Examining Subject-Dependent and Subject-Independent Human Affect Inference from Limited Video Data

Ravikiran Parameshwara, Ibrahim Radwan, Ramanathan Subramanian and Roland Goecke
Human-Centred Technology Research Centre, University of Canberra, Australia
{Ravikiran.Parameshwara, Ibrahim.Radwan, Ram.Subramanian}@canberra.edu.au
Roland.Goecke@ieee.org

Abstract— **Continuous human affect estimation from video data entails modelling the dynamic emotional state from a sequence of facial images. Though multiple affective video databases exist, they are limited in terms of data and dynamic annotations, as assigning continuous affective labels to video data is subjective, onerous and tedious. While studies have established the existence of *signature* facial expressions corresponding to the basic categorical emotions, individual differences in emoting facial expressions nevertheless exist; factoring out these idiosyncrasies is critical for effective emotion inference. This work explores continuous human affect recognition using AFEW-VA, an 'in-the-wild' video dataset with limited data, employing *subject-independent* (SI) and *subject-dependent* (SD) settings. The SI setting involves the use of training and test sets with mutually exclusive subjects, while training and test samples corresponding to the same subject can occur in the SD setting. A novel, dynamically-weighted loss function is employed with a Convolutional Neural Network (CNN)-Long Short-Term Memory (LSTM) architecture to optimise dynamic affect prediction. Superior prediction is achieved in the SD setting, as compared to the SI counterpart.**

I. Introduction and Background

Automatic emotion inference plays a critical role in building intelligent human-machine interfaces that can understand and respond to human emotions. Emotions are integral to perception, rational decision making, and other cognitive functions [25]. Previously, there have been tremendous efforts to develop methods that can accurately recognise and analyse human affect [12]. Research on emotion inference focuses on either *categorising* human emotions into the universal emotional classes [6], namely, happiness, fear, surprise, sadness, disgust, contempt, and anger, or mapping them onto a continuous *dimensional* plane [28], for example, spanned by the *valence* and *arousal* dimensions. Valence is defined as the degree of pleasantness or unpleasantness elicited by a stimulus, while arousal describes the extent of calmness or excitation (physiological activity) evoked by it. The continuous model is more representative of emotions as compared to the categorical counterpart, as it can lead to an accurate assessment of the natural affective state, as evoked emotions are often mixed, complex, subtle and ambiguous in real-world scenarios [8].

Since emotional expressions evolve dynamically [6], it is essential to model short and long-range dependencies among emotional expression features observed over a given time interval. A continuous emotional space not only describes complex emotional states, but naturally enables the representation by a temporal model [20]. Multiple studies have employed deep neural networks to learn spatio-temporal dependencies for emotion inference [13], [16], [30], [32]. Affect recognition systems typically use affective data captured in controlled settings [1], [23], but recent studies have focused on recognising 'in-the-wild' emotional expressions captured under naturalistic settings [17], [31].

Although multiple affective video databases exist [15], [17], [27], they are typically limited by the amounts of annotated data, unlike affective image databases that contain millions of samples [22]. The scarcity of large corpora of affective annotated video data can be attributed to (a) the difficulty in capturing emotional data under naturalistic conditions and, (b) the difficulty in assigning static and dynamic emotion labels to large amounts of data [8]. In addition, limited video data present a challenge as emotional patterns are to be learnt spatio-temporally, unlike image data where spatial emotion representations need to be learnt.

Early studies on basic emotions state the existence of a core facial configuration reflecting the emotional state of a person [5]. In contrast, other scientific frameworks posit that expressions of the same emotion vary substantially across individuals and situations [3]. For example, the typical expression of anger (eyebrows furrowed, eyes wide, lips tightened) might sometimes be accompanied with additional facial movements such as a widened mouth, while in other instances, a facial movement might be missing with respect to the prototype. Such variations are considered to be a meaningful part of an emotional expression, because facial movements are functionally tied to other factors such as external context and the person's internal affective state. Hence, while inferring affect computationally, it becomes essential to consider models, which are both subject-specific and subject-agnostic. Specifically, limited data is a serious impediment in learning emotion-specific facial expressions and it is, therefore, likely that machine learning and deep learning algorithms learn identity-specific characteristics for decoding observed expressions of emotions. Consequently, how training and test data are divided plays a vital role [9], [29] in determining recognition performance. A comparison of subject-specific vs subject-agnostic settings is, therefore, critical in lean data settings. We employ both *subject-dependent* (SD) and *subject-independent* (SI) settings to infer valence and arousal scores on the AFEW-VA [17], an in-the-

979-8-3503-4544-5/23/$31.00 ©2023 IEEE

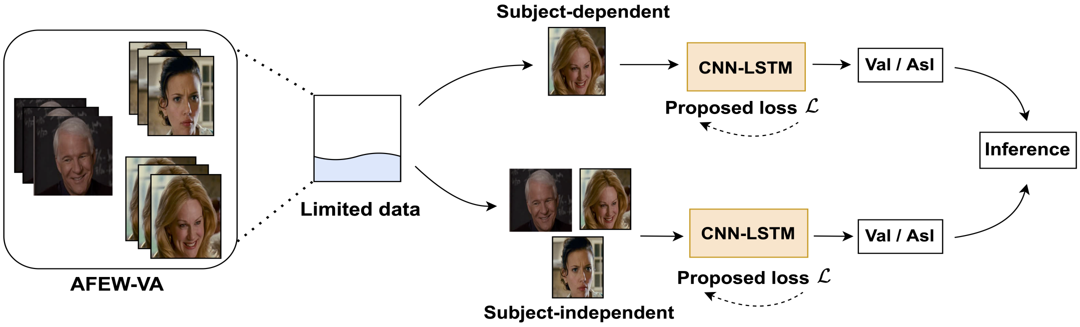

Fig. 1: **Approach overview** depicting continuous/dynamic valence (Val) and arousal (Asl) score prediction with limited data in the AFEW-VA [17] dataset. The proposed network and loss function are evaluated in SD and SI settings.

wild video dataset with limited data. The SI setting involves the use of training and test sets with mutually exclusive subjects, while training and test samples corresponding to the same subject can occur in the SD setting (see Fig. 1). Specifically, we make the following research contributions:

1) We examine the SI and SD settings for valence and arousal inference on the sample-limited AFEW-VA dataset. Given (a) the small size of AFEW-VA dataset, and (b) the inverse-exponential (e^{-x}) distribution observed for the number of samples (video snippets) available per subject (see Fig. 2 (left)), we note vastly different emotion inference performance in the SI and SD settings.
2) While both the SI and SD settings involve mutually exclusive training and test sets, these sets also involve *mutually exclusive subjects* in the SI setting as mentioned above. All performance metrics considered here substantially improve in the SD setting as compared to the SI setting. These results reveal that learning individual encoding is critical for accurate arousal and valence recognition on AFEW-VA.
3) We also propose a novel dynamically-weighted loss function $\mathcal{L}$ to simultaneously improve the correlation as well as minimise the error between the target values and predicted *valence* and *arousal* values via the Convolutional Neural Network (CNN)-Long Short-Term Memory (LSTM) network depicted in Fig. 3.
4) To the best of our knowledge, this study serves as an upper limit benchmark for affect inference in the SD setting on the AFEW-VA dataset.

II. Experiments

We now describe the AFEW-VA dataset, the pre-processing step of face extraction, our CNN-LSTM architecture, evaluation metrics, and the proposed loss function.

A. Dataset

To examine emotion inference using limited data, we use a publicly available video dataset, AFEW-VA [17], which is a subset of AFEW [4]. While AFEW is an affective database with categorical annotations of the six universal emotions plus a neutral class, AFEW-VA consists of a subset of 600 video sequences with continuous annotations of valence and arousal values in the range of $[-10, 10]$ for each frame in the video sequence, whose length varies between 10–145 frames. The videos in the database are collected from movies, which are closer to real-world scenarios than controlled lab settings. The 240 *subjects* in AFEW-VA denote the video actors, who mimic real-world human behaviour [4].

B. Methods

*1) **Pre-processing**:* As an initial step, we extract faces from each frame in the AFEW-VA videos. Given an input video, we employ Multitask Cascaded Convolutional Neural Networks (MTCNN) [35], which is a unified framework for both face detection and face alignment. Sometimes, face detection algorithms are prone to fail while detecting faces in-the-wild. In such a case, we employ the Contrast Limited Adaptive Histogram Equalization (CLAHE) [26] technique to enhance image contrast. The output of CLAHE is passed to MTCNN for face detection. If the face is still not detected, the bounding box of the neighbouring (preceding or succeeding) frame is positioned on the current frame, given that the face location differences in neighbouring frames are negligible. Rather than discarding a frame when the face is not detected via MTCNN, we thus ensure the inclusion of almost all frames in the AFEW-VA, which is relatively small to begin with.

In each video, we consider a sequence (snippet) of eight consecutive frames [2] with a stride of 1 as an input sample. The total number of derived samples from the dataset is $25,759$, with the input dimensionality of each sample being $8 \times 3 \times 128 \times 128$, *i.e.*, each input sample (or video snippet) has eight frames of size $3 \times 128 \times 128$. Input samples and affect labels are normalised to the $[0, 1]$ and $[-1, 1]$ range, respectively, before feeding them to a CNN-LSTM network. Fig. 2 (left) shows the distribution of the number of video snippets per subject.

*2) **Architecture**:* The architecture chosen for this study is illustrated in Fig. 3. We use a CNN-LSTM network, in which spatio-temporal patterns are learned using a 2D-CNN network for each frame followed by LSTM layers. For the CNN architecture to learn spatial patterns, we chose ResNet-18 and ResNet-50 [10] architectures, however, we report our experimental results for the ResNet-18 architecture, as obtained results were fairly similar in both cases. The final

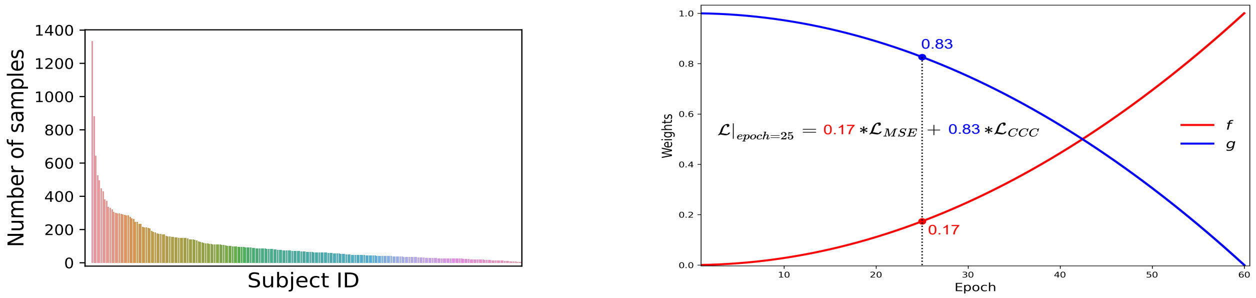

Fig. 2: **(Left)** Distribution of the number of input samples per subject in the AFEW-VA dataset. **(Right)** Illustration of the dynamic weight functions f and g used in Eq. 4 with $k = 2$, $\alpha = 1$, and $n = 60$.

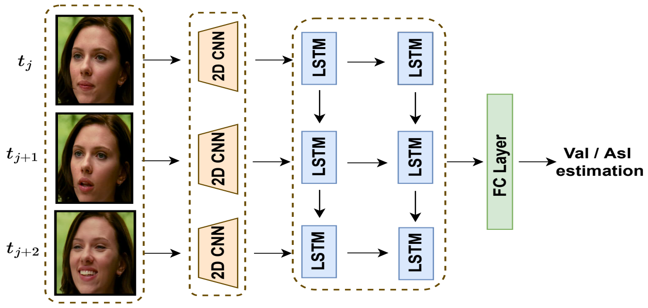

Fig. 3: Overview of the proposed CNN-LSTM network architecture for a 3-frame video snippet. t_j, t_{j+1}, and t_{j+2} denote the time steps corresponding to the three frames.

classification layer of ResNet-18 is replaced with a linear layer with 300 neurons (this number was empirically found to be optimal). The outputs of the respective CNN networks are fed as input to an LSTM layer, followed by another LSTM layer, both with 256 units. This is followed by a linear layer with 128 neurons and a final regression layer for estimating valence or arousal scores in $[-1, 1]$.

*3) **Metrics and Loss Function:*** Similar to past studies examining dimensional emotion estimation [14], [17], [33], the performance metrics used here are Root Mean Square Error (RMSE), Pearson Correlation Coefficient (PCC) and Concordance Correlation coefficient (CCC). If θ and $\hat{\theta}$ denote the ground truth and predicted labels, respectively, the metrics are defined as:

$$RMSE(\theta, \hat{\theta}) = \sqrt{\mathbb{E}[(\theta - \hat{\theta})^2]} \tag{1}$$

$$PCC(\theta, \hat{\theta}) = \frac{\mathbb{E}[(\theta - \mu_\theta)(\hat{\theta} - \mu_{\hat{\theta}})]}{\sigma_\theta \sigma_{\hat{\theta}}} \tag{2}$$

$$CCC(\theta, \hat{\theta}) = \frac{2\sigma_\theta \sigma_{\hat{\theta}} PCC(\theta, \hat{\theta})}{\sigma_\theta^2 + \sigma_{\hat{\theta}}^2 + (\mu_\theta - \mu_{\hat{\theta}})^2} \tag{3}$$

where μ_θ and σ_θ correspond to the mean and the standard deviation of θ, respectively, and $\mathbb{E}$ denotes the expected value.

In dimensional affect inference, the aim is to minimise RMSE, while simultaneously maximising PCC and CCC. The most common approach employed for regression model optimisation is to use individual loss functions, namely, Mean Squared Error (MSE) or inverse-CCC [11], [14], [18].

TABLE I: RMSE, PCC and CCC values of estimated valence using various loss functions in the subject-dependent setting.

Loss			$RMSE\downarrow$	$PCC\uparrow$	$CCC\uparrow$
MSE	PCC	CCC			
✓			0.17 ± 0.06	0.73 ± 0.26	0.68 ± 0.32
	✓		0.74 ± 0.08	0.52 ± 0.08	0.31 ± 0.05
		✓	0.25 ± 0.02	0.69 ± 0.11	0.69 ± 0.10
✓	✓	✓	0.22 ± 0.08	0.73 ± 0.08	0.73 ± 0.08
$\mathcal{L}$ (proposed)			**0.13± 0.01**	**0.89** ± **0.02**	**0.89** ± **0.02**

Other studies also use a combination of losses in addition to using the losses individually [16], [33]. For example, the authors of [16] use a weighted sum of the MSE, CCC and PCC losses, where weights are shake-shake regularisation coefficients [7] sampled randomly and uniformly in the range $[0, 1]$. Differently, we propose a dynamically weighted loss function $\mathcal{L}$, defined as:

$$\mathcal{L} = f * \mathcal{L}_{MSE} + g * \mathcal{L}_{CCC} \tag{4}$$

where $\mathcal{L}_{MSE}$ is the MSE loss, $\mathcal{L}_{CCC}$ = 1 - CCC, and f and g are dynamic weight functions given by $f = \alpha \left(\frac{i}{n}\right)^k$ and $g = 1 - \left(\frac{i}{n}\right)^k$, where i denotes the i^{th} epoch in the training phase of a total n epochs, and $\alpha \in \mathbb{R}$ and $k \in \mathbb{Z}^+$ are hyper-parameters controlling the normalisation and non-linearity, respectively. Fig. 2 (right) illustrates the weight functions f and g.

For estimating affect, generally, $\mathcal{L}_{CCC}$ is used to maximise the correlation between the ground-truth and the predicted values. When a combination of loss functions with *static* co-efficients is used, the model tries to simultaneously optimise the metrics, which may result in a sub-optimal model. Our proposed loss function employs dynamic weights to ensure that the network learns to maximise the correlation initially and, then, minimise the error. Empirically, the proposed loss function results in improved model performance as shown in Table I.

C. *Implementation*

The model is implemented using the open-source software library PyTorch [24] and is trained on an NVIDIA A100 GPU with 40GB memory. The Adam optimiser is used with a decrease of the learning rate by a factor of 10 for every

TABLE II: RMSE, PCC and CCC values of estimated valence and arousal for the SI and SD settings.

Mode	Valence			Arousal		
	$RMSE \downarrow$	$PCC \uparrow$	$CCC \uparrow$	$RMSE \downarrow$	$PCC \uparrow$	$CCC \uparrow$
Subject-independent	0.35 ± 0.02	0.12 ± 0.11	0.10 ± 0.10	0.31 ± 0.02	0.29 ± 0.12	0.26 ± 0.12
Subject-dependent	**0.13 ± 0.01**	**0.89 ± 0.02**	**0.89 ± 0.02**	**0.12 ± 0.00**	**0.93 ± 0.00**	**0.93 ± 0.00**

15 epochs, with the initial learning rate set to 10^{-3}. The models are trained for 60 epochs with a batch size of 128 and a dropout rate of 0.5. In the proposed loss function, fine-tuning is performed for hyper-parameters $k \in [1, 2, 3]$, and $\alpha \in [1, 2, 20]$. The results reported are the $\mu \pm \sigma$ values obtained via five-fold cross-validation.

III. Results and Discussion

Table I shows the RMSE, PCC, and CCC values obtained using the individual loss functions, a combination of the loss functions, and the proposed dynamically-weighted loss function for valence estimation within the SD setting. As mentioned earlier, a typical objective of regression models is to simultaneously minimise RMSE, while maximising PCC and CCC. When using the CCC loss function alone (row 3), the obtained RMSE is worse as compared to its MSE loss counterpart (row 1). Conversely, while using the MSE loss function alone, the model performs better in terms of minimising RMSE, as compared to its CCC loss counterpart. When the PCC loss alone (row 2) is used, the achieved PCC and CCC values are lower than the CCC loss counterpart. This is because, as can be seen in Eq. 3, CCC also incorporates the PCC value, but penalizes correlated signals with different means [19]. That is, if the predicted feature has a trend similar to the target feature, but the predicted value is far from the target value, implying a high error, a low CCC is obtained, although the PCC is high. It is also observed that the RMSE value is optimised better when the CCC loss is used, as compared to the PCC loss.

In comparison to the individual loss functions, when a weighted sum of the three loss functions is employed (where weights are the shake-shake regularisation coefficients [7], as in [33]), PCC and CCC values are higher than the individual loss counterparts. However, a trade-off is observed in terms of RMSE value as compared to the MSE loss counterpart. The proposed loss function performs the best as the RMSE value is the lowest, while the PCC and CCC values are the highest as compared to the other loss functions. To account for optimising all three metrics simultaneously, the proposed loss function is designed to learn the correlations initially and, later, to minimise the mean squared error, in a continuous fashion where the (non-)linearity factor is controlled by the hyper-parameter k.

In this study, we perform both subject-dependent and subject-independent experiments for valence and arousal inference using AFEW-VA. While prior studies only performed subject-independent experiments [13], [14], [16], [21], [33], [36], whether affect inference is required in an SI or SD setting may depend on the use-case. *E.g.*, if the affect inference system entails inferring affect from the end user, the model should be optimised for each user. Conversely, if the affect recognition is to be achieved independently of the end-user, the model should be optimised for the SI setting. We employ the proposed loss function in our experiments as it results in improved RSME, PCC and CCC values, as compared to the other loss functions.

The values in Table II are obtained using the proposed loss functions for the subject-independent and subject-dependent settings. As seen in the table, all performance metrics are improved in the SD setting as compared to the SI setting, as the SD setting results in the lowest RMSE and highest PCC and CCC values for both valence and arousal. In contrast, the model is unable to learn generalised features to discern diverse valence or arousal values in the subject-independent framework. To obtain further insights, we use t-SNE [34] to visualise the features learnt to estimate valence and arousal in the two settings (see Fig. 4). As can be seen, the learned features for valence and arousal prediction in the SD setting are better separated as compared to the SI setting, where features corresponding to high/low valence/arousal values overlap considerably.

Overall, the results in the SD setting indicate the upper limit for arousal and valence estimation on the AFEW-VA dataset. Moreover, the results in Table I demonstrate that the proposed loss function results in superior prediction performance. Employing this loss function on the AFEW-VA dataset, we observe considerably more precise valence and arousal estimation in the SD setting as compared to the SI setting. Cumulatively, the results reveal that for the small AFEW-VA dataset with a highly imbalanced distribution of input samples per subject, identity-specific characteristics substantially impact emotional inference. Conversely, emotion-specific representations cannot be efficiently learned across subjects as typified by the poor valence/arousal estimation in the SI setting.

IV. Conclusions and Future Work

In this study, we have examined the influence of limited video data and of an imbalanced distribution of samples per subject on continuous human affect (valence, arousal) inference using the AFEW-VA dataset. While some studies in psychology state the existence of unique emotional expressions for the basic emotions, on the contrary, others hypothesise that emotional expressions of the same emotion vary substantially across individuals (and often for the same individual) due to factors such as context, social environment, *etc*. Hence, to infer affect computationally, it is essential to examine the affect inference using

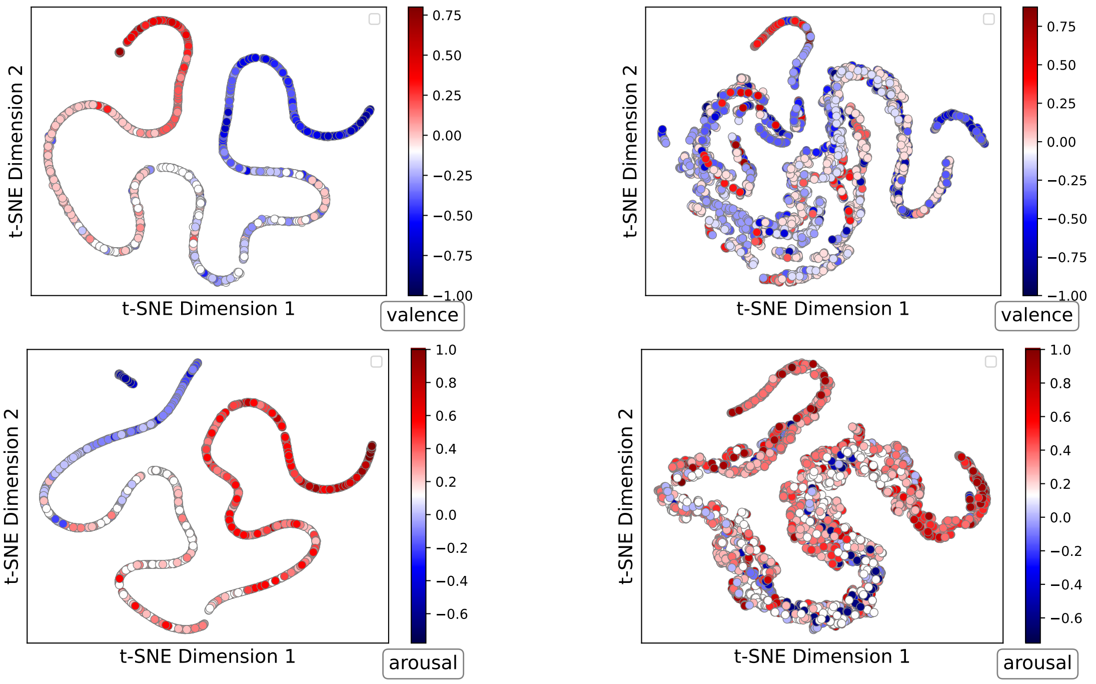

Fig. 4: Visualisations of the feature distribution generated by t-SNE for valence (top row) and arousal (bottom row) using subject-dependent (left) and subject-independent (right) frameworks.

subject-dependent and subject-independent settings. A novel dynamically-weighted loss function is proposed, and is found to enhance correlation as well as reduce error between the target and predicted values. Empirically, we observe that this loss function results in an improved performance than competing loss functions. Furthermore, superior performance in terms of RMSE, PCC, and CCC metrics is observed in the subject-dependent framework as compared to the subject-independent counterpart. The results indicate that the features of valence and arousal learnt by the model are not generalisable across subjects. Visualisations convey that the features of the subject-independent framework are not as discriminative as the subject-dependent setting.

In the future, we plan to explore the influence of individual-specific factors on affect inference using other affective video datasets, such as RECOLA [27] and AffWild2 [15]. We also plan to examine the effect of fusing multiple modalities, such as audio and context, for affect inference using subject-dependent and subject-independent settings.

V. Acknowledgments

This research was partially supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP190101294).

References

[1] M. K. Abadi, S. M. Kia, R. Subramanian, P. Avesani, and N. Sebe. User-centric affective video tagging from meg and peripheral physiological responses. In *Affective Computing and Intelligent Interaction*, pages 582–587, 2013. 1

[2] 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 *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,*, pages 172–181. INSTICC, SciTePress, 2021. 2

[3] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019. PMID: 31313636. 1

[4] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, 2012. 2

[5] P. Ekman and D. Cordaro. What is meant by calling emotions basic. *Emotion Review*, 3(4):364–370, 2011. 1

[6] P. Ekman, E. R. Sorenson, and W. V. Friesen. Pan-cultural elements in facial displays of emotion. *Science*, 164(3875):86–88, 1969. 1

[7] X. Gastaldi. Shake-shake regularization. *arXiv preprint arXiv:1705.07485*, 2017. 3, 4

[8] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136, 2013. Affect Analysis In Continuous Input. 1

[9] N. Hajarolasvadi, E. Bashirov, and H. Demirel. Video-based person-dependent and person-independent facial emotion recognition. *Signal, Image and Video Processing*, 15(5):1049–1056, 2021. 1

[10] 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*, pages 770–778, 2016. 2

[11] M. Hu, Q. Chu, X. Wang, L. He, and F. Ren. A two-stage spatiotemporal attention convolution network for continuous dimensional emotion recognition from facial video. *IEEE Signal Processing Letters*, 28:698–702, 2021. 3

[12] B. C. Ko. A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401, 2018. 1

[13] D. Kollias, S. Cheng, E. Ververas, I. Kotsia, and S. Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, 128(5):1455–1484, 2020. 1, 4

[14] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond.

Int. J. Comput. Vision, 127(6–7):907–929, jun 2019. 3, 4

[15] D. Kollias and S. Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018. 1, 5

[16] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6060–6069, 2020. 1, 3, 4

[17] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017. 1, 2, 3

[18] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1022–1040, 2019. 3

[19] I. Lawrence and K. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989. 4

[20] A. Metallinou and S. Narayanan. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013. 1

[21] A. Mitenkova, J. Kossaifi, Y. Panagakis, and M. Pantic. Valence and arousal estimation in-the-wild with tensor methods. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019. 4

[22] 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*, 10(1):18–31, 2017. 1

[23] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 317–321. IEEE, 2005. 1

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 3

[25] R. W. Picard. *Affective computing*. MIT press, 2000. 1

[26] A. M. Reza. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38(1):35–44, 2004. 2

[27] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013. 1, 5

[28] J. A. Russell. Evidence of convergent validity on the dimensions of affect. *Journal of personality and social psychology*, 36(10):1152, 1978. 1

[29] S. Scheurer, S. Tedesco, B. O'Flynn, and K. N. Brown. Comparing person-specific and independent models on subject-dependent and independent human activity recognition performance. *Sensors*, 20(13):3647, 2020. 1

[30] A. Shukla, S. S. Gullapuram, H. Katti, M. Kankanhalli, S. Winkler, and R. Subramanian. Recognition of advertisement emotions with application to computational advertising. *IEEE Transactions on Affective Computing*, 13(2):781–792, 2022. 1

[31] A. Shukla, S. S. Gullapuram, H. Katti, K. Yadati, M. Kankanhalli, and R. Subramanian. Evaluating content-centric vs. user-centric ad affect recognition. In *International Conference on Multimodal Interaction*, page 402–410, New York, NY, USA, 2017. Association for Computing Machinery. 1

[32] 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)*, pages 1–7. IEEE, 2019. 1

[33] 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*, 3(1):42–50, 2021. 3, 4

[34] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 4

[35] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 2

[36] R. Zhao, T. Liu, Z. Huang, D. P. Lun, and K.-M. Lam. Spatial-temporal graphs plus transformers for geometry-guided facial expression recognition. *IEEE Transactions on Affective Computing*, 2022. 4