A Weakly Supervised Approach to Emotion-change Prediction and Improved Mood Inference

Soujanya Narayana[1*], Ibrahim Radwan[1], Ravikiran Parameshwara[1], Iman Abbasnejad[2], Akshay Asthana[2], Ramanathan Subramanian[1] and Roland Goecke[1]
[1]Human-Centred Technology Research Centre, University of Canberra, Australia
[2]Seeing Machines Ltd., Australia

***Abstract*—Whilst a majority of affective computing research focuses on inferring emotions, examining mood or understanding the *mood-emotion interplay* has received significantly less attention. Building on prior work, we (a) deduce and incorporate emotion-change (Δ) information for inferring mood, without resorting to annotated labels, and (b) attempt mood prediction for long duration video clips, in alignment with the characterisation of mood. We generate the emotion-change (Δ) labels via metric learning from a pre-trained Siamese Network, and use these in addition to mood labels for mood classification. Experiments evaluating *unimodal* (training only using mood labels) vs *multimodal* (training using mood plus Δ labels) models show that mood prediction benefits from the incorporation of emotion-change information, emphasising the importance of modelling the mood-emotion interplay for effective mood inference.**

***Index Terms*—Mood inference, Emotion change, Siamese network, Contrastive Loss, Teacher-student network, Unimodal, Multimodal**

I. Introduction

Over the past two decades, there has been an enormous increase in the research on inferring affective states (characterised by emotions, moods, *etc*.) from unimodal and multimodal data. Several studies emphasise on the importance of emotional regulation for the successful functioning of human mind [2], [3], as they play an indispensable role in rational decision-making, perception, attention, and other diverse cognitive functions [4]. While the terms *emotion* and *mood* are often used synonymously, the two affective phenomena are distinct in terms of duration, intensity, attribute, and behavioural impact. *Emotion* is a short-term affective state, lasting for at most a few minutes, and is typically elicited by a contextual event/stimulus. On the contrary, *mood* is considered to be a long-term diffuse affective state lasting for hours, which may emerge without an apparent cause [5].

Akin to emotions, mood has an impact on cognitive processes like human creativity, evaluative judgement, and memory retrieval, etc [4]. Mood is also known to generate cognitive bias and influence human emotion recognition [6]. Going further, mood disorders like depression and bipolar disorder result in the impairment of emotional processing abilities [7], altered facial expression understanding [8], and olfactory perception [9]. Recently, the focus of numerous studies is on building emotionally-aware systems to better understand human behaviour, and facilitate enhanced human-computer interactions. Advancements in machine learning techniques have enabled automatic emotion recognition from unimodal and multimodal data, for instance, facial expressions [10], physiological signals [11]–[13], videos [14], *etc*. Although substantial progress has been made in psychology to understand mood, negligible work has focused on computationally inferring mood. Furthermore, the psychology literature recognises an association between emotions and mood [15]; theories state that despite being distinct mechanisms, they affect one another repeatedly and continuously. Nevertheless, hardly any research has been devoted towards computationally modelling the mood-emotion interplay for mood inference.

Preliminary studies on mood recognition [16], [17] infer mood via body posture and 3D pose data in a controlled setting. As a step towards *in-the-wild* mood prediction, [18] uses deep learning to perform mood classification on affective videos. It observes that mood prediction improves on utilising emotion-change information. The premise in [18] is the existence of continuous emotion (valence) labels along with mood annotations during the training and testing phases. Reliance on continuous emotion labels represents a significant overhead, as these labels may not be available in real-world settings.

This work is inspired by and extends the idea proposed in [18], and obviates the need for ground-truth emotion labels by *deducing* emotional change information and utilising it for mood inference. Our mood inference framework is illustrated in Fig. 1. Specifically, emotion change is modelled in terms of emotional (dis)similarity between a pair of video frames via a Siamese Network with contrastive loss.

We present mood inference results on the *AffWild2* database [1], where (a) video mood labels are derived as in [18]; (b) *pseudo* emotion-change (Δ) labels are derived via a pre-trained Siamese Network; (c) a 3-dimensional Convolutional Neural Network (3D-CNN) with a ResNet18 backbone and projection head is trained with mood labels; (d) a 3D-CNN with ResNet18 backbone and branched projection heads is trained with both mood and Δ labels, and (e) a Teacher-Student (TS) network [19] is employed, where the teacher distills the privileged Δ-specific knowledge to the student for mood inference.

Consistent with [18], we observe that mood prediction performance improves when emotion similarity information is

This research is partially funded by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP190101294).
*Corresponding author: soujanya.narayana@canberra.edu.au

Final Author Version

979-8-3503-2743-4/23/$31.00 ©2023 IEEE

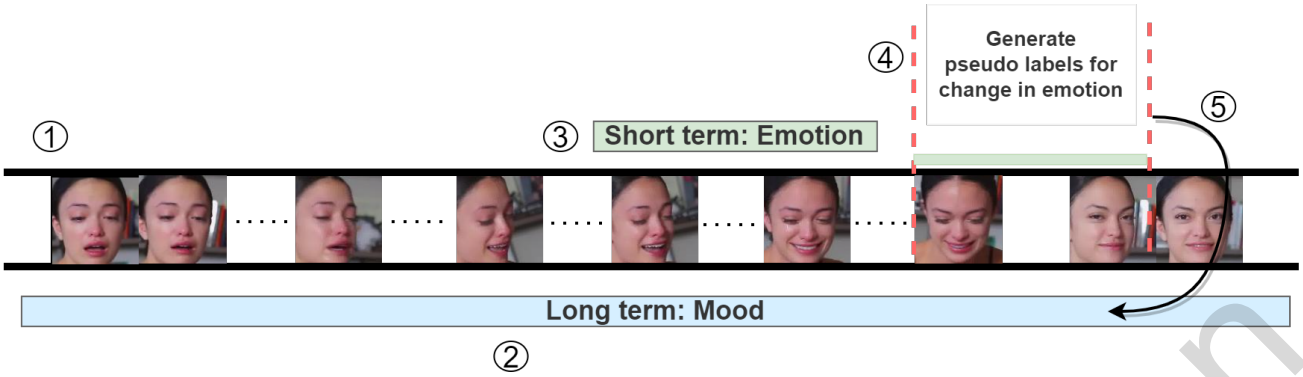

Fig. 1. **Study Overview:** (1) We consider the publicly available *AffWild2* [1] video dataset for automated mood inference. (2) A mood label is derived for the video, aligning with the characterisation of mood as a long-term affective state. (3) Our study also seeks to automatically capture affective emotion-change (Δ) information over shorter durations (time windows). (4) We generate pseudo-emotion-change (Δ) labels via metric learning obviating the need for valence annotations. (5) We then incorporate the generated Δ labels to automatically infer the mood class as *positive*, *negative* or *neutral*.

incorporated, emphasising the prominence of short-term affect (emotion-change) for long-term affect (mood) inference. To summarise, the main contributions of this work are as follows:

1) We propose to infer mood employing *emotional similarity*, which models emotion change between a pair of images. To this end, we train a 3D-CNN with two branches, which are respectively trained via video frames annotated with mood and Δ labels.
2) We weakly label the AffWild2 [1] dataset, by generating Δ labels for video frame pairs using a pre-trained Siamese Network with contrastive loss.
3) Assessing various models, we demonstrate that incorporating emotion-change (Δ) information via emotion similarity benefits mood recognition and enhances mood prediction performance. Similar trends are observed in the experiments employing ground truth Δ_{GT} labels.
4) Through an ablation study, we verify the effectiveness of the various components of the proposed mood classification framework outlined in Fig. 1.

II. Related Work

In this section, we present studies examining the affective phenomena of emotion and mood (Sec. II-A), the various affective databases available for affect inference, and machine learning studies examining mood inference (Sec. II-B). The motivation for this study given the literature context is presented in Sec. II-C.

A. Emotion, Mood and their Interconnectedness

While there are multiple definitions of emotion, the following characterisation appears to be consistent. Emotion is considered to be an episode of neurophysiological and cognitive change in response to an external or internal stimulus [20]. The concepts of emotion and mood are distinguished based on the factors of duration, trigger, intensity, and behavioural impact [21]. The former are short-term, lasting for a few seconds and are elicited based on stimulus events. The level of response to the stimuli and the corresponding emotional expression is recognised to be of relatively high intensity [22]. In contrast, moods are considered to be enduring affective states, lasting for hours or even days without being instantiated by a stimulus. They are regarded to be diffuse with low levels of intensity [21].

The human physiological state and mind are both influenced by and reflective of mood, as it directly influences human health and well-being. Besides having an impact on evaluative judgements, mood also governs memory retrieval [23]. Mood-congruency, which refers to the match between a person's mood and his/her thoughts [24], is observed in [25], where the authors examine mood effects on emotion recognition. Happy mood impedes the recognition of mood-incongruent sad emotions, while sad mood obstructs the recognition of happy emotions. The interplay between mood and emotion is described by the mood-emotion loop [23], a theory which proposes that mood and emotion are distinct mechanisms forming a loop, and are reciprocally influencing one another.

B. Computational Studies on Mood

Most research on affective state inference has focussed on emotions, as opposed to mood. Likewise, the prevailing affective databases to aid behavioural and computational studies with various modes of data predominantly target emotions.

1) ***Databases****:* Audio-visual databases, such as AFEW [26], DECAF [27], Ascertain [13] *etc*., comprise videos of emotional episodes with categorical emotion annotations, namely, happy, sad, fear, disgust, anger, surprise, and neutral. RECOLA [28], AFEW-VA [29], AffWild2 [1], *etc*., manifest continuous emotion annotations of *valence* (degree of pleasantness or unpleasantness) and *arousal* (degree of excitement or calmness). These databases, designed by considering factors such as recording environment, duration, and annotation type, best suit emotion inference tasks. The test set of the AVEC 2013 challenge, annotated for the level of depression, is one of the few databases with mood annotations. However, depression, a mood disorder, is not a commonly observed mood state. EMMA [30] is an acted video database recorded in a controlled setting with mood annotations.

Final Author Version

*2) **Computational Approaches**:* In [17], the authors use body posture and head movement features to capture the affective state while listening music. A vertical position of the head is observed in a positive mood and a downward position otherwise. A 3-dimensional pose tracker is used in [16] to infer physical attributes and the mood of the person by capturing walking motions. The authors of [31] perform automatic mood recognition from recognised emotions, and show that clustered emotions in the valence-arousal space are better predictors of a single mood as compared to multiple moods within a video.

Mood prediction using various 3D-CNNs are performed in [18] using the AFEW-VA [29] dataset. Utilising the valence annotations, the authors compute the valence differential to infer mood and demonstrate that incorporating valence change improves mood prediction performance. Although this study is a promising step towards automatic mood inference, only video clips of very short duration ($\approx$ 0.04 seconds) are considered, which may not adequately depict subject mood in the video.

C. Novelty of our Study

A thorough examination of the literature reveals the following: (a) While significant research has been conducted to infer emotions automatically, mood inference and modelling the mood-emotion interplay have been neglected from a computational perspective; (b) Existing affective databases are richly annotated for emotions, while labelled data for mood inference are sparse; and (c) Existing studies, which infer mood using emotions, require continuous valence annotations and consider clips of very short duration ($\approx$ 0.04 seconds) for this purpose.

Differently, this study uses *deduced* emotional similarity information in lieu of valence-differential (Δ) labels, obviating the need for ground truth Δ annotations. This setting resembles real-world scenarios where valence annotations may not be accessible for inferring mood. This work proposes to (1) deduce emotion change (Δ) labels using a Siamese Network trained with contrastive loss, and (2) incorporate Δ labels for inferring mood. In addition, ablation studies are performed to empirically examine the effect of different components of the proposed mood inference framework.

III. Databases and Label Generation

This section describes the databases considered in this study as well as the labelling procedure.

A. AffWild2 Database

We consider AffWild2 [1], a publicly available affective video database with continuous dimensional (valence, arousal) and categorical emotion annotations for performing mood classification. AffWild2 comprises 564 *in-the-wild* videos collected from *YouTube*. There are a total of 2,816,832 frames with 455 subjects (277 male, 178 female). The videos are annotated by four experts for continuous valence and arousal values, and the average of the four raters is considered as the final rating, while three experts annotated the categorical emotion labels. The valence and arousal annotations are in the range $[-1, 1]$, and the emotional categories are *happy*, *sad*, *disgust*, *anger*, *fear*, *surprise*, *neutral*, and *other*. The frames with annotated values outside this range are discarded as suggested by the dataset providers. The partitioning of the database into training, validation and test sets is done in a subject-independent manner, so that every subject is present in one of the three partitions. This partitioning results in 341, 71 and 152 videos respectively in the training, validation and test sets. The validation set is utilized for evaluating our proposed approach, as the test set has not been released.

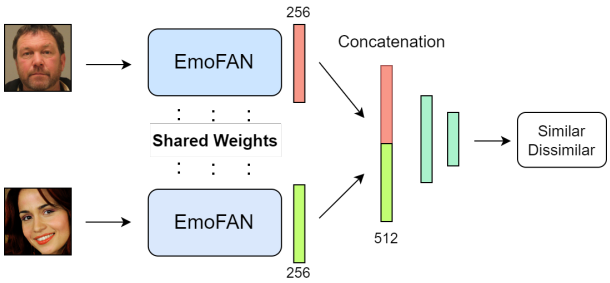

Fig. 2. Siamese network for modelling emotion-change (Δ).

B. Mood labels

Since current *in-the-wild* affective video databases lack mood annotations, as the closest alternative, we utilize valence annotations to assign mood labels for each video in the AffWild2 database, as done in [18]. The three mood categories considered are *positive* (+1), *negative* (-1), and *neutral* (0). As mood denotes a long-term affective state [4], the most persistent valence value (maximum number of consecutive frames) is considered for assigning a mood label. Mood label is assigned to -1, 0, or $+1$ if the valence values for maximum number of consecutive frames respectively lie in the range $[-1, -0.3)$, $[-0.3, 0.3]$, and $(0.3, 1]$.

C. AffectNet Database

In order to automatically generate the emotion change (Δ) labels, we employ the *AffectNet* [32] dataset to train a Siamese Network (described in Sec. III-D). AffectNet, an affective database curated for automatic facial emotion recognition tasks, comprises around 420,300 facial images captured under natural conditions. Twelve experts annotated the data with continuous valence and arousal values, and eight emotion categories (*happy*, *sad*, *disgust*, *anger*, *fear*, *surprise*, *neutral*, and *contempt*). Partitioning the dataset with the criteria of having 500 images for each of the nine emotion categories in the validation set, results in 287,151 and 4500 images in the training and test sets respectively.

D. Emotion Change (Δ) Labels

This work seeks to eliminate the need for valence annotations (as proposed in [18]) to generate Δ labels. Differently, we propose to use a Siamese Network with contrastive loss to deduce emotion-change (Δ) between video frames in terms of similarity (little change in emotion) or dissimilarity (significant

Final Author Version

change in emotion). A Siamese network is a neural network that discriminates if a pair of input data samples are similar or dissimilar [33]. Multiple emotion-related tasks have employed a Siamese network, and observed promising performance [34], [35]. We use contrastive loss in the Siamese network, which learns a similarity metric that minimises the distance between similar image pairs, while maximising the distance between dissimilar pairs. Distance between the pairs is compared to a *margin* value, and the contrastive loss function enforces a smaller distance between similar pairs, and a larger distance between dissimilar pairs.

*1) **Siamese Network**:* In this study, *similarity* refers to little or no change in the emotional facial expression between a pair of frames. Fig. 2 shows the architecture of the Siamese network, which comprises identical sub-networks for classifying the input frames as *similar* or *dissimilar*. Each sub-network involves an EmoFAN [36] network as the encoder, $E(\cdot)$, which maps the input image to a vector. We employ a pre-trained EmoFAN model as it has demonstrated high performance in emotion recognition tasks. For two images x_1 and x_2 we obtain, $v_1 = E(x_1) \in \mathcal{R}^{\mathcal{D}_\mathcal{E}}$ and $v_2 = E(x_2) \in \mathcal{R}^{\mathcal{D}_\mathcal{E}}$, where $\mathcal{D}_\mathcal{E} = 256$. The embeddings v_1 and v_2 are concatenated, $v = v_1 \| v_2 \in \mathcal{R}^{\mathcal{D}_\mathcal{C}}$, where $\mathcal{D}_\mathcal{C} = 512$. v is fed into a projection head $P(\cdot)$, which maps it to a vector, $u = P(v) \in \mathcal{R}^{\mathcal{D}_\mathcal{P}}$, where $\mathcal{D}_\mathcal{P} = 2$. $P(\cdot)$ is a Multi-Layer Perceptron (MLP) with three fully connected (fc) layers, comprising 256, 128, and 2 neurons, respectively. The neurons in the last fc layer refer to the two classes, *similar* and *dissimilar*. The inputs to the fc layers are normalised with zero mean and unit variance, before feeding to the ReLU activation.

*2) **Contrastive Loss**:* As opposed to using the cross-entropy loss, $\mathcal{L}_\mathcal{B}$, alone for binary classification (similar/dissimilar), we additionally consider using the contrastive loss, $\mathcal{L}_\mathcal{C}$, given by,

$$\mathcal{L}_\mathcal{C} = \frac{1}{N}\sum_{i=1}^{N} y_i(1-d_i) + (1-y_i)\max(0, d_i - m) \quad (1)$$

where N is the batch size, y_i is the label indicating whether the two input samples are similar (1) or dissimilar (0), d_i is the cosine distance between the embeddings v_1 and v_2, and m is the margin. The total loss is given by,

$$\mathcal{L}_\mathcal{T} = \lambda\mathcal{L}_\mathcal{B} + (1-\lambda)\mathcal{L}_\mathcal{C} \quad (2)$$

where λ is a training hyperparameter.

*3) **Deducing the Δ Label**:* The Siamese network is trained on the images in the AffectNet database. The groundtruth labels for the Siamese network y_i are derived as 1 if the emotion categories of the input pair are the same, 0 otherwise. The network is trained for 40 epochs, using the Adam optimiser with the learning rate decreased by a factor of 10 for every 10 epochs, with the initial learning rate as 0.0001. The batch size is set to 64, with the dropout rate as 0.3, margin m in the contrastive loss set to 0.25, and λ set to 0.5.

The Siamese network achieves an accuracy of 68%. This model is used to generate Δ labels for the AffWild2 video frame pairs.

E. Generating input samples for mood inference

The AffWild2 dataset comprises videos of long duration, with an average length of $\approx$ 3 minutes (minimum duration of 0.03 minutes, maximum duration of 26.22 minutes). From each video, using a sliding window approach, we generate clips with a stride s, where each clip is a collection of sampled frames. Each clip c is of temporal length t, which refers to the duration of the clip (number of frames). Constructing clips by including all the frames in c increases the computational load and time substantially. Hence, to address this computational impediment, we significantly reduce the number of frames in c, and sample n frames at equal intervals of time. Clips generated from each video contain frames from the parent video alone, and no frames from other videos.

c is assigned the mood label of its parent video, implying all clips generated from a source video are assigned the same mood label. To generate the Δ label for c, the first frame and the last frame of c are fed to the trained Siamese Network (described in Sec. III-C). The model returns 0 or 1 as the Δ label, by evaluating the dissimilarity or similarity of the emotion displayed in this pair of frames. Hence, each clip is associated with a mood label $\in$ {0, +1, -1}, and a Δ label $\in$ {0, 1}.

IV. Mood Classification Approach

The capability of 3D-CNNs to capture the temporal dependencies in the input data, along with spatial information, has resulted in their extensive usage. ResNet18-3D [37] (R3D), a 3D variant of the ResNet architecture, is commonly used as a backbone network in many 3D-CNN architectures for facial emotion inference [38], [39]. Leveraging the interplay between mood and emotions, we utilise the mood and Δ labels to perform mood classification. The various models used in this study are described as follows.

A. ResMood

Fig. 3 (left) shows ResMood, a model trained with mood labels alone, which consists of a ResNet18-3D, $R3D(\cdot)$, as the backbone and a projection head, $P(\cdot)$. The backbone maps each input sample x to a representation vector, $v = R3D(x) \in \mathcal{R}^{\mathcal{D}_\mathcal{B}}$, where $\mathcal{D}_\mathcal{B} = 1024$. The projection head $P(\cdot)$ further maps v to a vector $z = P(v) \in \mathcal{R}^{\mathcal{D}_\mathcal{P}}$, where $\mathcal{D}_\mathcal{P} = 3$. $P(\cdot)$ is instantiated as a Multi-Layer Perceptron (MLP), with three fully-connected (fc) layers comprising 512, 256, and 3 neurons, respectively. The 3 neurons in the last fc layer denote the three mood classes, *positive*, *negative*, and *neutral*. The inputs to each layer in the MLP are normalised batch-wise with zero mean and unit variance before feeding them to the ReLU activation function.

B. ResMoodEmo

Fig. 3 (left) shows ResMoodEmo, a model for performing mood classification trained with both mood and Δ labels. Distinct from ResMood, ResMoodEmo is composed of $R3D(\cdot)$ as the backbone, and two projection heads $P_M(\cdot)$ and $P_\Delta(\cdot)$, branching out for mood and Δ classification, respectively.

Final Author Version

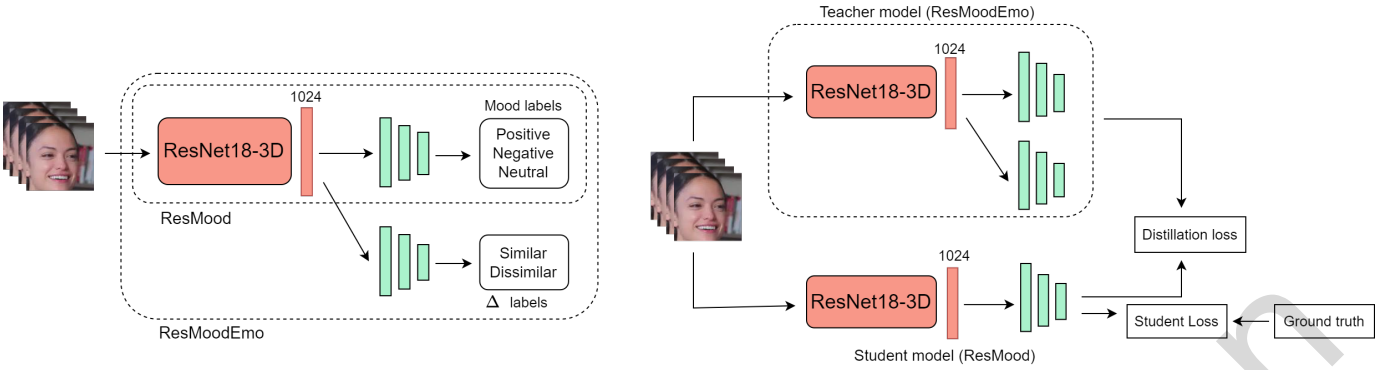

Fig. 3. (Left) The architecture of ResMood is shown within the inner dashed rectangle. The outer dashed rectangle represents ResMoodEmo. (Right) The architecture of the TS-Net. (Best viewed in colour).

$R3D(\cdot)$ maps an input x to a vector $v = R3D(x) \in \mathcal{R}^{\mathcal{D}_\mathcal{B}}$, where $\mathcal{D}_\mathcal{B} = 1024$. Further, the projection $P_M(\cdot)$ maps v to a vector $z_1 = P_M(v) \in \mathcal{R}^{\mathcal{D}_\mathcal{M}}$, and $P_\Delta(\cdot)$ maps v to a vector $z_2 = P_M(v) \in \mathcal{R}^{\mathcal{D}_\Delta}$, where $\mathcal{D}_\mathcal{M} = 3$ and $\mathcal{D}_\Delta = 2$, respectively. $P_M(\cdot)$ and $P_\Delta(\cdot)$ are both configured as MLPs with three fc layers, but differing in the number of neurons in the last layer. As $P_M(\cdot)$ is the branch used for mood classification, the last fc layer has 3 neurons for classifying the three mood classes *positive*, *negative*, and *neutral*, whereas $P_\Delta(\cdot)$ used for classifying Δ labels has two neurons denoting the *similar* and *dissimilar* classes. In both projection heads, the inputs to each layer are normalised with zero mean and unit variance prior to being input to the ReLU activation function. Compared to ResMood, ResMoodEmo has an additional branch after the $R3D(\cdot)$ to incorporate the emotion change (Δ) information. The losses of each branch $\mathcal{L}_\mathcal{M}$ and $\mathcal{L}_\Delta$ are summed up and the cumulative loss $\mathcal{L} = \mathcal{L}_\mathcal{M} + \mathcal{L}_\Delta$ is optimised.

C. Teacher-Student Network

Similar to [18], we employ knowledge distillation [19], a technique used to transfer knowledge from a larger (teacher) model to a smaller (student) model. In this method, the goal is to train the student model to mimic the output probabilities of the teacher model, in addition to the predicting the true labels. Fig. 3 (right) presents the Teacher-Student Network (TS-Net), where ResMoodEmo is used as the teacher model (see Sec. IV-B), and ResMood is used as the student model (see Sec. IV-A). The teacher model, trained with both mood and Δ labels, distills knowledge to the student, which is only trained with mood labels. Since the performance of the student model alone is evaluated, Δ labels are not utilised during the testing phase. The SoftMax layer of the student model has a hyper-parameter called *temperature* (T), which regulates the softness of the output class probabilities. Using low temperature values produces a sharper probability distribution, facilitating the student to focus on the relative differences in the probabilities of the classes. A weighted sum of the distillation loss, $\mathcal{L}_\mathcal{D}$, measuring the difference between the outputs of the teacher and student models, and the student loss, $\mathcal{L}_\mathcal{S}$, a typical supervised loss is optimised in the TS-Net, $\mathcal{L} = \alpha\mathcal{L}_\mathcal{S} + (1-\alpha)\mathcal{L}_\mathcal{D}$, where α is a training hyperparameter.

D. Implementation details

All experiments are based on using the open-source library PyTorch. The models are trained on Nvidia GeForce RTX 3090 GPU with 24GB memory. We use the videos with cropped and aligned faces provided in the AffWild2 database. To generate the input samples (see Sec. III-E), we set the temporal length $t = 100$, with the number of frames in each sample $n = 5$, and the stride $s = 3$. The models ResMood, ResMoodEmo, and the TS-Net are trained using the Adam optimiser with the learning rate reduced by a factor of 10 for every 10 epochs, and the base learning rate set to 0.0001. The models are trained for 30 epochs with a batch size of 128 and the dropout rate is 0.5. In the TS-Net, we validated with the temperature values $\in \{3, 5, 7\}$ and $\alpha \in \{0.05, 0.1, 0.15, 0.2\}$, as shown in Table VI.

V. Results and Discussion

Due to an imbalance in mood classes in the test set, we use weighted F1-score as the performance evaluation metric in all our experiments. Table I shows the results of ResMood, ResMoodEmo, and TS-Net. While ResMood is trained with mood labels alone, ResMoodEmo is trained with both mood and Δ labels. In the TS-net, ResMoodEmo is the teacher model, and ResMood is the student model, implying that the teacher is pre-trained with both mood and Δ labels, while the student is trained with mood labels alone.

ResMoodEmo yields a higher F-score as compared to ResMood, indicating that the ResMoodEmo learning temporal short-term emotion changes, better predicts mood than ResMood, which only employs mood labels. Without resorting to the valence differential for gathering emotion change information as done in [18], training a Siamese Network with contrastive loss, and generating the Δ labels results in a competitive performance in mood prediction. Further, by using the emotion change labels, we are capturing variations over a short duration simultaneously for characterising mood. The results indicate that (local) emotion variations contribute towards understanding the (global) mood.

A similar trend is observed in the TS-Net, as it yields a higher F-score than the ResMood model. This shows that

Final Author Version

TABLE I
PERFORMANCE RESULTS (WEIGHTED F-SCORE) OF THE MODELS.

Model	Train labels	F-score
ResMood	Mood	0.65
ResMoodEmo	Mood and Δ	**0.78**
TS-Net	Mood (Student)	**0.78**

TABLE II
ABLATION STUDY RESULTS WITH Δ_{GT} LABELS

Model	Train labels	F-score
ResMood	Mood	0.63
ResMoodEmo	Mood and Δ_{GT}	0.73
TS-Net	Mood (Student)	0.66

the teacher, possessing the privileged knowledge concerning emotion change, is able to effectively distil knowledge to the student (ResMood) during the training phase. With the Δ labels being implicitly given as soft labels from ResMoodEmo, the performance of the student increases, as compared to the standalone ResMood model. Cumulatively, these results show that using the pseudo-emotion-change information enhances the mood prediction performance.

For the Siamese Network trained on the AffectNet dataset, generating effective Δ labels from AffWild2 is crucial. The obtained results show that the Δ labels generated are reliable as they improve mood prediction performance; the investigation of an optimal architecture for the Siamese network is left to future work. Overall, our results confirm that despite not using the valence differential labels to denote emotion changes, performing weak supervision using the Δ labels in ResMoodEmo and TS-Net improves mood prediction performance.

A. Ablation Studies

To corroborate the above findings, we perform the following ablation studies and examine the effectiveness of various components of our approach.

*1) **Using Ground Truth Emotion-change (Δ_{GT}) Labels**:* The increase in the F-score using ResMoodEmo as compared to ResMood, could be attributed to the contribution of the Δ labels for mood inference or the implicit efficiency of the Siamese Network. For better comprehension of the results, we use the eight emotion labels available for each video frame in the AffWild2 dataset to obtain Δ_{GT} (*similar* or *dissimilar*) labels. However, since not all videos have the categorical emotion annotations, the total number of samples reduced from 191,552 to 53,026. These samples are obtained as described in Sec. III-E. For each video sample, there is an assigned mood label and Δ_{GT} label (similarity between first frame and last frame). The results of ResMood and ResMoodEmo are shown in Table II. It is noteworthy that a similar trend to Table I is observed here. This indicates that, (a) Δ labels generated by the Siamese Network are effective, as they result in a competitive performance, (b) emotion information positively contributes in mood inference in a dataset-agnostic manner (Δ labels are deduced from the Siamese Network trained using AffectNet, whereas the Δ_{GT} labels are obtained from AffWild2.), and (c) achieving a comparable result without using the Δ_{GT} highlights the robustness of the proposed mood inference approach.

*2) **Number of Frames in the Input Video Sample**:* Table III shows the results of varying the number of frames in the input samples, while fixing the temporal length (t) to 100. Temporal information plays a crucial role in examining mood, and variation in the number of frames in each sample clarifies if increasing the information provided to the model facilitates mood inference. The performance of ResMoodEmo increases when the number of frames are set to 3, 5, or 7, but decreases when the number of frames is increased to 9. Using the TS-Net, when the input samples have 3 frames, no change is observed, while with 7 and 9 frames, the performance reduces. Overall, the comparison shows that 5 frames in the input sample maximally increases mood prediction performance.

*3) **Temporal Length (t) of the Sample**:* Table III presents the results of varying t of the input samples while fixing the number of frames (n) in each sample to be 5. Since mood is an enduring affect, it is important to consider long sequences of data for automatic mood inference. For $t = 50$, and $t = 150$, the F-score obtained with ResMood increases, while the F-score remains the same for $t = 200$, as compared to using $t = 100$. Although varying t results in a comparable F-score with ResMood, it decreases in ResMoodEmo and TS-Net with lengths of 50, 150, and 200, as shown in the respective *% increase to ResMood* columns of Table IV. The highest F-score with ResMood and the least F-score with ResMoodEmo is observed using $t = 150$. Since emotion is a short-term affect, observing the changes in the emotion over longer sequence of time causes a detrimental effect on mood inference.

*4) **Varying the Backbone Architecture**:* Table V reports the results when the backbone architecture in the models is changed. The F-score for ResMood increases slightly as the depth of the ResNet increases. With ResNet18 and ResNet50, a general trend of increase in the mood prediction performance using ResMoodEmo and TS-Net is observed. Using ResNet34, an increase in F-score is observed with ResMoodEmo as compared to ResMood, but with TS-Net, the F-score remains the same. ResNet18, a lighter architecture as compared to its counterparts, results in the maximum F-score for ResMoodEmo and TS-Net, and largest % increase from ResMood.

*5) **Temperature and α**:* The results with varying temperature (T) and α values are shown in Table VI. For varying α, as T increases, the F-score reduces. This is due to the fact that high values of temperature soften the output class probabilities, while with low temperature values, the relative differences in the probabilities are captured. For $T = 5$ and $T = 7$, the F-score either remains the same or increases as α increases, while for $T = 3$, no general trend is observed. As described in Sec. IV-C, α is the weight of the student loss function, indicating that lower values of α imply a higher weighting for the distillation loss, enabling the student model to get closer

Final Author Version

TABLE III
ABLATION STUDY VARYING NUMBER OF FRAMES (n) IN THE INPUT SAMPLES. BEST RESULTS OBTAINED ARE HIGHLIGHTED IN BOLD.

Number of frames	ResMood	ResMoodEmo		TS-Net	
	F-score	F-score	% increase to ResMood	F-score	% increase to ResMood
3	0.67	0.68	+1.49	0.67	0
5	0.65	**0.78**	**+20**	**0.78**	**+20**
7	0.68	0.72	+5.88	0.65	-4.41
9	0.70	0.63	-10	0.52	-25.71

TABLE IV
ABLATION STUDY RESULTS FOR VARIOUS TEMPORAL LENGTHS (t) OF THE INPUT SAMPLE.

Temporal length	ResMood	ResMoodEmo		TS-Net	
	F-score	F-score	% increase to ResMood	F-score	% increase to ResMood
50	0.67	0.65	-2.99	0.66	-1.49
100	0.65	**0.78**	**+20**	**0.78**	**+20**
150	0.69	0.65	-5.80	0.64	-7.25
200	0.65	0.64	-1.54	0.64	-1.54

TABLE V
ABLATION STUDY RESULTS FOR VARIOUS BACKBONE ARCHITECTURES.

Backbone	ResMood	ResMoodEmo		TS-Net	
	F-score	F-score	% increase to ResMood	F-score	% increase to ResMood
ResNet18	0.65	**0.78**	**+20**	**0.78**	**+20**
ResNet34	0.66	0.70	+6.06	0.66	0
ResNet50	0.68	0.75	+7	0.75	+7

TABLE VI
ABLATION STUDY RESULTS FOR THE TS-NET WITH VARIOUS TEMPERATURE AND ALPHA VALUES.

T/α	0.05	0.1	0.15	0.2
3	**0.78**	0.72	0.69	0.72
5	0.64	0.64	0.68	0.69
7	0.64	0.66	0.66	0.71

to the teacher model.

Overall, the ablation study results confirm that the generated Δ labels are an effective alternative to Δ_{GT} and produce a comparable effect, as incorporating these labels improves mood prediction performance.

VI. CONCLUSION

The aim of this study is to examine mood from a computational perspective by incorporating emotion similarity information. Different from prior studies, without using the valence differential, this study proposes to use emotion change information by employing a metric learning approach. To this end, a Siamese network is trained using the AffectNet database and the trained model is used to generate pseudo-Δ labels for a pair of frames in the AffWild2 database. For mood classification, we employ ResMood, a model trained with mood labels alone, ResMoodEmo trained with mood labels and Δ labels, and TS-Net, a teacher-student network with ResMoodEmo as the teacher to distil knowledge to ResMood. Higher F-scores are observed with models trained with both mood and Δ labels as compared to models trained with mood labels alone. This indicates that the emotion change labels are generated effectively and contribute positively to the mood prediction performance. Our claim is further confirmed by similar trends when performing corresponding experiments employing Δ_{GT} labels.

ETHICAL IMPACT STATEMENT

This study aims at examining mood from a computational perspective by using emotional similarity information. The data for the study reuses publicly available databases, AffectNet and AffWild2, to conduct computational modeling experiments. This study is designed towards answering a theoretical question regarding the interaction between mood and emotion. While facial information is revealed from images and videos in the databases, we neither use identity-specific information, nor base our claims on a specific religion, race or gender. The proposed framework is non-obtrusive, using the images, and videos present in the databases. One of the most crucial applications of this framework is in healthcare, to detect early signs of mood disorders such as depression, and monitoring the mood of the patients by observing their emotional patterns. Other application include education, gaming technology, marketing, etc [4].

Final Author Version

Although we aim at developing robust mood inference technology, as with any other affect recognition system, there could be potential ethical concerns. Mood inference could reveal sensitive information about an individual's mental state, and could be used against the person. Mood detection system could be also used inappropriately to influence or manipulate individuals' behavior or emotions. Additionally, the use of mood detection in contexts such as employment could lead to discrimination or bias. Finally, we acknowledge that there could be intrinsic bias, as we train our models on the databases which may be biased towards facial expressions of individuals from a specific location/culture.

REFERENCES

[1] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," in *30th British Machine Vision Conference 2019, Cardiff, UK*. BMVA Press, 2019, p. 297.

[2] E. Hudlicka and J.-M. Fellous, "Review of computational models of emotion," *Arlington, MA: Psychometrix Associates*, 1996.

[3] D. DeSteno, J. J. Gross, and L. Kubzansky, "Affective science and health: the importance of emotion and emotion regulation." *Health Psychology*, vol. 32, no. 5, p. 474, 2013.

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

[5] K. R. Scherer, "What are emotions? and how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[6] M. Siemer, "Mood-specific effects on appraisal and emotion judgements," *Cognition & Emotion*, vol. 15, no. 4, pp. 453–485, 2001.

[7] P. Panchal, A. Kaltenboeck, and C. J. Harmer, "Cognitive emotional processing across mood disorders," *CNS spectrums*, vol. 24, no. 1, pp. 54–63, 2019.

[8] H. R. Venn, S. Watson, P. Gallagher, and A. H. Young, "Facial expression perception: an objective outcome measure for treatment studies in mood disorders?" *International Journal of Neuropsychopharmacology*, vol. 9, no. 2, pp. 229–245, 2006.

[9] S. Lombion-Pouthier, P. Vandel, S. Nezelof, E. Haffen, and J.-L. Millot, "Odor perception in patients with mood disorders," *Journal of affective disorders*, vol. 90, no. 2-3, pp. 187–191, 2006.

[10] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," *Procedia Computer Science*, vol. 108, pp. 1175–1184, 2017.

[11] M. Bilalpur, S. M. Kia, M. Chawla, T.-S. Chua, and R. Subramanian, "Gender and emotion recognition with implicit user signals," in *ICMI '17: Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 379—-387.

[12] A. Shukla, S. S. Gullapuram, H. Katti, K. Yadati, M. Kankanhalli, and R. Subramanian, "Affect Recognition in Ads with Application to Computational Advertising," in *MM '17: Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1148–1156.

[13] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2018.

[14] R. Parameshwara, I. Radwan, R. Subramanian, and R. Goecke, "Examining Subject-Dependent and Subject-Independent Human Affect Inference from Limited Video Data," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2023, pp. 1–6.

[15] W. N. Morris, "A functional analysis of the role of mood in affective systems," in *Emotion*. Sage Publications, Inc, 1992, pp. 256–293.

[16] L. Sigal, D. J. Fleet, N. F. Troje, and M. Livne, "Human attributes from 3d pose tracking," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III 11*. Springer, 2010, pp. 243–257.

[17] M. Thrasher, M. D. Van der Zwaag, N. Bianchi-Berthouze, and J. H. Westerink, "Mood recognition based on upper body posture and movement features," in *Affective Computing and Intelligent Interaction: 4th International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part I 4*. Springer, 2011, pp. 377–386.

[18] S. Narayana, R. Subramanian, I. Radwan, and R. Goecke, "To improve is to change: Towards improving mood prediction by learning changes in emotion," in *Companion Publication of the 2022 International Conference on Multimodal Interaction*, 2022, pp. 36–41.

[19] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[20] K. R. Scherer, "Toward a dynamic theory of emotion: The component process model of affective states," *Geneva studies in Emotion and Communication*, vol. 1, pp. 1–98, 1987.

[21] P. Ekman, "Expression and the nature of emotion," *Approaches to emotion*, vol. 3, no. 19, p. 344, 1984.

[22] K. Oatley, D. Keltner, and J. M. Jenkins, *Understanding emotions*. Blackwell publishing, 2006.

[23] M. Y. Wong, "The mood-emotion loop," *Philosophical Studies*, vol. 173, pp. 3061–3080, 2016.

[24] J. D. Mayer, Y. N. Gaschke, D. L. Braverman, and T. W. Evans, "Mood-congruent judgment is a general effect." *Journal of personality and social psychology*, vol. 63, no. 1, p. 119, 1992.

[25] P. C. Schmid and M. Schmid Mast, "Mood effects on emotion recognition," *Motivation and Emotion*, vol. 34, pp. 288–292, 2010.

[26] A. Dhall, R. Goecke, S. Lucey, T. Gedeon *et al.*, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 19, no. 3, p. 34, 2012.

[27] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 209–222, 2015.

[28] 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)*. IEEE, 2013, pp. 1–8.

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

[30] C. Katsimerou, J. Albeda, A. Huldtgren, I. Heynderickx, and J. A. Redi, "Crowdsourcing empathetic intelligence: The case of the annotation of emma database for emotion and mood recognition," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 4, pp. 1–27, 2016.

[31] C. Katsimerou, I. Heynderickx, and J. A. Redi, "Predicting mood from punctual emotion annotations on videos," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 179–192, 2015.

[32] 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.

[33] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," *Advances in Neural Information Processing Systems*, vol. 6, 1993.

[34] Z. Lian, Y. Li, J. Tao, and J. Huang, "Speech emotion recognition via contrastive loss under siamese networks," in *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and First Multi-Modal Affective Computing of Large-Scale Multimedia Data*, 2018, pp. 21–26.

[35] M. Sabri and T. Kurita, "Facial expression intensity estimation using siamese and triplet networks," *Neurocomputing*, vol. 313, pp. 143–154, 2018.

[36] 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.

[37] C. Lu, W. Zheng, C. Li, C. Tang, S. Liu, S. Yan, and Y. Zong, "Multiple spatio-temporal feature learning for video-based emotion recognition in the wild," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI2018)*, 2018, pp. 646–652.

[38] D. Kim and B. C. Song, "Emotion-aware Multi-view Contrastive Learning for Facial Emotion Recognition," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*. Springer, 2022, pp. 178–195.

[39] H. Zhou, D. Meng, Y. Zhang, X. Peng, J. Du, K. Wang, and Y. Qiao, "Exploring emotion features and fusion strategies for audio-video emotion recognition," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 562–566.

Final Author Version