising continuous emotions and action unit (AU) intensities from face videos requires a spatial and temporal understanding of expression dynamics. Existing works primarily rely on 2D face appearances to extract such dynamics. This work focuses on a promising alternative based on parametric 3D face shape alignment models, which disentangle different factors of variation, including expression-induced shape variations. We aim to understand how expressive 3D face shapes are in estimating valence-arousal and AU intensities compared to the state-of-the-art 2D appearance-based models. We benchmark four recent 3D face alignment models: ExpNet, 3DDFA-V2, DECA, and EMOCA. In valence-arousal estimation, expression features of 3D face models consistently surpassed previous works and yielded an average concordance correlation of .739 and .574 on SEWA and AVEC 2019 CES corpora, respectively. We also study how 3D face shapes performed on AU intensity estimation on BP4D and DISFA datasets, and report that 3D face features were on par with 2D appearance features in AUs 4, 6, 10, 12, and 25, but not the entire set of AUs. To understand this discrepancy, we conduct a correspondence analysis between valence-arousal and AUs, which points out that accurate prediction of valence-arousal may require the knowledge of only a few AUs.

Index Terms—Facial Expression Analysis, Dimensional Affect Recognition, Action Unit Intensity Estimation, 3D Morphable Models

1 Introduction

Facial expressions are important social signals produced through coordinated movements of different facial muscle groups along spatio-temporal dimensions. Automatic recognition of facial expressions from video data is a fundamental task in Affective Computing with a wide range of applications, including but not limited to psychotherapy and well-being [1], educational analytics [2], naturalistic human-computer [3], and human-robot interaction [4]. The problem of automated facial expressive behaviour analysis has been extensively studied in the last two decades [5], [6], [7], [8]. The two most common video-based facial expression analysis approaches are based on Russell's circumplex model of dimensional emotions [9] and Facial Action Coding System (FACS) [10]. The circumplex model represents emotions in a continuous space composed of two orthogonal axes, namely valence and arousal dimensions. In contrast, FACS encodes the movements of different facial muscle groups by defining the occurrence and intensity values of their corresponding Action Units (AUs).

A common challenge encountered in video-based facial expression analysis in naturalistic conditions is to disentangle expression-induced facial variations from a wide range of other factors of variation in a given 2D face image sequence. Expression-irrelevant facial variations typically include head pose changes, facial geometry that contains identity information, or fine-scale details such as wrinkles, etc. In the era of deep representation learning, most state-of-the-art methods depend on end-to-end learning from 2D face appearances. Such 2D appearance-based expression features achieved impressive performance on both valence-arousal and AU intensity estimation tasks [11], [12], [13].

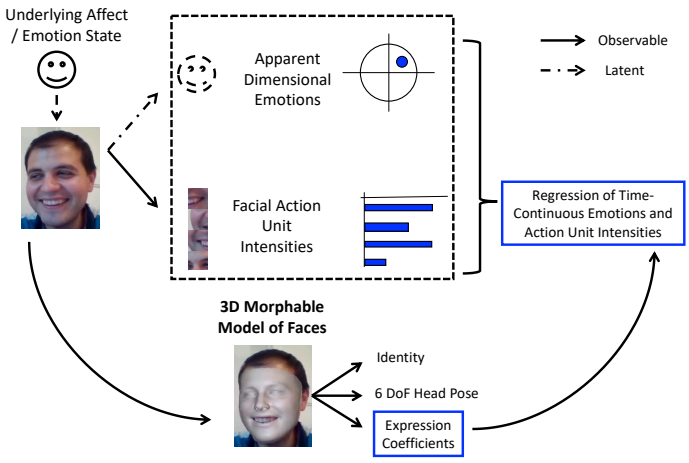

Fig. 1: Recognition of dimensional emotions (valence-arousal) and facial action unit intensities directly from the expression coefficients of 3D Morphable Models (3DMM).

- *Mani Kumar Tellamekala and Michel Valstar are with the Computer Vision Lab, School of Computer Science, University of Nottingham, UK. E-mail:{mani.tellamekala,michel.valstar}@nottingham.ac.uk*
- *Ömer Sümer and Elisabeth André are with the Chair for Human-Centered Artificial Intelligence, University of Augsburg, Germany. E-mail:{oemer.suemer, andre}@@informatik.uni-augsburg.de*
- *Björn W. Schuller is with the Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany, and GLAM – the Group on Language, Audio, & Music, Imperial College London, UK Email: schuller@uni-a.de*
- *Timo Giesbrecht is with Unilever R&D Port Sunlight, UK. Email: timo.giesbrecht@unilever.com*

Manuscript submitted on July 04, 2022

However, they heavily rely on the manual annotations of emotions or AU intensities for vast amounts of visual data to extract the facial expression features and their temporal dynamics. On the other hand, analysis-by-synthesis methods such as 3D Morphable Models (3DMM) [14] of faces offer an interesting alternative to distil the expression-induced facial shape variations in a more principled approach. Such analysis-by-synthesis methods, most importantly, do not need the labels of emotions or AU intensities to extract expression features from 2D face image sequences.

Several parametric 3D face alignment models [15], [16], [17], [18] based on 3DMM formulation achieved significant improvements in recent years by leveraging the advancements in data-driven representation learning. Some recent works on 3D face alignment methods [17], [18], [19] even attempted to reconstruct the facial expressions with high fidelity. Despite such advancements in 3D face alignment solutions, to the best of our knowledge, the idea of utilising 3DMM expression information for video-based facial expression analysis received limited attention compared to 2D appearance-based approaches. Inspired by a recent attempt to study the impact of 2D face alignment on expression analysis tasks [20], we pose the following questions in this work:

- Are 3D face shapes expressive enough to estimate AU intensities as well as dimensional emotions (valence-arousal) from face video data?
- Where do 3D face shape expression features stand w.r.t. 2D face appearance features that are directly learned for estimating AU intensities and dimensional emotions in an end-to-end fashion?

To answer these questions, as Fig. 1 illustrates, we train AU intensity estimation and dimensional emotion recognition based on the temporal dynamics of 3D facial expressions. We extensively evaluate the quality of 3DMM based expression features on the datasets of valence-arousal estimation (SEWA [21], AVEC 2019 CES [7]) and AU intensity estimation (BP4D [22] and DISFA [23]). We apply a simple bi-directional GRU network to model the temporal dynamics of 3DMM expression features extracted from four dense 3D face alignment models: ExpNet [19], 3DDFA-V2 [24], DECA [17], and EMOCA [18]. We compare the recognition performance of different 3D face shape models with the 2D face appearance baselines and models that currently have state-of-the-art performance on both tasks.

Our experimental analysis shows that in the case of continuous emotion recognition, 3D face expression features outperform the existing benchmarks as well as the 2D appearance baselines evaluated in this work. However, on the task of AU intensity prediction, 3D face shape models perform poorly compared to the existing state-of-the-art benchmarks based on appearance features. Further, we conduct a correspondence analysis between different AUs and valence-arousal dimensions to explain the performance discrepancy of 3D face models between emotion recognition and AU intensity prediction tasks. Thus, this work comprehensively illustrates the current state of the 3D face shape expression features in terms of their ability to model video-based facial expression dynamics.

2 Background and Related Work

As the main focus of our work is the analysis of 3D face shape models for video-based facial expression analysis, we review the literature on 3D morphable models of faces and expression analysis tasks tackled by using 3D face alignment models. Here, our particular interest is face geometry-aware approaches in video-based facial expression analysis tasks: valence-arousal estimation and facial action unit intensity estimation in videos.

2.1 3D Morphable Models of Faces

Estimating 3D shape models from 2D measurements is a fundamental problem in computer vision. Mainly focusing on face analysis, Blanz and Vetter [14] initially addressed this problem and proposed a 3D Morphable Model (3DMM) to generate 3D face shape and appearance. 3DMM can be considered a representation of facial shape and colour, separating them from external factors. 3DMM is a statistical approach learned from dense one-to-one correspondences of representative 3D shapes and 2D appearance data. The original work performed face registration from unregistered 3D scenes using gradient-based optical flow and created a 3D face model, learning an optimisation problem to synthesise 2D appearance from a linear combination of 3D shapes using PCA decomposition. We refer the interested readers to [25] for a detailed review of 3D morphable face models.

The main factors of variation in 3DMM are geometric shape and texture. The original formulation can be given as follows:

$$S = \bar{S} + A_s\alpha_s + A_t\alpha_t, \tag{1}$$

where S is a 3D face, $\bar{S}$ is the mean 3D shape, $A_{shape}, A_{texture}$ are geometric shape and texture bases, and $\alpha_{shape}, \alpha_{texture}$ are the parameters. After the 3D face is reconstructed with this model, it is projected back to the image plane using a scale orthographic model:

$$V_{2d}(\mathbf{p}) = f * \mathbf{Pr} * \mathbf{R} * (\bar{\mathbf{S}} + \mathbf{A}_s\alpha_s + \mathbf{A}_t\alpha_t) + t_{2d}, \tag{2}$$

where $V_{2d}(\mathbf{p})$ generates the 2D locations of 3D model vertices. The head pose information comes from the scale factor f, orthographic projection matrix, $\mathbf{Pr}$ and rotation matrix, $\mathbf{R}$. α_s contains identity-related shape information, whereas texture details by α_t may correspond to other information that can also contain facial expression. However, the original and many following 3DMM approaches focused on reconstructing 3D shapes per identity and neglected the facial expression variations specifically in their optimisation.

2.1.1 Modeling Facial Expressions in 3DMM

Looking into the methods that incorporate facial expression information in 3DMM, 3DFFA [26] trained the bases, for instance, A_e, on the 3D models of faces with various expressions on the FaceWarehouse database [27] containing 3D scans of 150 people of diverse ages and ethnic background. Subsequently, [24] improved the parameter optimisation by Vertex Distance Cost (VDC) and Weighted Parameter Distance Cost (WPDC) and used more compact backbone regressors; however, expression modelling remained the same.

Similar to [26], Chang et al. [19] proposed a landmark-free approach. They estimated 3DMM shape and pose parameters using CNN-based models. By leveraging the identity information and assuming that the shape parameters of a person's different images will remain the same, they acquired the expression deformation using Gauss-Newton optimisation. Subsequently, they extracted expression codes on large-scale face datasets and regressed them with ResNet-101 deep network architecture.

Recently, Li et al. (FLAME model, [28]) used a large amount of training data to capture intrinsic shape deformations and the deformation related to pose chances, to some extent, modelling the muscle activations and respective facial expressions. In order to decouple expressions from pose variations, they estimated pose coefficients, applied an inverse transform, and normalised pose to reduce its effect on expression parameters. Feng et al. (DECA, [17]), by building on top of the FLAME model, disentangled static and dynamic facial details utilising in the wild images. After reconstructing the course shape, they swap the person-specific details and jaw pose parameters between the different images of the same person and disentangle from expression.

2.2 3D Face Models in Expression Analysis

Our main focus is on the use of 3D face models in video-based facial expression recognition tasks, namely, valence-arousal and FACS action unit intensity estimation. This section presents a brief summary of the previous works that used 3D face features to address these problems.

As alternatives to the standard 2D appearance-based facial expressive features (e.g. [29], [30], [31], [32]), the features derived from 3D face alignment (3D landmarks [33] and their displacements [34], parametric forms of 3D face shapes [35], [36], [37]), were explored for shape-based expression analysis approaches. Several implicit and explicit decoupling methods were used for disentangling facial expression information from the identity information of 3D faces in the existing works. For example, linear disentanglement methods were used, such as the one proposed in [38] to leverage the intrinsically low-rank property of the consecutive frames of a face video to decouple the identity from residual expressions. In contrast, 3DMM [39] used explicit factorisation models that disentangle camera parameters, 3D head pose, and identity variables from the expression-induced shape variations.

Note that our method is not the first to use 3DMM expression coefficients as facial expression features for emotion recognition. Several works [33], [34], [35], [36] in the past already explored the applications of 3DMM face models to discrete and continuous emotion recognition tasks. For instance, to recognise categorical discrete emotions, Chen et al. [35], and Koujan et al. [36] proposed to jointly learn 3D face reconstruction from 2D images and the emotion recognition modules in an end-to-end fashion. In [35], 3D facial shape information was fused with image data, whereas in [36], 28-dimensional expression coefficients of the 3DMM model are directly used for emotion recognition. Similarly, in [34], first, and second-order temporal differential vectors (87D) of 3DMM expression coefficients were used as input features for discrete and continuous emotion recognition models. In [33], emotion recognition models used 3D facial landmarks and were first projected onto 2D face images for more accurate feature extraction.

Based on [17], recently, Danecek, Black and Bolkart [18] proposed a perceptual emotion consistency loss between the emotion features of input images and those of rendered ones. They trained an emotion recognition model on RGB images and then optimised L_2 loss between DECA emotion embedding (the output of a multilayer perceptron based on emotion and detail vectors) and previously trained image-based embedding.

In contrast to the existing methods, our objective is to systematically evaluate the current state of recent 3D face alignment models on video-based facial expression analysis. Towards this objective, in valence-arousal estimation and action unit intensity estimation, we focus on the usability of 3D face models' expression embeddings in video-based learning.

3 3D Shape vs. 2D Appearance Features for Continuous Facial Expression Analysis

The face is essentially a 3D volumetric surface that undergoes rigid (e.g., head pose changes) and non-rigid (e.g., talking and raising eyebrows) deformations. Capturing such non-rigid deformations that correspond to emotional expressions along spatio-temporal dimensions is at the core of video-based facial expressive behaviour analysis. Here, our goal is to comprehensively compare and analyse the performance of standard 2D CNN-based facial appearance features learned using task-specific target labels and expression-related facial features derived from dense 3D face alignment models. To this end, we model the temporal dynamics of 3D shape-based features and 2D appearance-based features extracted from face image sequences for learning video-based facial expression analysis tasks. In particular, we consider time-continuous dimensional emotion (valence-arousal) recognition and action unit (AU) intensity estimation as representative tasks for video-based facial expression analysis. In this comparison, it is worth noting that expression features in 3D face models are learned with the objective of accurate shape reconstruction, whereas 2D face appearance features are directly optimised to predict the task-specific target label sequences (valence-arousal or AU intensities).

3.1 Expression Embeddings from 3D Face Shapes

For extracting expression-specific 3D face shape features, we consider 3D Morphable Models (3DMM) [39] of faces, for they offer a principled approach to factorise the facial expression information. In the standard linear representation of 3DMM used in face alignment, as shown in Eq. 1, the shape component can be further decomposed as follows:

$$\mathbf{A}_s\alpha_s = \mathbf{A}_{id}\alpha_{id} + \mathbf{A}_{ex}\alpha_{ex}, \tag{3}$$

where $\mathbf{A}_{id}$ and $\mathbf{A}_{ex}$ are the basis matrices of face shapes and expressions, whereas α_{id} and α_{ex} are their corresponding coefficient vectors. Here, we refer to the coefficient vectors α_{ex} as expression embeddings.

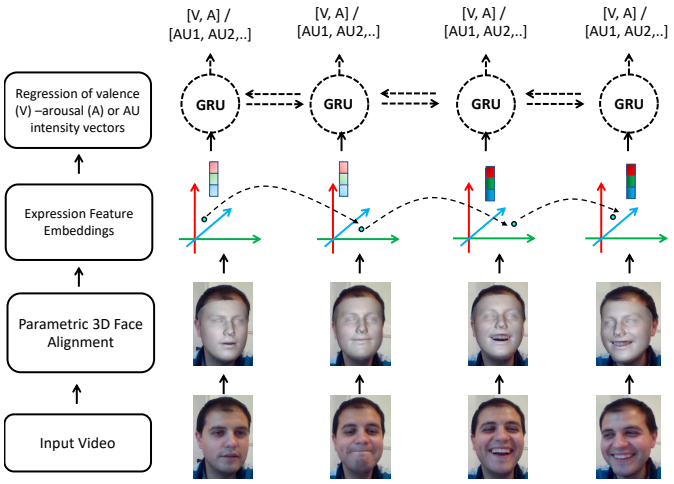

Fig. 2: Modelling the temporal dynamics of 3DMM expression coefficients using bidirectional GRU for video-based dimensional emotion recognition and AU intensity estimation.

Given a 2D face image as input, to extract its expression embedding from its 3D face shape, we consider four different approaches that learn 3DMM parameters: ExpNet [19], 3DDFA-V2 [16], DECA [17], and EMOCA [18]. The criteria for selecting these four models are as follows: EMOCA [18] is the current state-of-the-art in capturing 3D facial expressions, building on the ability of DECA [17] formulation in modelling the detailed facial expressions. DECA develops this ability by adopting a consistency loss to effectively disentangle details specific to a person from wrinkles induced by expressions. Next, we choose 3DDFA-V2 [24] since its global shape reconstruction error is very close to that of DECA [17], [40]. The last model that we evaluate here is ExpNet [19], which directly regresses the 3DMM expression coefficients inferred using the 3DDFA model [41], a predecessor to the 3DDFA-V2 formulation.

Though all these four models output the 3DMM expression embeddings (α_{ex}), their dimensionality varies from model to model: 29 D for ExpNet (same as in the original 3DDFA), 10 D for 3DDFA-V2, and 50 D for both DECA and EMOCA. It is important to note that the fidelity of facial expressions captured by these models does depend not only on the expression embedding dimensionality but also on various other factors such as their corresponding CNN backbone complexity and optimisation procedure followed during their training, etc.

3.2 Expression Embeddings from 2D Face Images

We use end-to-end learning models based on the standard CNN + RNN architectures as 2D appearance baselines. For this purpose, we adopt two strong CNN backbones that are extensively used for end-to-end facial feature learning in several recent works [11], [29], [31], [42].

ResNet-50 [43], particularly, a version of it pre-trained on the VGG-Face database [44], is a commonly used CNN backbone for feature extraction from face images for emotion recognition [29], [30]. When implementing this backbone CNN, we flatten the output feature maps of its last convolutional layer into 2056-dimensional feature vectors, further mapped to 512-dimensional vectors (using an additional fully connected layer) and used as facial embeddings. Considering the relatively small-scale training datasets for AU intensity estimation tasks, we use a ResNet-18 architecture in line with previous works [12], [42], [45].

EmoFAN [11], [45], a recently proposed 2D CNN model, is designed for facial feature extraction using only convolution layers to make the model more efficient in the number of trainable parameters. A pre-trained variant of this backbone on 2D face alignment tasks is found to be very effective for transfer learning [11], [12]. To extract the facial features with better generalisation capacity, we use a variant of this CNN backbone pre-trained on image-based emotion recognition using the AffectNet dataset [46], in addition to the 2D face alignment task. Following prior works [11], [45], we use this backbone to extract 512-dimensional facial embedding vectors.

3.3 Temporal Dynamics of Expression Embeddings

Fig. 2 illustrates the steps that we follow for video-based expression analysis tasks. Modelling the temporal dynamics of frame-wise expression features in a video is critical for dimensional emotion recognition and AU intensity estimation tasks. For this purpose, we use a simple bidirectional 2-layer GRU network with two hidden layers of 128 dimensions. We use the same temporal network for the expression embeddings from the 3D shape and 2D appearance models for a fair comparison. Note that the dimensionality of the input embeddings varies across the different models. On top of the last layer of GRU block output, there is a single fully connected layer to map the per-frame hidden state vector to the final output vector of dimensional emotions or AU intensities. The output is two-dimensional in valence-arousal intensity, whereas it differs in the number of action units in AU intensity estimation models (5-dimensional in the BP4D dataset and 12-dimensional in the DISFA dataset).

3.4 Datasets

Dimensional Emotion Recognition. For video-based valence and arousal estimation, we use two large-scale video datasets: SEWA [21] and the AVEC'19 Cross-cultural Emotion Sub-Challenge (CES) Corpus [7].

SEWA data was collected during computer-based naturalistic dyadic interactions and contains 538 face videos of 398 subjects from 6 different cultures. Each video is annotated with per-frame continuous-valued valence and arousal annotations in the range of -1 to 1 at 50 frames per second (FPS). The numbers of videos used for training, validation, and testing[1] are 431, 53, and 53, respectively, with the duration in the range of 10 s to 30 s.

AVEC'19 CES Corpus is a multimodal in-the-wild affect recognition dataset captured in cross-cultural settings and consists of German, Hungarian, and Chinese subjects. All videos in this dataset are also annotated with continually varying valence and arousal ratings at 50 FPS in the range

1. The details of the train, validation, and test partitions were kindly provided by the database owners.

[-1,1], the same as in the case of SEWA. It provides 64 videos for training, and 32 videos for validation, with a total duration of roughly 160 minutes and 65 minutes, respectively. We report the results on its validation set since the test set labels are not publicly available.

Action Unit Intensity Estimation. We use two video-based AU intensity labelled datasets, DISFA [23] and BP4D [22] with different number of AUs.

DISFA has 27 videos of 27 subjects; each video contains approximately 4844 frames annotated with the intensity values of 12 AUs. As there are no predefined training, validation, and test partitions, a subject-independent 3-fold cross-validation is a commonly used evaluation protocol on this dataset. To compare with the state-of-the-art results on DISFA, following the existing works (e.g. [12], [42], [45]), we also perform the same 3-fold cross-validation; each fold containing 18 videos for training and 9 videos for evaluation.

BP4D contains 487 videos of 41 subjects, containing approximately 140,000 frames annotated with the intensity values of 5 AUs. It was the main corpus of the FERA 2015 challenge [6]. We use the same training (168 videos), validation (160 videos), and test (159 videos) sets that were originally used by the FERA 2015 challenge participants [6].

3.5 Evaluation Metrics

Dimensional Emotion Recognition performance is measured using Lin's Concordance Correlation Coefficient (CCC) [47] that computes the agreement between target emotion labels e^* and their predicted values e^o

$$CCC = \frac{\rho_{e^*e^o}.\sigma_{e^*}.\sigma_{e^o}}{(\mu_{e^*} - \mu_{e^o})^2 + {\sigma_{e^*}}^2 + {\sigma_{e^o}}^2}, \quad (4)$$

where $\rho_{e^*e^o}$ denotes the Pearson's coefficient of correlation between e^* and e^o, and $(\mu_{e^*}, \mu_{e^o}), (\sigma_{e^*}, \sigma_{e^o})$ denote their mean and standard deviation values, respectively.

AU Intensity Estimation is evaluated using two standard metrics: Intra-class Correlation Coefficient (ICC) and Mean Square Error (MSE), computed for each AU individually.

3.6 Training Details

Loss Functions. To train the dimensional emotion recognition models, we use inverse-CCC + MSE loss, following the objective function originally proposed in [48]. Whereas for the AU intensity estimation, we use MSE alone as the loss function, similar to the existing methods [12], [45]. In both cases, the per-frame loss is accumulated over an input image sequence in computing the total loss per mini-batch.

Optimisation. We use the Adam optimiser [49] to train all the models evaluated in this work. Note that in the 2D appearance baselines, CNN backbones and GRU blocks are trained end-to-end, unlike in the case of 3D face models. During training, the dropout values in the GRU and the final FC layers are set to 0.5 and 0.25, respectively. Each mini-batch is composed of 4 sequences, with each sequence containing 100 frames. The initial learning rate value is 1e-4, and it is tuned using a cosine annealing based scheduler with warm restarts enabled [50]. In training all the models presented in this work, L_2 regularisation is applied by setting the weight decay value to 1e-4.

Method	Mean (in mm)	Std. Dev (in mm)	Median (in mm)
ExpNet (3DMM-CNN [51])	2.33	2.05	1.84
3DDFA-V2 [24]	1.57	1.39	1.23
DECA [17]	1.38	1.18	1.09
EMOCA [18]†	1.38	1.18	1.09

TABLE 1: Monocular 3D face reconstruction error (scan-to-mesh distance) values reported in the leader board of NoW [52], [53] evaluation repository (†EMOCA has the same reconstruction performance as DECA)

Model	Valence CCC ↑	Arousal CCC ↑	Avg. CCC ↑
Mitenkova et al. [54]	0.469	0.392	0.415
Toisoul et al. [11]	0.650	0.610	0.630
Kossaifi et al. [48]	0.750	0.520	0.635
APs [45]	0.750	0.640	0.686
ResNet-50+GRU†	0.550	0.552	0.551
EmoFAN+GRU†	0.715	0.568	0.641
ExpNet+GRU	0.638	0.510	0.574
3DDFA+GRU	0.710	0.646	0.678
DECA+GRU	0.755	0.682	0.718
EMOCA+GRU	**0.775**	**0.716**	**0.739**

TABLE 2: Dimensional emotion recognition results on the SEWA test set (†denotes in-house baselines of 2D face appearance baselines).

4 Results and Discussion

4.1 Task-wise Performance Analysis

Dimensional Emotion Recognition on SEWA. Table 2 presents the results of valence and arousal estimation on the SEWA test set in three groups: recent state-of-the-art (SOTA) benchmarks, in-house evaluated 2D CNN baselines using ResNet-50 and EmoFAN backbones, and 3D face models. Note that GRU modules of the same modelling capacity are used on top of the in-house evaluated 2D CNN features as well as 3D face features. We can clearly see that EMOCA features outperform all the remaining models

Model	Valence CCC ↑	Arousal CCC ↑	Avg. CCC ↑
Zhao et al.* [55]	0.579	**0.594**	**0.586**
ResNet-50+GRU†	0.495	0.522	0.508
EmoFAN+GRU†	0.527	0.564	0.545
ExpNet+GRU	0.534	0.505	0.519
3DDFA-V2+GRU	**0.590**	0.544	0.567
DECA+GRU	0.561	0.565	0.563
EMOCA+GRU	0.580	0.568	0.574

TABLE 3: Dimensional emotion recognition results (CCC) on the AVEC'19 validation set. (*denotes visual-only results reported in the AVEC'19 CES Winners [55] and †denotes in-house baselines of 2D face appearance baselines)

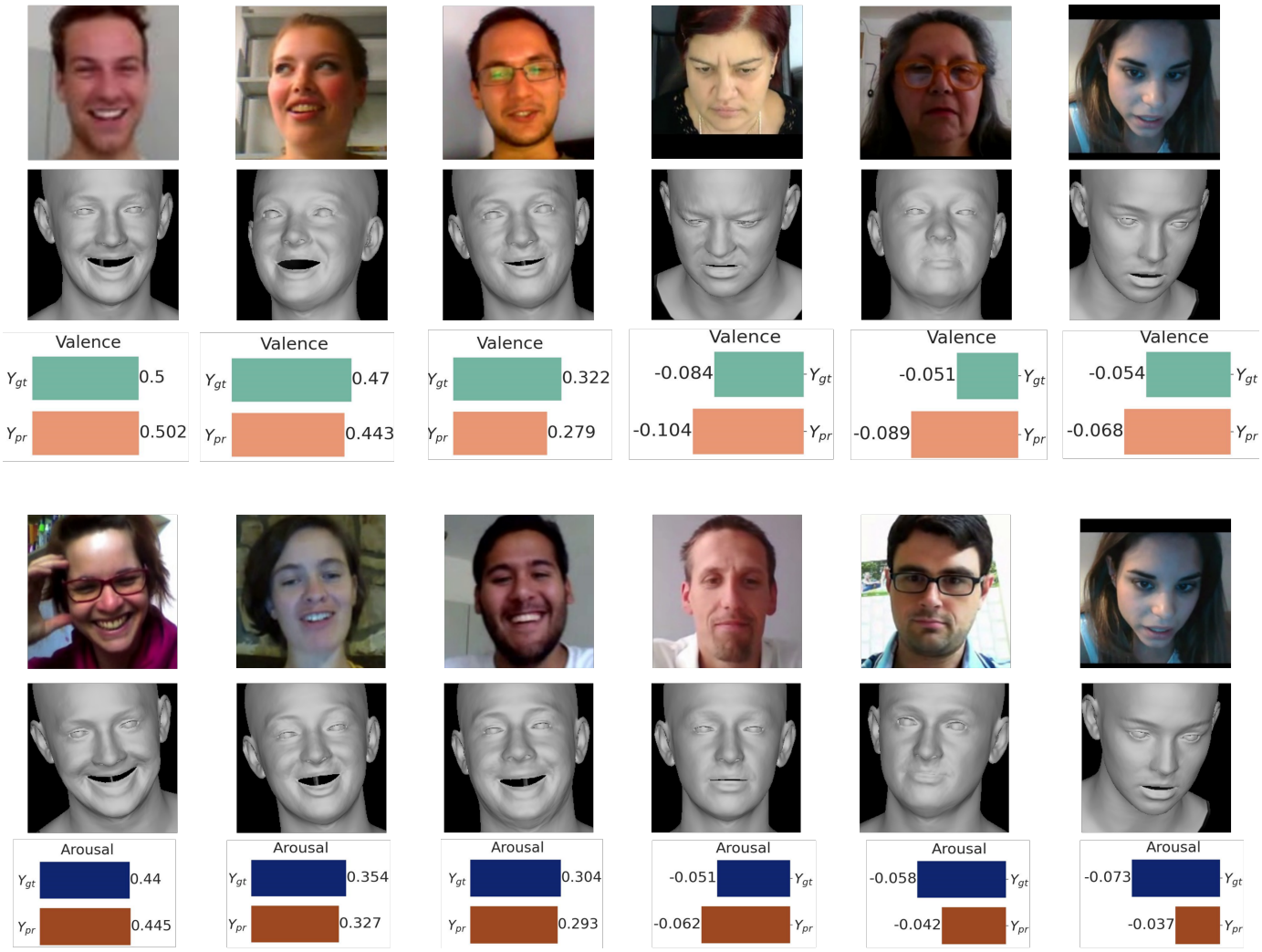

Fig. 3: Dimensional emotion recognition results on AVEC'19 validation examples using EMOCA [18] (Y_{gt} and Y_{pr} denote the ground truth and prediction values, respectively).

listed in Table 2 by considerable margins in terms of both valence and arousal dimensions. Let us consider the best 2D CNN baseline, EmoFAN, as a reference with the CCC values of 0.715 and 0.682 for valence and arousal, respectively. EMOCA expression features outperform EmoFAN by +.060 and +.148 in valence and arousal CCC scores. EMOCA even outperforms all the existing benchmarks on SEWA with improved mean CCC values in the range of +.324 ([54]) to +.053 [45]. Similarly, DECA and 3DDFA-V2 have improved the mean CCC values by +.077 and +.037 compared to the mean CCC of EmoFAN. Further, we observe that 3DDFA-V2 is on par with the SOTA method (APs [45]), which is based on a more complicated stochastic temporal context modelling. The overall performance of 3D face features in Table 2 is in line with their corresponding 3D shape reconstruction errors (see Table 1) reported in the NoW evaluation repository leader board[2].

2. Based on the challenge results provided at https://now.is.tue.mpg.de/nonmetricalevaluation.html. As EMOCA builds on the identity and shape encoders originally learned in DECA, they have the same reconstruction errors on the NoW evaluation repository.

To summarise, the above discussed results on the SEWA demonstrate that the 3D face features are expressive enough to perform superior to the 2D appearance features in recognising time-continuous dimensional emotions. It is worth noting that in contrast to all 2D appearance-based approaches that applied transfer learning on CNN backbones, the training procedure of 3D face models does not rely on any labelled facial expression data. One exception is the EMOCA that used AffectNet [46] pretraining and an additional emotion recognition module. Still, even less accurate 3D face models perform either on par or better than all previous 2D appearance-based methods.

Dimensional Emotion Recognition on AVEC'19 CES. Evaluation results on the AVEC'19 CES database, as shown in Table 3, exhibit similar trends. All four 3D face models perform far above the ResNet-50+GRU baseline, except for the ExpNet model; the rest also outperform a stronger 2D CNN baseline, EmoFAN+GRU. The AVEC'19 CES challenge winners, Zhao et al. [55] achieve a mean CCC score of +.012 above the best performing 3D face model, i.e. EMOCA. In

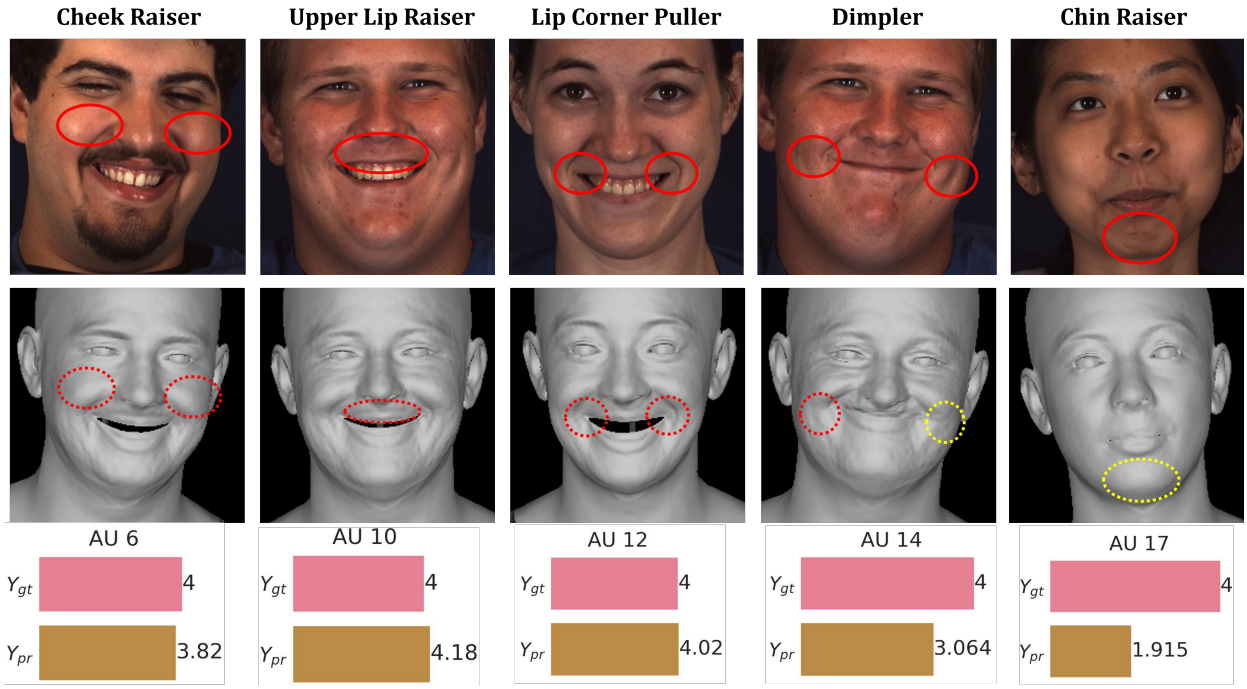

Fig. 4: Action Unit intensity estimation results on BP4D validation examples using EMOCA [18] (Y_{gt} and Y_{pr} denote the ground truth and prediction values, respectively). Facial regions of AUs with low errors in the intensity prediction and with better 3D reconstruction quality are enclosed in red colored ellipses. The regions enclosed in yellow coloured ellipses highlight somewhat poorly reconstructed facial regions of AUs with high errors in the intensity prediction.

valence estimation, unlike in the case of SEWA, the best performing method is 3DDFA-V2.

Figure 3 illustrates the qualitative results of EMOCA in valence-arousal estimation on some of the validation examples from the AVEC'19 corpus. Interestingly, the emotion recognition performance is slightly worse in negative valence and arousal cases compared to their positive counterparts. This could be possible because fewer training examples are available in general for negative quadrants in the existing dimensional emotion datasets [7], [21].

To further validate the efficacy of 3D face expression features in recognising apparent emotions, we also extend our experimental analysis to categorical emotion recognition. Refer to Appendix A for additional experimental results on discrete emotion recognition using the 3D face models and a correspondence analysis between different AUs and discrete emotion categories.

AU Intensity Estimation on BP4D. Table 4 presents the results of existing SOTA benchmarks using our in-house evaluated 2D CNN-based features and 3D face expression features on the test set of BP4D. By comparing the average ICC and MSE scores of the different listed here, we can clearly see that 3D face models have the lowest performance. Most importantly, unlike in valence-arousal estimation, where 3D face models are either on par or better than 2D appearance-based models, and they fall behind in estimating AU intensities. Among the 3D face models, ExpNet has the poorest results in terms of average ICC and MSE scores.

All 3D face models show inferior performance consistently in predicting the intensities of AU 14 and AU 17. Figure 4 qualitatively illustrates the performance of EMOCA on the BP4D test set examples, from which we can notice a clear correspondence between less accurate predictions made for AUs such as dimpler (AU 14) and chin raiser (AU 17) and somewhat poor 3D reconstructions of their corresponding facial regions (enclosed in yellow coloured ellipses in Figure 4). For instance, in the case of AU 17, the details of the chin region are poorly reconstructed in its 3D faces, which explains the poor performance of EMOCA in predicting AU 17 for that example.

AU Intensity Estimation on DISFA. In Table 5, we compare the aggregated results of 3-fold cross-validation of different SOTA methods, 2D CNN baselines, and 3D face models. Similar to the BP4D results, here also we can clearly see that 3D face models have the poorest performance among all the models in terms of their mean ICC and MSE values. Among the 3D face models, ExpNet has the worst performance, whereas EMOCA is the best one. But, unlike in the case of BP4D, EMOCA achieves significantly higher performance (ICC score of +.14 w.r.t. DECA) compared to the remaining 3D face models. Compared to the best performing 2D appearance-based models, APs [45], EMOCA has lower performance by a margin of -.15 ICC score.

Based on the combined results on BP4D and DISFA datasets, we segregate all the AUs into three different groups based on the best ICC values among the 3D face models. We observe that only five AUs (12, 10, 25, 6) listed in group 1 are captured well in the 3D face expressions. Most importantly, the subtler the AUs (e.g. AU 17 – chin raiser)

Metric	Model	6	10	12	14	17	Avg.
ICC ↑	CDL [56]	0.69	0.73	0.83	0.50	0.37	0.62
	ISIR [57]	0.79	0.80	**0.86**	**0.71**	0.44	0.72
	HR [12]	**0.82**	**0.82**	0.80	**0.71**	0.50	0.73
	APs [45]	**0.82**	**0.80**	**0.86**	0.69	**0.51**	**0.74**
	ResNet-18+GRU	0.75	0.71	0.79	0.63	0.45	0.66
	EmoFAN+GRU†	0.78	0.76	0.83	0.62	0.50	0.70
	ExpNet+GRU	0.57	0.56	0.69	0.35	0.38	0.51
	3DDFA-V2+GRU	0.73	0.67	0.87	0.36	0.31	0.59
	DECA+GRU	0.72	0.68	0.83	0.42	0.23	0.58
	EMOCA+GRU	0.73	0.68	0.86	0.34	0.27	0.58
MSE ↓	CDL [56]	-	-	-	-	-	-
	ISIR [57]	0.83	0.80	0.62	1.14	0.84	0.85
	HR [12]	**0.68**	**0.80**	0.79	**0.98**	0.64	0.78
	APs [45]	0.72	0.84	**0.60**	1.13	**0.57**	**0.77**
	ResNet-18+GRU†	0.81	0.90	0.82	1.17	0.82	0.91
	EmoFAN+GRU†	0.79	0.85	0.76	1.19	0.78	0.87
	ExpNet+GRU	1.5	1.56	1.33	1.59	0.88	1.37
	3DDFA-V2+GRU	0.91	1.22	0.61	1.48	0.86	1.01
	DECA+GRU	0.98	1.14	0.8	1.57	1.18	1.13
	EMOCA+GRU	0.82	1.08	0.6	1.75	0.96	1.04

TABLE 4: BP4D test set results (†denotes in-house evaluation).

Metric	Model	1	2	4	5	6	9	12	15	17	20	25	26	Avg.
ICC ↑	G2RL [58]	**0.71**	0.31	**0.82**	0.06	0.48	**0.67**	0.68	0.21	0.47	0.17	**0.95**	**0.75**	0.52
	RE-Net [59]	0.59	**0.63**	0.73	**0.82**	0.49	0.50	0.73	0.29	0.21	0.03	0.90	0.60	0.54
	VGP-AE [60]	0.48	0.47	0.62	0.19	0.50	0.42	0.80	0.19	0.36	0.15	0.84	0.53	0.46
	2DC [61]	0.70	0.55	0.69	0.05	**0.59**	0.57	**0.88**	0.32	0.10	0.08	0.90	0.50	0.50
	HR [12]	0.56	0.52	0.75	0.42	0.51	0.55	0.82	**0.55**	0.37	0.21	0.93	0.62	0.57
	APs [45]	0.35	0.19	0.78	0.73	0.52	0.65	0.81	0.49	**0.61**	**0.28**	0.92	0.67	**0.58**
	ResNet-18+GRU	0.21	0.16	0.71	0.65	0.55	0.59	0.78	0.41	0.54	0.22	0.90	0.64	0.53
	EmoFAN+GRU†	0.23	0.13	0.77	0.70	0.53	0.64	0.82	0.42	0.58	0.25	0.92	0.69	0.56
	ExpNet+GRU	-0.03	-0.07	0.16	-0.02	0.25	0.12	0.38	0.04	0.09	-0.01	0.57	0.31	0.15
	3DDFA-V2+GRU	0.17	0.23	0.19	0.0	0.52	0.32	0.78	0.01	-0.01	0.02	0.78	0.42	0.28
	DECA+GRU	0.09	-0.03	0.33	0.07	0.54	0.30	0.79	0.14	0.14	0.0	0.73	0.33	0.29
	EMOCA+GRU	0.31	0.17	0.76	0.57	0.48	0.52	0.85	0.21	0.16	-0.01	0.84	0.28	0.43
MSE ↓	G2RL [58]	-	-	-	-	-	-	-	-	-	-	-	-	-
	RE-Net [59]	-	-	-	-	-	-	-	-	-	-	-	-	-
	VGP-AE [60]	0.51	**0.32**	1.13	0.08	0.56	0.31	0.47	0.20	0.28	**0.16**	0.49	0.44	0.41
	2DC [61]	**0.32**	0.39	0.53	0.26	0.43	0.30	**0.25**	0.27	0.61	0.18	0.37	0.55	0.37
	HR [12]	0.41	0.37	0.70	0.08	0.44	0.30	0.29	0.14	0.26	**0.16**	0.24	0.39	0.32
	APs [45]	0.68	0.59	**0.40**	**0.03**	0.49	**0.15**	0.26	**0.13**	0.22	0.20	0.35	**0.17**	**0.30**
	ResNet-18+GRU	0.88	0.71	0.54	0.13	0.38	0.26	0.39	0.20	0.28	0.25	0.32	0.41	0.39
	EmoFAN+GRU†	0.85	0.79	0.48	0.06	0.47	0.19	0.34	0.18	0.23	0.21	0.30	0.40	0.37
	ExpNet+GRU	0.93	0.99	1.77	0.09	0.74	0.57	1.09	0.2	0.37	0.18	1.43	0.6	0.75
	3DDFA-V2+GRU	0.53	0.39	1.39	0.07	0.57	0.27	0.5	0.2	0.33	0.14	0.83	0.54	0.48
	DECA+GRU	0.61	0.65	2.23	0.08	0.45	0.40	0.39	0.18	0.37	0.17	0.92	0.5	0.57
	EMOCA+GRU	0.66	0.64	0.63	0.05	0.53	0.29	0.28	0.18	0.36	0.20	0.52	0.56	0.41

TABLE 5: Aggregated 3-fold cross validation results on DISFA dataset († denotes in-house evaluation).

Group 1: $0.6 < ICC < 1.0$	AU 4	Brow Lowerer
	AU 12	Lip Corner Puller
	AU 10	Upper Lip Raiser
	AU 25	Lips Part
	AU 6	Cheek Raiser
Group 2: $0.4 < ICC < 0.6$	AU 5	Upper Lid Raiser
	AU 9	Nose Wrinkler
	AU 14	Dimpler
	AU 26	Jaw Drop
Group 3: $ICC < 0.4$	AU 1	Inner Brow Raiser
	AU 2	Outer Brow Raiser
	AU 15	Lip Corner Depressor
	AU 17	Chin Raiser
	AU 20	Lip stretcher

TABLE 6: Segregation of AUs, from both BP4D and DISFA, according to their best performance with 3D face expression features.

are, the worse their ICC scores are with 3D face models. One possible explanation is that such subtle expression-specific 3D reconstruction errors are likely to get suppressed the typical global reconstruction loss for training the 3D face alignment models. Further, the size of facial regions affecting specific AUs varies widely, and their occurrence in the 3D face models' training data is also likely to be uneven.

Overall, the trends in the AU intensity estimation results clearly show that the 3D faces are still far from capturing the fine-grained facial expressions that are critical to fully understanding expressive facial behaviour. While the dimensional emotion recognition performance of all 3D face models is superior on both the SEWA and AVEC'19 CES datasets, it is interesting to note their inability to recognise a wide range of action units. We performed an AU-wise correspondence analysis between different facial actions and emotion labels to investigate this discrepancy, as discussed below.

4.2 Correspondence Analysis: Dimensional Emotions and Action Units

We investigate the significance of different AUs for recognising dimensional emotions to reconcile the observations above regarding 3D face models' performance on emotion and AU estimation tasks. To this end, we perform a simple linear regression analysis – in which AU intensities are used as input features to predict their corresponding emotion labels. By comparing the coefficient values of different AUs, we interpret the importance of each AU in predicting the target emotion labels. For this purpose, we use a recently released in-the-wild emotion recognition corpus, Aff-wild-2 [8], in which video data is annotated with both valence-arousal values and their corresponding AU occurrences. Although the Aff-Wild-2 dataset seems to be a more suitable candidate for the 3D face expression evaluation, it is composed of highly challenging, in-the-wild recording conditions. As depicted in Figure 7, all the 3D face models evaluated in this work perform poorly on the Aff-wild-2 face images, except for EMOCA, which shows slightly

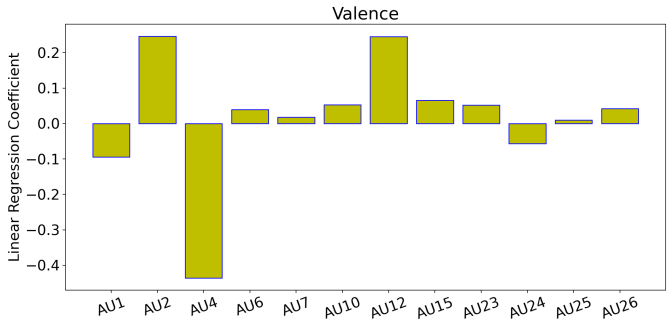

Fig. 5: Coefficients of a linear regression model predicting **valence** from AUs.

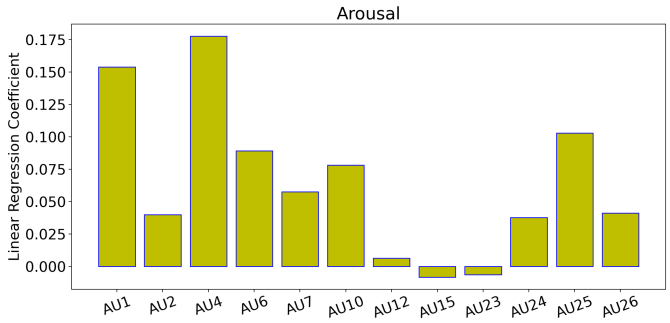

Fig. 6: Coefficients of a linear regression model predicting **arousal** from AUs.

better performance w.r.t. capturing the facial expressions. Rather than pushing the limits of 3D face models to perform well in such in-the-wild conditions, our focus here is to investigate the current status of existing 3D face alignment models where they could attain an acceptable shape-fitting performance.

Figure 5 and Figure 6 illustrate AU wise regression coefficients for valence and arousal respectively. From these illustrations, we infer that the presence of AU 2 or AU 12 or the absence of AU 4 seems to be highly critical for predicting valence. Whereas for arousal prediction, the presence of AU 1 or AU 4 or AU 25 looks important. All four 3D face models perform well (see Group 1 in Table 6) on at least one of the aforementioned AUs critical for valence and arousal prediction. Thus, the superior emotion recognition performance of the 3D face features is clearly explainable based on their relatively high ICC values for AU 4 (Table 4) and AU 25 (Table 5).

4.3 Summary and Discussion

Based on all the above discussed results on dimensional emotion recognition and AU intensity estimation tasks, and the correspondence analysis between valence-arousal and AU intensities, we draw the following conclusions:

- 3D face shapes are expressive enough to regress dimensional representations of facial emotion. They are also good at capturing categorical emotion information (see Appendix. A)

- In AU intensity estimation, 3D faces are far from describing the complete set of facial actions and they fall behind the 2D appearance-based features and most previous methods.

ExpNet, 3DDFA-V2, and DECA do not use any emotion labels, and they are not trained or transferred representations from affect-related tasks. On the other hand, the EMOCA model is based on the DECA model; however, it uses a valence-arousal estimation model pretrained in AffectNet [46]. EMOCA additionally optimises an additional loss component, perceptual emotion consistency loss between the emotion features of RGB input and another valence-arousal estimator on the DECA model's expression and detail coefficients. Our results show even though the use of AffectNet pretraining and emotion consistency, EMOCA demonstrates poor performance in capturing the facial actions corresponding to several AUs, as listed in Table 6.

It is important to note that the size of areas for AUs and their occurrences varies. Constructing 3D face scans captured in conditions eliciting various AUs and making 3D face models more able to reconstruct AU-relevant temporal facial deformations would potentially close the gap. However, the current AU labelled video datasets are limited in terms of the total duration and the number of subjects. To address this significant challenge, leveraging the naturally available supervision cues, such as temporal coherency of facial actions in a video [62], [63], is an alternative approach worth considering to make the 3D face more expressive in a label-efficient manner.

Another important consideration is to increase the affect expressiveness of 3D face models by modelling capacity of 3DMMs. To this end, increasing the dimensionality of expression coefficients is one obvious solution. However, as mentioned in several prior 3D face alignment works (e.g. [24]), higher dimensional expression coefficient vectors may negatively impact the shape reconstruction loss optimisation, hence slowing down the convergence of model training. Thus, to make the 3D face shape models expressive enough to capture the complete set of facial actions, discrete or continuous emotion labels alone as additional supervision signals do not suffice.

Ethical Considerations and Limitations. Automated facial expression analysis, particularly in affective computing, has valuable use cases for society. For example, human-computer interaction, learning analytics, mental health and well-being and teleconferencing are only a few of these beneficial applications for facial expression analysis. However, there exist potential use cases raising ethical questions such as surveillance and military applications.

From the algorithmic fairness point of view, building emotion recognition based on 3D face models has advantages over CNN models that learn emotions directly from RGB images and videos. Most datasets are imbalanced in gender, ethnicity, age and other appearance-relevant traits. Even though algorithmic bias is still a significant and open issue, learning from emotion coefficients of 3D face models discards all additional information that appearance-based CNN models jointly learn and condition on. Together with this benefit, 3D face models require good quality of images for alignment (for instance, see the qualitative performance of all compared 3D face models in Fig. 7), and it may limit their use cases.

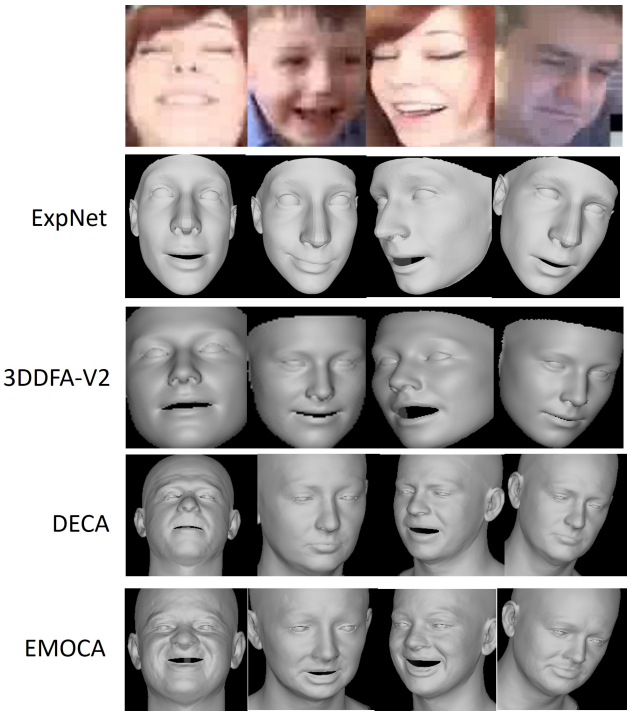

Fig. 7: Shape fitting results of different 3D face alignment models for the representative images sampled from the Aff-wild 2 dataset.

5 Conclusion

We systematically investigated the ability of 3D face models to capture expression-induced shape deformations. By evaluating the 3D face expressions on the standard emotion recognition and AU intensity corpora, we presented a detailed exposition of their current strengths and limitations compared to state-of-the-art models based on 2D face image sequences. Our key findings in this study pointed out that expression features from 3D face models can achieve state-of-the-art results on time-continuous dimensional emotion recognition by outperforming most previous works and strong 2D face appearance baselines. However, the poor performance of 3D face models in AU intensity estimation indicates that 3D face models are far from describing the complete set of facial actions.

Acknowledgments

The work of Mani Kumar Tellamekala was supported by the Engineering and Physical Science Research Council project (2159382) and Unilever U.K. Ltd, and the work of Michel Valstar was supported by the Nottingham Biomedical Research Centre (BRC). This work was also partially funded by the European Union Horizon 2020 research and innovation programme, grant agreement 856879 (Present), and the German Research Foundation DFG, grant agreement AN 559/8-1 (Panorama).

REFERENCES

[1] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald, "Social risk and depression: Evidence from manual and automatic facial expression analysis," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.

[2] C.-H. Wu, Y.-M. Huang, and J.-P. Hwang, "Review of affective computing in education/learning: Trends and challenges," *British Journal of Educational Technology*, vol. 47, no. 6, pp. 1304–1323, 2016.

[3] F. Dornaika and B. Raducanu, "Facial expression recognition for hci applications," in *Encyclopedia of Artificial Intelligence*. IGI Global, 2009, pp. 625–631.

[4] D. R. Faria, M. Vieira, F. C. Faria, and C. Premebida, "Affective facial expressions recognition for human-robot interaction," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 805–810.

[5] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *CVPR Worksh.* IEEE, 2006, pp. 149–149.

[6] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn, "Fera 2015-second facial expression recognition and analysis challenge," in *IEEE FG*, vol. 6. IEEE, 2015, pp. 1–8.

[7] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner *et al.*, "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 3–12.

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

[9] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[10] P. Ekman, "Facial action coding system," 1977.

[11] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence*, vol. 3, no. 1, pp. 42–50, 2021.

[12] I. Ntinou, E. Sanchez, A. Bulat, M. Valstar, and Y. Tzimiropoulos, "A transfer learning approach to heatmap regression for action unit intensity estimation," *IEEE Transactions on Affective Computing*, 2021.

[13] D. Kollias and S. P. Zafeiriou, "Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset," *IEEE Trans. on Affect. Comput.*, 2020.

[14] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '99. USA: ACM Press/Addison-Wesley Publishing Co., 1999, p. 187–194. [Online]. Available: https://doi.org/10.1145/311535.311556

[15] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3d total solution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 78–92, 2017.

[16] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*. Springer, 2020, pp. 152–168.

[17] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," *ACM Trans. Graph.*, vol. 40, no. 4, jul 2021. [Online]. Available: https://doi.org/10.1145/3450626.3459936

[18] R. Danecek, M. J. Black, and T. Bolkart, "EMOCA: Emotion driven monocular face capture and animation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[19] F.-J. Chang, A. Tuan Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "Expnet: Landmark-free, deep, 3d facial expressions," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 2018, pp. 122–129.

[20] R. Belmonte, B. Allaert, P. Tirilly, I. M. Bilasco, C. Djeraba, and N. Sebe, "Impact of facial landmark localization on facial expression recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[21] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. W. Schuller *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[22] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[23] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Trans. on Affect. Comput.*, vol. 4, no. 2, pp. 151–160, 2013.

[24] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 152–168.

[25] B. Egger, W. A. P. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, C. Theobalt, V. Blanz, and T. Vetter, "3d morphable face models—past, present, and future," *ACM Trans. Graph.*, vol. 39, no. 5, jun 2020. [Online]. Available: https://doi.org/10.1145/3395208

[26] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3d total solution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 78–92, 2019.

[27] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "FaceWarehouse: a 3D facial expression database for visual computing," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 3, pp. 413–425, 2014. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2013.249

[28] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM transactions on graphics*, vol. 36, no. 6, pp. 1–17, Nov. 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3130800.3130813

[29] A. Kumar, S. Eslami, D. J. Rezende, M. Garnelo, F. Viola, E. Lockhart, and M. Shanahan, "Consistent generative query networks," *arXiv preprint:1807.02033*, 2018.

[30] D. Kollias, M. A. Nicolaou, I. Kotsia, G. Zhao, and S. Zafeiriou, "Recognition of affect in the wild using deep neural networks," in *CVPR Worksh.* IEEE, 2017, pp. 1972–1979.

[31] D. Kollias and S. Zafeiriou, "Aff-wild2: Extending the aff-wild database for affect recognition," *arXiv preprint:1811.07770*, 2018.

[32] E. Sánchez-Lozano, G. Tzimiropoulos, and M. Valstar, "Joint action unit localisation and intensity estimation through heatmap regression," *arXiv preprint arXiv:1805.03487*, 2018.

[33] D. Fabiano and S. Canavan, "Deformable synthesis model for emotion recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–5.

[34] E. Pei, M. C. Oveneke, Y. Zhao, D. Jiang, and H. Sahli, "Monocular 3d facial expression features for continuous affect recognition," *IEEE Transactions on Multimedia*, 2020.

[35] H. Chen, J. Li, F. Zhang, Y. Li, and H. Wang, "3d model-based continuous emotion recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1836–1845.

[36] M. R. Koujan, L. Alharbawee, G. Giannakakis, N. Pugeault, and A. Roussos, "Real-time facial expression recognition "in the wild" by disentangling 3d expression from identity," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 24–31.

[37] J. Kittler, P. Huber, Z.-H. Feng, G. Hu, and W. Christmas, "3d morphable face models and their applications," in *International Conference on Articulated Motion and Deformable Objects*. Springer, 2016, pp. 185–206.

[38] X. Xiang and T. D. Tran, "Linear disentangled representation learning for facial actions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 12, pp. 3539–3544, 2017.

[39] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[40] Z.-H. Feng, P. Huber, J. Kittler, P. Hancock, X.-J. Wu, Q. Zhao, P. Koppen, and M. Rätsch, "Evaluation of dense 3d reconstruction from 2d face images in the wild," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 780–786.

[41] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146–155.

[42] M. K. Tellamekala, T. Giesbrecht, and M. Valstar, "Modelling stochastic context of audio-visual expressive behaviour with affective processes," *IEEE Transactions on Affective Computing*, 2022.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[44] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.
[45] E. Sanchez, M. K. Tellamekala, M. Valstar, and G. Tzimiropoulos, "Affective processes: stochastic modelling of temporal context for emotion and facial expression recognition," in *CVPR*, 2021.
[46] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.
[47] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.
[48] 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 *CVPR*, June 2020.
[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint:1412.6980*, 2014.
[50] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint:1608.03983*, 2016.
[51] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni, "Regressing robust and discriminative 3d morphable models with a very deep neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5163–5172.
[52] W. Zielonka, T. Bolkart, and J. Thies, "Towards metrical reconstruction of human faces," *arXiv preprint arXiv:2204.06607*, 2022.
[53] S. Sanyal, T. Bolkart, H. Feng, and M. Black, "Learning to regress 3D face shape and expression from an image without 3D supervision," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 7763–7772.
[54] A. Mitenkova, J. Kossaifi, Y. Panagakis, and M. Pantic, "Valence and arousal estimation in-the-wild with tensor methods," in *IEEE FG*. IEEE, 2019, pp. 1–7.
[55] J. Zhao, R. Li, J. Liang, S. Chen, and Q. Jin, "Adversarial domain adaption for multi-cultural dimensional emotion recognition in dyadic interactions," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 37–45.
[56] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *IEEE FG*, vol. 6. IEEE, 2015, pp. 1–6.
[57] J. Nicolle, K. Bailly, and M. Chetouani, "Facial action unit intensity prediction via hard multi-task metric learning for kernel regression," in *IEEE FG*, vol. 6. IEEE, 2015, pp. 1–6.
[58] Y. Fan and Z. Lin, "G2rl: Geometry-guided representation learning for facial action unit intensity estimation," 2020.
[59] H. Yang and L. Yin, "Re-net: A relation embedded deep model for au occurrence and intensity estimation," in *ACCV*, 2020.
[60] S. Eleftheriadis, O. Rudovic, M. P. Deisenroth, and M. Pantic, "Variational gaussian process auto-encoder for ordinal prediction of facial action units," in *ACCV*. Springer, 2016, pp. 154–170.
[61] D. Linh Tran, R. Walecki, S. Eleftheriadis, B. Schuller, M. Pantic *et al.*, "Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding," in *ICCV*, 2017, pp. 3190–3199.
[62] M. K. Tellamekala and M. Valstar, "Temporally coherent visual representations for dimensional affect recognition," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7.
[63] L. Lu, L. Tavabi, and M. Soleymani, "Self-supervised learning for facial action unit recognition through temporal consistency," in *BMVC*, 2020.
[64] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 94–101.
[65] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.
[66] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 94–101.
[67] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014. [Online]. Available: https://www.pnas.org/content/111/15/E1454

APPENDIX A
DISCRETE EMOTION RECOGNITION

In discrete emotion recognition tasks, we evaluate all four 3D face alignment models considered in this work: ExpNet, 3DDFA-v2, DECA, and EMOCA, on the CK+ [64] and CFEE [65] datasets. Both these datasets are acquired in controlled lab settings. Note that here our objective is not to aim for a state-of-the-art performance but to compare the expression representations derived from the 3D face models, as an ablation study.

CK+ [66] contains 327 video clips starting from a neutral state and ending at the apex point of anger (An), contempt (Co), disgust (Di), fear (Fe), happy (Ha), sadness (Sa), and surprise (Su). We use the apex frames in our evaluation.

CFEE [67] contains still images of 230 subjects from diverse ethnic backgrounds with 22 basic and compound emotions categories. We us all samples (1375 images) labelled with basic emotions: anger (An), disgust (Di), fear (Fe), happy (Ha), sadness (Sa), and surprise (Su).

As an ablation study to compare the expression coefficients of 3D face models, we normalise the expression coefficients according to the quantile range of the values and used a simple kNN classifier (k=5) with leave-one-out cross-validation and report the confusion matrices and emotion recognition accuracies.

Discrete Emotion Recognition. Figure 9 and Figure 10 compares class-wise performance of all four 3D face models on CFEE and CK+ datasets respectively. While all the 3D models achieved reasonably good classification accuracy, EMOCA demonstrates the best performance in terms of mean accuracy on both the datasets. Although ExpNet builds on the original 3DDFA formulation that has a higher reconstruction error than 3DDFA-V2, it achieves better accuracy than the 3DDFA-V2. This could be due to the reason that ExpNet is trained particularly for the expression task unlike 3DDFA-V2. Overall, we observe that all the 3D face models, except for EMOCA, perform consistently well on the positive emotions, i.e., the happy and surprise classes, whereas for negative emotions (angry, sad, fear, disgust), they have relatively poor performance in most of the cases. t-SNE distributions of the 3D face expressions features illustrated in Figure 8 illustrate similar trends.

Correspondence Analysis: Discrete Emotions and AUs Figure 11 presents a comparison of class-wise regression coefficients of AUs. As we can see from this illustration, similar to the case of continuous emotions, performance of 3D face models on discrete emotion recognition can be clearly explained by their performance on some specific AUs. For instance, the happiness class for which 3D face models have the best accuracy, is governed by the presence of AU 6 or AU 12, or absence of AU 4. All three 3D face models have good performance on AU 4 (Group 1 in Table 6) and AU 12 (Group 1 in Table 6), explaining their superior performance w.r.t. happiness prediction. On the other hand, fear, one of the classes with lowest accuracy, requires the presence of AU 1 or AU 2, (Group 3 in Table 6) for which the 3D face models have very poor performance.

Fig. 8: t-SNE distributions of the samples with basic emotions in CFEE database using ExpNet, 3DDFA-v2, DECA and EMOCA features

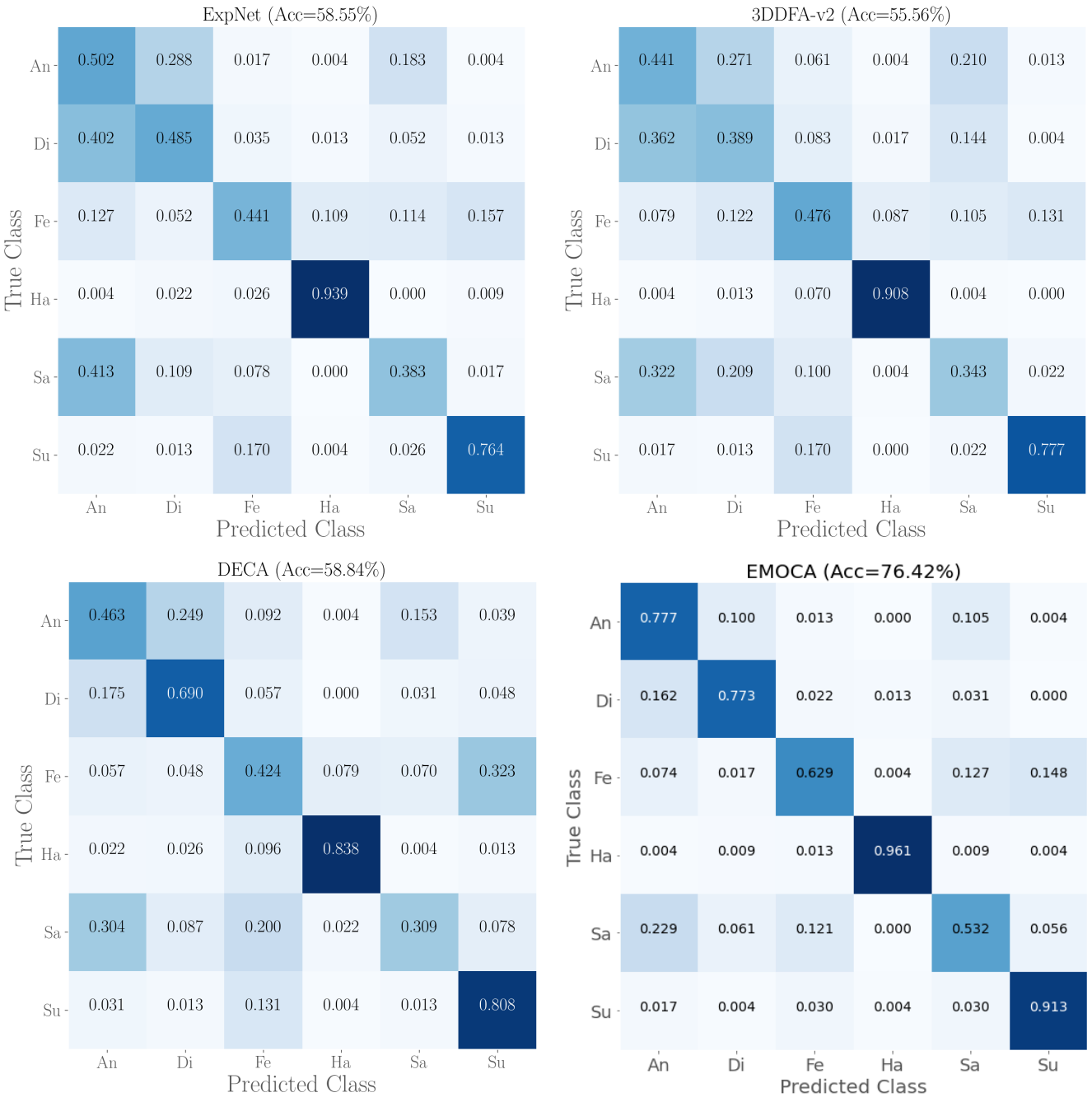

Fig. 9: Discrete Emotion Recognition Results on **CFEE Dataset**

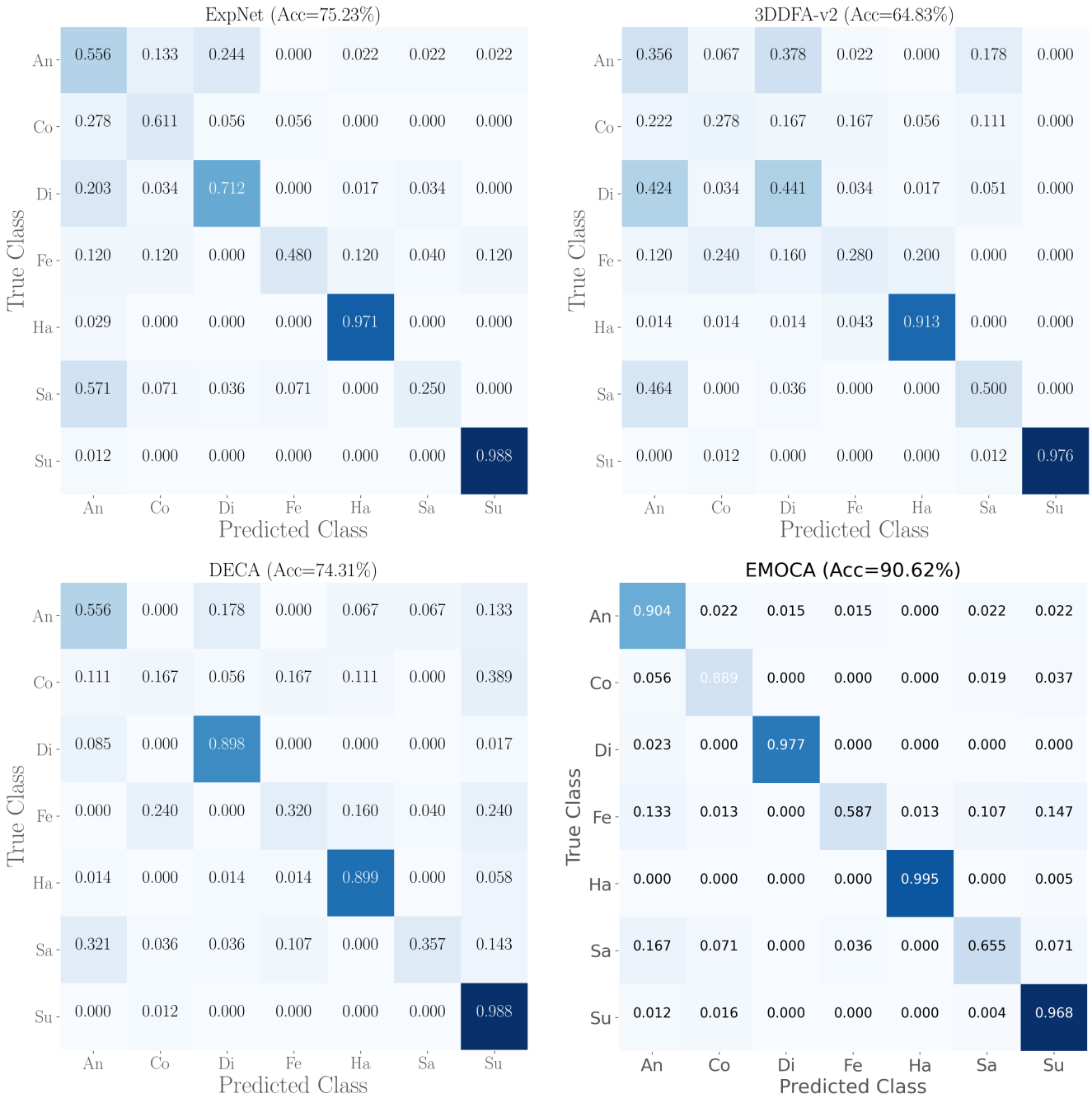

Fig. 10: Discrete Emotion Recognition Results on **CK+ Dataset**.

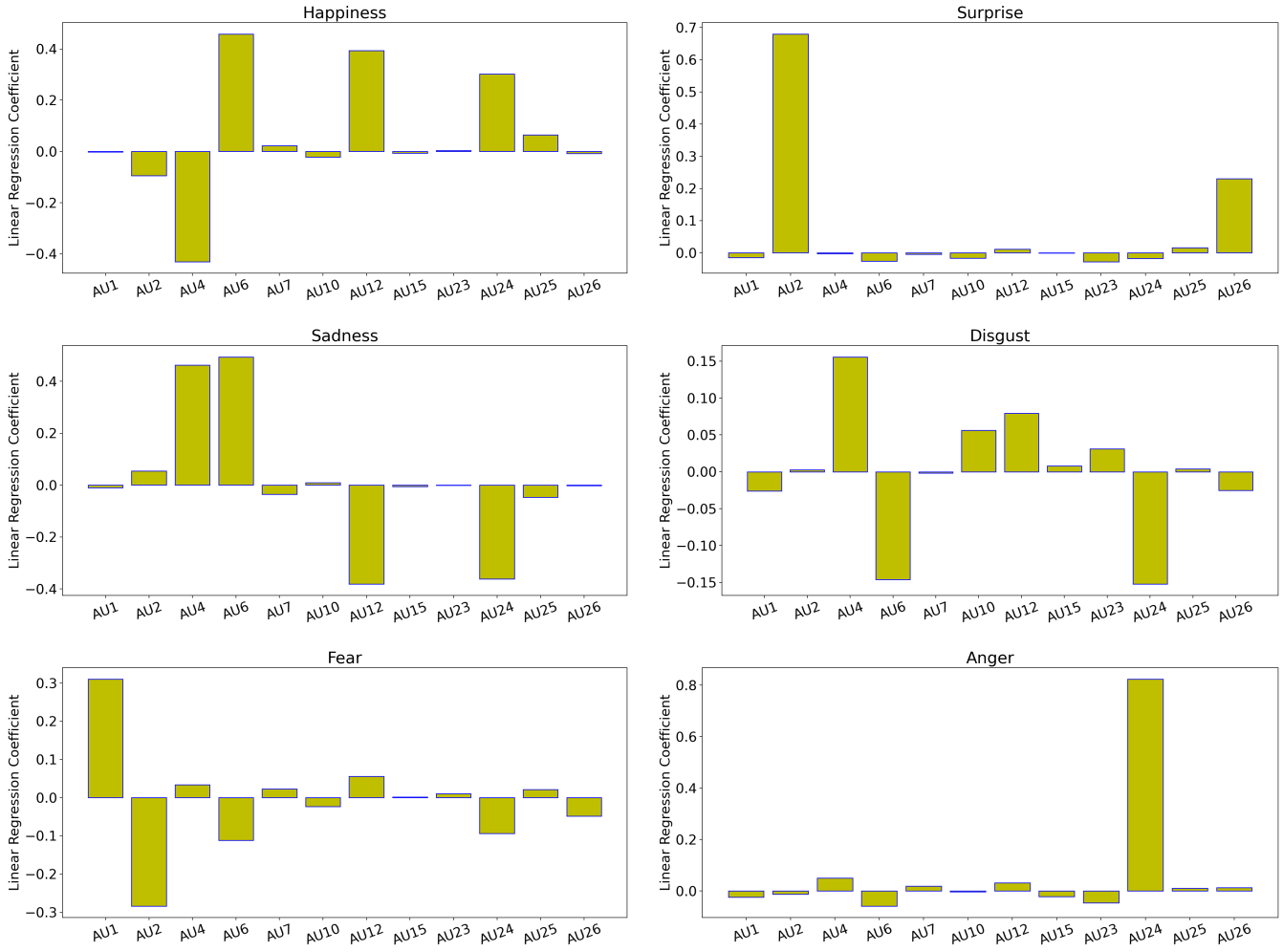

Fig. 11: Coefficients of linear regression model predicting **discrete emotions** from AUs.