To Improve Is to Change: Towards Improving Mood Prediction by Learning Changes in Emotion

SOUJANYA NARAYANA, University of Canberra, Australia
RAMANATHAN SUBRAMANIAN, University of Canberra, Australia
IBRAHIM RADWAN, University of Canberra, Australia
ROLAND GOECKE, University of Canberra, Australia

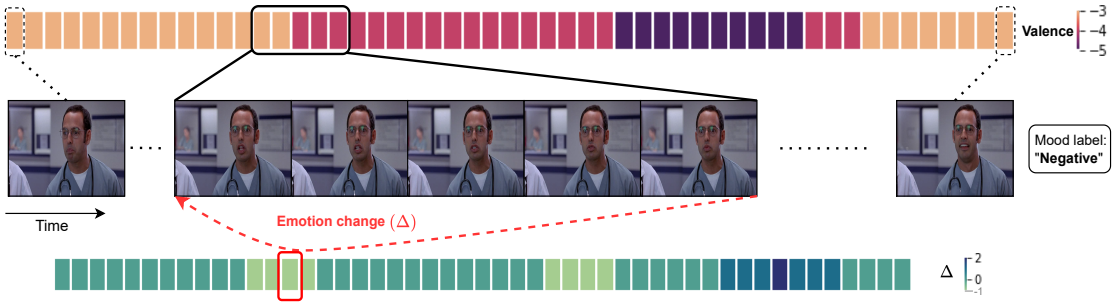

Fig. 1. **Problem Illustration:** Figure depicts emotion changes in an input video sample having a negative mood label. The top colour bar denotes per-frame valence values for the video, while the bottom colour bar depicts emotional valence change (Δ) labels over a window of five frames (best viewed in colour).

Although the terms *mood* and *emotion* are closely related and often used interchangeably, they are distinguished based on their duration, intensity and attribution. To date, hardly any computational models have (a) examined mood recognition, and (b) modelled the interplay between mood and emotional state in their analysis. In this paper, as a first step towards mood prediction, we propose a framework that utilises both dominant emotion (or *mood*) labels, and emotional change labels on the AFEW-VA database. Experiments evaluating unimodal (trained only using mood labels) and multimodal (trained with both mood and emotion change labels) convolutional neural networks confirm that incorporating emotional change information in the network training process can significantly improve the mood prediction performance, thus highlighting the importance of modelling emotion and mood simultaneously for improved performance in affective state recognition.

CCS Concepts: • **Human-centered computing** → **Ambient intelligence**; • **Computing methodologies** → **Supervised learning by classification**; • **Applied computing** → *Psychology*.

Additional Key Words and Phrases: Mood; Emotion; Convolution neural network; Unimodal; Multimodal

ACM Reference Format:
Soujanya Narayana, Ramanathan Subramanian, Ibrahim Radwan, and Roland Goecke. 2022. To Improve Is to Change: Towards Improving Mood Prediction by Learning Changes in Emotion. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22 Companion), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3536220.3563685

Manuscript submitted to ACM

1 INTRODUCTION

There is mounting evidence that emotions play an essential role in rational and intelligent behaviour. Besides contributing to a richer quality of interaction, they directly impact a person's ability to interact in an intelligent way [21]. Quite often, the terms *mood* and *emotion* are used interchangeably, although differences in duration, intensity and attribution exist. While ***emotion*** is a short-term affective state induced by a source which can last for a few minutes, ***mood*** refers to a longer-term affective state that can last for hours or even days, and be without a causal source [13]. Most research in affective computing has focused on inferring emotional states, while very little research has so far been devoted to automated mood recognition [15] or the joint modelling of the interplay between emotion and mood for improved affective state recognition. Psychological studies on mood have made substantial progress. An eye-tracking study has revealed that positive mood results in better global information processing than a negative mood [23]. The authors in [22] have observed a mood-congruity effect, where positive mood hampers the recognition of mood-incongruent negative emotions and vice-versa. The mood-emotion loop is a theory that posits mood and emotion as distinct mechanisms, which affect each other repeatedly and continuously. This theory argues that mood is a high-level construct activating latent low-level states such as emotions [29]. Recognising the interactions between mood and emotion has the potential to lead to a better understanding of affective phenomena, such as mood disorders and emotional regulation.

On the contrary, mood recognition has rarely been addressed from a computational perspective and only a few studies have explored mood [15]. Body posture and movement correlates of mood have been explored in [27]. User mood is induced via musical stimuli and the authors have observed that head posture and movements characterise happy and sad mood. Katsimerou *et al.* [15] have examined automatic mood prediction from recognised emotions, showing that clustered emotions in the valence-arousal space predict single moods much better than multiple moods within a video. Research on mood prediction has also neglected to investigate the interplay between mood and emotion, though the psychological literature recognises a relationship between the two [20].

From an affective computing viewpoint, developing a mood recognition framework requires ground-truth mood labels for model training, but only very few databases record the user mood (directly or indirectly via an observer). Widely used affective corpora, such as AFEW-VA [16], HUMAINE [9], SEMAINE [19] and DECAF [1] only contain dimensional and/or categorical emotion labels. One of the few datasets with mood ratings is EMMA [14], where the annotations developed represent the overall emotional impression of the human annotator (or observer) for the examined stimulus [14]. Machine learning approaches have been extensively used for inferring emotions from visual, acoustic, textual and neurophysiological data [4, 5, 17, 26, 28]. Contemporary studies emphasise the improved performance of multimodal approaches to the detection of emotional states vis-á-vis unimodal ones [8]. Recent studies characterise mood disorders, such as depression, by examining speech style, eye activity, and head pose [2, 3, 25]. Deng *et al.* [6] propose a multitask emotion recognition framework that can deal with missing labels employing a teacher-student paradigm. *Knowledge Distillation* (KD) is a technique that enables the transfer of knowledge between two neural networks, unifying model compression and learning with privileged information [12, 18]. KD techniques have been employed for facial expression recognition where the teacher has access to a fully visible face, whereas the student model only has access to occluded faces [10].

While our research is ultimately aimed towards mood prediction and understanding the interplay between mood and emotions from video data, the present study is an initial step on this path. We use the AFEW-VA dataset to derive (a) *dominant emotion* labels, which refer to the emotion persisting for most consecutive frames (termed *mood* labels), and (b) Δ or emotion change labels, which represent the change in emotion over a fixed window size. Given the sparsity of

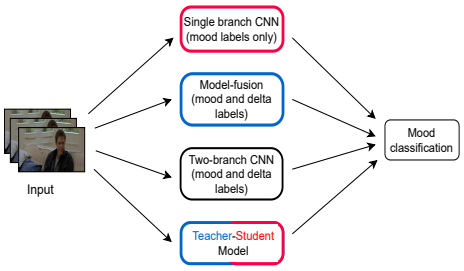

Fig. 2. Overview of the proposed mood recognition framework.

in-the-wild data with mood annotations and the preliminary nature of this study, the dominant emotion labels are used here in lieu of actual mood labels. In the future, we will be using actual mood labels derived from expert annotators. Fig. 1 illustrates how emotion change is captured for an exemplar video clip, while Fig. 2 overviews our dominant emotion or mood prediction framework. A unimodal 3D Convolutional Network Network (3D CNN) is trained using only mood labels, while a two-branch (multimodal) CNN model, multi-layer perceptron, and a teacher-student model are evaluated for fusing emotion-mood information for mood prediction. Empirical evaluation reveals that incorporating emotion change information improves mood prediction performance by as much as 54%, confirming the salience of fine-grained emotional information for coarse-grained mood prediction. This study makes the following contributions:

- To the best of our knowledge, from a computational modelling perspective, this is the first study to examine mood prediction incorporating both mood and emotional information. Mood labels are derived from valence annotations, instead of subjective impressions provided by a human annotator.
- The experimental evaluation of multiple models shows that incorporating emotional change information is beneficial and can produce a significant improvement in mood prediction performance.

2 MATERIALS

2.1 Dataset

Here, the AFEW-VA [16] dataset, a subset of the AFEW [7], comprising 600 video clips extracted from feature films at a rate of 25 frames per second, was used. Video clips in this dataset range from very short sequences ($\sim$ 10 frames) to longer videos (145 frames), and depict various facial expressions. The videos are captured in both indoor and outdoor naturalistic settings [7]. There are 240 subjects in AFEW-VA, who are the actors in the videos. Each clip has per-frame valence and arousal annotations in the $[-10, 10]$ range, annotated by two expert annotators (1 male, 1 female). To examine the interplay between mood and emotion, and perform mood prediction, we assigned a mood label to each clip.

2.2 Labels

2.2.1 ***Mood labels:*** The valence value prevailing over most consecutive frames is considered as the dominant emotion label for a video. As mentioned before, given the lack of in-the-wild datasets with both mood and emotion labels and given the preliminary nature of this study, we consider the dominant emotion label as the mood label in this paper to assess the proposed framework (see Sec. 1). We assigned the valence ranges of $[-10, -3)$, $[-3, 3]$ and $(3, 10]$, to labels of -1 (negative), 0 (neutral), and 1 (positive), respectively. Based on this annotation scheme, the AFEW-VA dataset comprises 59, 400, and 141 videos with positive, neutral, and negative mood labels, respectively. We consider an overlapping sequence of k frames as one input sample. For example, considering a video clip with $n = 10$ frames and $k = 3$, the

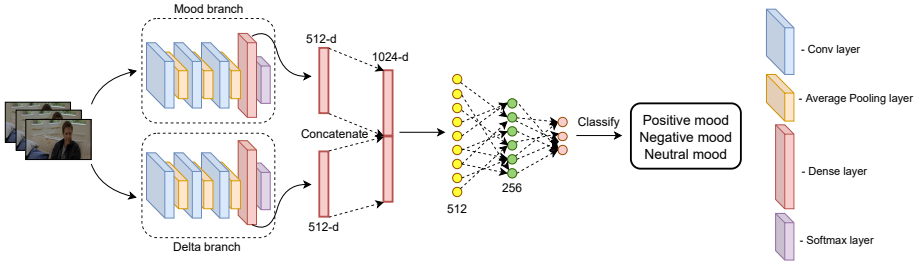

Fig. 3. Architecture of the model fusing mood and Δ information. The legend on the right side shows the unique colours used to describe the layers in the architecture. (Best viewed in colour)

input samples include frames [1–3], [2–4], [3–5], . . ., [8–10]. With $k = 5$, the data comprises a total of 27,651 samples. Each sample is assigned the mood label of its source clip.

2.2.2 ***Emotion Change (Δ) labels:*** Apart from the mood label, we also associate a Δ to every sample, which denotes the change in emotional valence over k frames. For a video clip with n frames, Δ is computed as the valence difference between the t^{th} and $(t-k+1)^{th}$ frames for $t = k, k+1, ..., n$. *E.g.*, considering $k = 5$, if $v_5 = -5$ and $v_1 = -3$, where v_5 and v_1 denote the valence of the fifth and the first frame, respectively, $\Delta = v_5 - v_1 = -2$. As a first step towards mood prediction, we assign the *sign* of the Δ value to be the Δ label (Δ = -1 in the previous example). Δ label is computed as:

$$\operatorname{sgn}\Delta = \begin{cases} -1 & \text{if } \Delta < 0, \\ 0 & \text{if } \Delta = 0, \\ 1 & \text{if } \Delta > 0. \end{cases} \tag{1}$$

Thus, each input video sample has both mood and Δ labels $\in [-1, 0, 1]$.

3 METHODS

Our work leverages the interplay and mutual influence between mood and emotion [15, 29]. We utilise valence annotations to gather information on mood, and perform mood classification using Convolutional Neural Networks. This section describes the unimodal (1-CNN) and multimodal (2-CNN, 2-CNN+MLP, and Teacher-Student (TS) network) models, their architectures and the hyper-parameters employed for model training.

3.1 Single-Branch Mood Classification

For mood classification, we feed a single-branch three-layered CNN (or 1-CNN) with input video samples and mood labels as described in Sec. 2.1. The 1-CNN is denoted using a dashed-rectangle in Fig. 3. Each convolutional layer convolves the input sample with a stride of 3, and the three convolutional layers comprise 16, 32 and 32 kernels of size $3 \times 3 \times 3$, respectively. Each of these layers is followed by average pooling over 2-pixel regions. The output of the third convolutional layer is flattened, followed by batch normalisation. The dense layer comprises 512 neurons, followed by a SoftMax layer with three neurons corresponding to the three mood classes considered in this study.

The input dimensionality for the 3D-CNN is $5 \times 32 \times 32 \times 3$, with each input sample comprising five frames of size $32 \times 32 \times 3$. We use categorical cross-entropy as the loss function for training the model. The fine-tuned hyper-parameters include a learning rate $\in \{10^{-5}, 10^{-3}\}$, batch size $\in \{64, 128, 256\}$, and dropout rate $\in \{0.4, 0.5\}$. The Adam optimizer is used for optimising model learning with stochastic gradient descent.

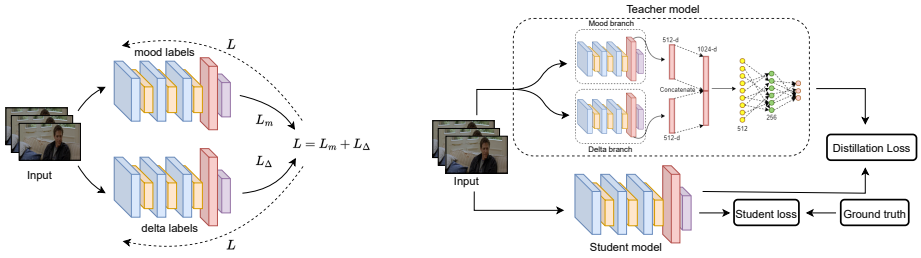

Fig. 4. **(Left)** Architecture of the 2-CNN model, composed of two 1-CNN models. **(Right)** Architecture of the TS-Network. The layers of the model are described in Fig. 3. (Best viewed in colour)

.

3.2 Model Fusion

Given the success of fusion-based emotion inferring approaches [3, 17], we fused the mood and Δ information for mood prediction. To examine the influence of emotion change or Δ on mood, we employ a two-branch CNN model with a Multi-Layer Perceptron (2-CNN+MLP). As shown in Fig. 3, the two three-layered 1-CNNs in each branch are trained independently, with Δ labels fed to one branch and mood labels to the other. From the trained models in each branch, we gather a 512-dimensional vector from the penultimate CNN layer. The vectors from the two branches are concatenated to form a 1024-dimensional feature, which is then passed to an MLP for mood classification. This architecture assumes the availability of Δ labels in both the training and test phases.

The two branches involve identical 1-CNN networks and only differ with respect to their input labels. The input dimensionality and the hyper-parameters for each branch are identical to 1-CNN. The MLP has two dense layers with 512 and 256 neurons, respectively, and a SoftMax layer with three neurons to classify the mood. Overall, we fuse representations learned in the two branches and feed them to an MLP for mood prediction.

3.3 Two-Branch Mood Classification

An alternative to the feature fusion model is to perform end-to-end optimisation [11, 28] with the mood and Δ branch networks. To this end, we employ a 2-CNN model (Fig. 4 (left)), a two-branch model with end-to-end learning, with a three-layered 1-CNN model in each branch. As for the fusion model, the 2-CNN model is composed of the mood and Δ 1-CNN branches. However, this model differs from 2-CNN+MLP with respect to the training process. The categorical cross-entropy losses from the mood and Δ branches (L_m and L_Δ) are summed up, and the cumulative loss is minimised in this model.

Unlike 2-CNN+MLP, which requires Δ labels in the test phase, Δ labels serve as auxiliary information and are only incorporated during 2-CNN model training and back-propagation. Only mood labels are utilised during the test phase, as the mood branch alone is activated for inference.

3.4 Teacher-Student Network

In addition to model fusion and end-to-end optimisation, we employ knowledge distillation [12] to transfer knowledge from a larger teacher network to a smaller student network. Fig. 4 (right) presents the Teacher-Student Network, where the teacher and student networks are as described in Sec. 3.2 and 3.1, respectively. The pre-trained teacher network utilizing both mood and Δ labels (2-CNN+MLP) distills knowledge, while training the student network (1-CNN) only requires mood labels. As inference is again based on the student network, Δ labels are not required during test time.

Table 1. Performance comparison of different models with a 5-frame input sequence ($k = 5$).

Model	1-CNN		2-CNN	2-CNN + MLP	TS-Network	
Label	Mood	Delta	Mood	Mood	Mood (student without teacher)	Mood (student with teacher)
Accuracy ($\mu \pm \sigma$)	0.35 ± 0.10	0.53 ± 0.11	0.73 ± 0.06	0.87 ± 0.15	0.35 ± 0.10	**0.89 ± 0.09**

The student model's SoftMax layer involves a hyper-parameter called the *temperature* T, which controls the smoothness of the output probabilities. Setting $T > 1$ increases the weight of smaller logit[1] values, thus revealing more information about inter-class relations than the one-hot labels [12]. The Kullback–Leibler (KL) divergence is used to compute the distillation loss, while sparse categorical cross-entropy is used as the student loss function. The overall loss of the teacher-student model L_{TS} is the weighted sum of the student loss L_{stu} and distillation loss L_{dis}:

$$L_{TS} = \alpha L_{\text{stu}} + (1 - \alpha) L_{\text{dis}} \tag{2}$$

where α is a training hyper-parameter. The fine-tuned hyper-parameters include a batch size $\in \{16, 64, 128\}$, $T \in \{3, 5, 7\}$ and $\alpha \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$.

3.5 Performance Measures

All models are evaluated via subject-independent 5-fold cross-validation to ensure the same subject does not appear simultaneously in two different folds. We report the mean accuracy over the five folds as the metric for performance evaluation.

4 RESULTS AND DISCUSSION

Table 1 shows the results obtained for each of the models implemented. The 1-CNN model is trained independently with mood and Δ labels. The model trained with Δ labels yields a higher accuracy as compared to the model trained with mood labels, implying that the 1-CNN model is able to learn temporal emotional changes better than the high-level mood construct. The 2-CNN+MLP model, combining information from the mood and Δ CNN branches for mood prediction, achieves a much higher accuracy than the 1-CNN model. This result validates our hypothesis that incorporating emotion change (Δ) information improves mood prediction. Further, the 2-CNN model, with its independent Δ and mood branches, performs better than the 1-CNN model but worse than the 2-CNN model, again confirming that fusing mood and Δ information is beneficial. Finally, the TS-Network composed of the 2-CNN+MLP as the teacher model trained with mood and auxiliary Δ labels, and the student 1-CNN network trained with mood labels gives the highest accuracy among all models. Cumulatively, these results confirm that modelling emotional change information is beneficial for mood prediction.

Fig. 5 qualitatively illustrates how the incorporation of Δ information induces better attention from the classifier. The 1-CNN model trained using only mood labels is unable to focus on the face and bases its incorrect prediction on both the facial and background information. In contrast, the 2-CNN model trained with both mood and Δ labels is able to focus on the face to make the correct mood prediction, consequently improving mood prediction performance.

As Δ is computed over a fixed window size, it captures emotion change over a shorter duration than the actual video. Incorporating Δ information for mood classification implies that the focus is not just on the person's mood, which lingers on for a longer duration, but also on the shifts and variations in short-term emotions that are inherent. This

[1]The logit function is the quantile function associated with the standard logistic distribution.

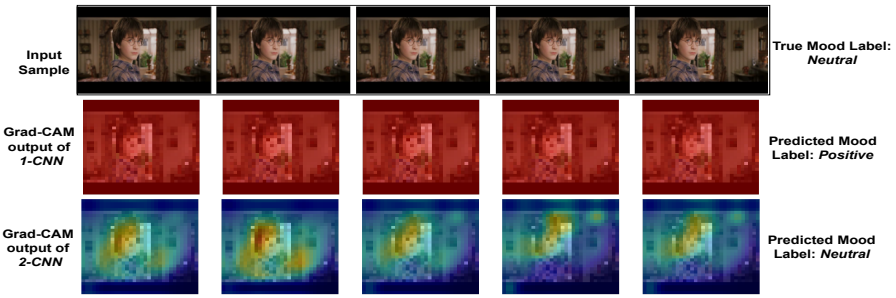

Fig. 5. GradCAM maps [24] depicting improved mood prediction when Δ is learnt by focusing on relevant face parts. *(Top)* An input sample with the ground truth mood label being neutral. *(Centre)* GradCAM maps of the 1-CNN model, with the predicted label as positive. *(Bottom)* Correct prediction using the 2-CNN model. (Best viewed in colour).

aligns with real-world scenarios where a person in a particular mood could experience emotional fluctuations under different circumstances. For example, while in a negative mood, a person could experience a positive emotion following exposure to a positive stimulus, or in a positive mood, one could experience a negative emotion following a negative event. Here, the interplay between emotion and mood would have a dampening effect. The emotion need not necessarily be of opposite valence; an emotion of the same valence as the mood could create a variation (amplification) as well. With the experimental results obtained, it is noteworthy that apart from mood, which speaks of the holistic affective state (as it lingers in the background of one's consciousness for a long duration), the local affective state captured as emotion change is positively contributing to mood prediction, which supports our hypothesis that emotion and mood should be modelled simultaneously for improved affective state prediction.

5 CONCLUSIONS

Increasing evidence in psychology shows a close association between the affective states, mood and emotion, while little has been done in this regard from a computational modelling perspective. While our long-term aim is to examine mood and investigate the interplay between mood and emotions, as a first step, we use the AFEW-VA database, which has per-frame annotations of valence and arousal, and derive the dominant emotion labels from the valence values as an approximation of the mood labels. In addition to these labels, we explore the potential of temporal emotion change for mood prediction. A unimodal CNN and multimodal feature fusion (2-CNN+MLP), 2-CNN and TS networks have been explored in this study. The experimental results demonstrate that learning the emotion change greatly improves mood prediction.

The present study is limited with respect to finding the existence of a relation between the two affective states. The work is also limited in considering the window size for the input sample. In the future, we will investigate the interplay between mood and emotions, by considering actual mood labels derived from expert annotators. We also plan to examine the mutual influence of emotion on mood by taking into account the polarity of the affective states. Exploring the effect of different window sizes for capturing emotion change is also left to future work.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe. 2015. DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses. *IEEE Transactions on Affective Computing* 6, 3 (2015), 209–222.

[2] Sharifa Alghowinem. 2013. From Joyous to Clinically Depressed: Mood Detection Using Multimodal Analysis of a Person's Appearance and Speech. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Vol. 19. IEEE, Geneva, Switzerland, 648–654. https://doi.org/10.1109/ACII.2013.113

[3] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Matthew Hyett, Gordon Parker, and Michael Breakspear. 2016. Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing* 9, 4 (2016), 478–490.

[4] Maneesh Bilalpur, Seyed Mostafa Kia, Manisha Chawla, Tat-Seng Chua, and Ramanathan Subramanian. 2017. Gender and Emotion Recognition with Implicit User Signals. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI '17)*. Association for Computing Machinery, New York, NY, USA, 379–387. https://doi.org/10.1145/3136755.3136790

[5] Ricardo A Calix, Sri Abhishikth Mallepudi, Bin Chen, and Gerald M Knapp. 2010. Emotion recognition in text for 3-D facial expression rendering. *IEEE Transactions on Multimedia* 12, 6 (2010), 544–551.

[6] Didan Deng, Zhaokang Chen, and Bertram E. Shi. 2020. Multitask Emotion Recognition with Incomplete Labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE Press, Buenos Aires, Argentina, 592–599. https://doi.org/10.1109/FG47880.2020.00131

[7] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2012. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE MultiMedia* 19, 3 (2012), 34–41. https://doi.org/10.1109/MMUL.2012.26

[8] Sidney D'Mello and Jacqueline Kory. 2012. Consistent but Modest: A Meta-Analysis on Unimodal and Multimodal Affect Detection Accuracies from 30 Studies. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (Santa Monica, California, USA) *(ICMI '12)*. Association for Computing Machinery, New York, NY, USA, 31–38. https://doi.org/10.1145/2388676.2388686

[9] Ellen Douglas-Cowie, Cate Cox, Jean-Claude Martin, Laurence Devillers, Roddy Cowie, Ian Sneddon, Margaret McRorie, Catherine Pelachaud, Christopher Peters, Orla Lowry, et al. 2011. The HUMAINE database. In *Emotion-Oriented Systems*. Springer, 243–284.

[10] Mariana-Iuliana Georgescu and Radu Tudor Ionescu. 2021. Teacher-Student Training and Triplet Loss for Facial Expression Recognition under Occlusion. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, Italy, 2288–2295. https://doi.org/10.1109/ICPR48806.2021.9412493

[11] Alex Graves and Navdeep Jaitly. 2014. Towards End-to-End Speech Recognition with Recurrent Neural Networks. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14)*. JMLR.org, Beijing, China, II–1764–II–1772.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. https://doi.org/10.48550/ARXIV.1503.02531

[13] Jennifer M Jenkins, Keith Oatley, and Nancy Stein. 1998. *Human emotions: A reader*. Wiley-Blackwell.

[14] Christina Katsimerou, Joris Albeda, Alina Huldtgren, Ingrid Heynderickx, and Judith A Redi. 2016. Crowdsourcing empathetic intelligence: the case of the annotation of EMMA database for emotion and mood recognition. *ACM Trans. Intelligent Systems and Technology* 7, 4 (2016), 1–27.

[15] Christina Katsimerou, Ingrid Heynderickx, and Judith A Redi. 2015. Predicting mood from punctual emotion annotations on videos. *IEEE Transactions on Affective Computing* 6, 2 (2015), 179–192.

[16] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. 2017. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing* 65 (2017), 23–36.

[17] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. 2018. Multi-Feature Based Emotion Recognition for Video Clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 630–634. https://doi.org/10.1145/3242969.3264989

[18] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. 2015. Unifying distillation and privileged information. *preprint arXiv:1511.03643* (2015).

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

[20] William N Morris. 1992. A functional analysis of the role of mood in affective systems. *Emotion* (1992), 256–293.

[21] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[22] Petra Claudia Schmid and Marianne Schmid Mast. 2010. Mood effects on emotion recognition. *Motivation and Emotion* 34, 3 (2010), 288–292.

[23] Petra C Schmid, Marianne Schmid Mast, Dario Bombari, Fred W Mast, and Janek S Lobmaier. 2011. How mood states affect information processing during facial emotion recognition: An eye tracking study. *Swiss Journal of Psychology/Schweizerische Zeitschrift für Psychologie/Revue Suisse de Psychologie* 70, 4 (2011), 223.

[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, Italy, 618–626. https://doi.org/10.1109/ICCV.2017.74

[25] Mohammed Senoussaoui, Milton Sarria-Paja, João F. Santos, and Tiago H. Falk. 2014. Model Fusion for Multimodal Depression Classification and Level Detection. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* (Orlando, Florida, USA) *(AVEC '14)*. Association for Computing Machinery, New York, NY, USA, 57–63.

[26] Abhinav Shukla, Shruti Shriya Gullapuram, Harish Katti, Mohan Kankanhalli, Stefan Winkler, and Ramanathan Subramanian. 2022. Recognition of Advertisement Emotions With Application to Computational Advertising. *IEEE Transactions on Affective Computing* 13, 2 (2022), 781–792. https://doi.org/10.1109/TAFFC.2020.2964549

[27] Michelle Thrasher, Marjolein D. Van der Zwaag, Nadia Bianchi-Berthouze, and Joyce H. D. M. Westerink. 2011. Mood Recognition Based on Upper Body Posture and Movement Features. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction* (Memphis, TN) *(ACII'11)*. Springer-Verlag, Berlin, Heidelberg, 377–386.

[28] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Shanghai, China, 5200–5204. https://doi.org/10.1109/ICASSP.2016.7472669

[29] Muk Yan Wong. 2016. The mood-emotion loop. *Philosophical Studies* 173, 11 (2016), 3061–3080.