Pattern Recognition 130 (2022) 108833

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Visual-to-EEG cross-modal knowledge distillation for continuous emotion recognition

Su Zhang[a], Chuangao Tang[b], Cuntai Guan[a,*]

[a] School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore
[b] Key Laboratory of Child Development and Learning Science (Ministry of Education), School of Biological Science and Medical Engineering, Southeast University, Nanjing, 210096, China

ARTICLE INFO

Article history:
Received 26 August 2021
Revised 3 May 2022
Accepted 3 June 2022
Available online 3 June 2022

Keywords:
Continuous emotion recognition
Knowledge distillation
Cross-modality

ABSTRACT

Visual modality is one of the most dominant modalities for current continuous emotion recognition methods. Compared to which the EEG modality is relatively less sound due to its intrinsic limitation such as subject bias and low spatial resolution. This work attempts to improve the continuous prediction of the EEG modality by using the dark knowledge from the visual modality. The teacher model is built by a cascade convolutional neural network - temporal convolutional network (CNN-TCN) architecture, and the student model is built by TCNs. They are fed by video frames and EEG average band power features, respectively. Two data partitioning schemes are employed, i.e., the trial-level random shuffling (TRS) and the leave-one-subject-out (LOSO). The standalone teacher and student can produce continuous prediction superior to the baseline method, and the employment of the visual-to-EEG cross-modal KD further improves the prediction with statistical significance, i.e., p-value < 0.01 for TRS and p-value < 0.05 for LOSO partitioning. The saliency maps of the trained student model show that the brain areas associated with the active valence state are not located in precise brain areas. Instead, it results from synchronized activity among various brain areas. And the fast beta and gamma waves, with the frequency of $18 - 30Hz$ and $30 - 45Hz$, contribute the most to the human emotion process compared to other bands. The code is available at https://github.com/sucv/Visual_to_EEG_Cross_Modal_KD_for_CER.

1. Introduction

Continuous emotion recognition (CER) is the process of identifying human emotion in a temporally continuous manner. The emotional state, once understood, can be used in various areas including entertainment, e-healthcare, recommender system, and e-learning. To describe the human state of feeling, psychologists have developed the categorical and the dimensional models. The categorical model aims to obtain a discrete estimate of emotional category. It features simplicity and universality and has been extensively exploited in affective computing. The dimensional model, on the other hand, aims to obtain a continuous estimate in a dimensional space. It can describe more complex and subtle emotions. This paper focuses on developing a CER method based on the dimensional model.

CER can utilize information from various modalities. The visual modality, usually featured by facial expressions [1,2], is one of the most dominant modalities for emotion recognition. By utilizing either a finely hand-crafted descriptor, e.g., facial action coding system (FACS) [3], or a powerful convolutional neural network, e.g., the Resnet for feature extraction, an emotion recognition method can achieve promising results. In recent years, Electroencephalography (EEG) has drawn considerable attention from researchers [4], due to its simple, cheap, portable, and easy-to-use solution for identifying emotions [5]. In addition to the visual and EEG information, the audio/speech, text, and some other physiological signals (e.g., heart rate, blood pressure, and eye gaze) are also widely used.

Two general differences between the visual and EEG modalities are of the most relevance to our interest. First, facial expressions and gestures are overt and determined, whereas the EEG signal is covert and highly subject-dependent. As a result, it is feasible to directly label the emotion based on the visual modality from an annotator, yet for EEG modality it is done either by predefined experiment protocol or by subjects themselves. Second, the visual modality usually has high and low resolutions on spatial and temporal dimensions (e.g., $40 \times 40 \times 3$ and 30 fps, respectively, whereas the EEG modality is high on temporal resolutions (e.g., 256 Hz) yet low on spatial resolutions (e.g., 32 electrodes). The greater the resolution is, the more detailed structural or phase

* Corresponding author.
E-mail addresses: sorazcn@gmail.com (S. Zhang), 230169620@seu.edu.cn (C. Tang), ctguan@ntu.edu.sg (C. Guan).

https://doi.org/10.1016/j.patcog.2022.108833
0031-3203/

changes in response to emotional stimuli can be studied. Based on the differences of modalities and the assumption that incorporating multimodal data will produce results that are superior to unimodal data, it is natural to utilize the multimodal data which can essentially increase the amount of available data and hopefully attenuate the defects of each modality.

Knowledge distillation (KD) is one of the promising solutions to combining multimodal data. In deep learning, KD is an effective technique that has been widely used to transfer information from one network to another network whilst training constructively [6]. Many cross-modal KD methods have been proposed to leverage the synchronization of visual and audio information in the video data. A joint embedding can be learned by distilling the knowledge between RGB/depth, face/voice, and CT/MRI images. However, to the best of the authors' knowledge, there is no prior work relevant to visual-to-EEG cross-modal KD on CER.

We, therefore, pose a question: Can the CER performance of the EEG modality be improved if we transfer the knowledge from the visual modality? Given a dataset containing synchronous facial videos and EEG signals of different subjects, the facial video modality tends to have stronger relevance with respect to the expert-labeled continuous trace. The reasons are two-fold. First, the experts conduct the labeling according to the subject's facial expression. Second, the EEG signal has a low information-to-noise ratio, large bias can be existing among signals recorded at a different time or from different subjects. It inspires us to teach the EEG modality using the visual knowledge.

In this work, we explore to what extent can the EEG modality gain from the visual modality using the cross-modal knowledge distillation (CKD) for CER. A teacher model is firstly trained in the visual modality using the facial video. Its intermediate features, a.k.a. dark knowledge, are then used to supervise the student model training in an offline manner. Specifically, the teacher and student models comprise a cascade spatiotemporal and a single temporal network, respectively. The inputs to the teacher and student are facial video frames and the synchronous EEG average band power. The temporal embeddings from the trained teacher's temporal component are taken as dark knowledge. During the training of the student, its temporal embeddings are guided by the dark knowledge using L1 loss. Together with the concordance correlation coefficient (CCC) loss, which punishes the inconsistency between the prediction and label sequences by scaling the correlation coefficient with their mean square difference, the student is able to learn from the visual and EEG modalities simultaneously. (A formal definition of CCC is provided by Eq. (3)). During the test of the student, it infers based on the EEG modality and the learned visual knowledge. Results from experiments manifest statistical significance (p-value < 0.01 or 0.05 depending on the data partitioning scheme) on root mean square error (RMSE), Pearson correlation coefficient (PCC), and CCC, comparing to its counterpart without KD.

The remainder of the paper is structured as follows. Section 2 discusses the related works on CKD and CER. Section 3 elaborates the pre-requisite knowledge regarding the brain, emotion, and EEG. Section 4 details the proposed deep neural networks and the CKD. Section 5 elaborates the complete pipeline, i.e., the data preprocessing and model training. Section 6 reports and analyzes the experiment results. Section 7 concludes the paper.

2. Related works

2.1. Cross-modal knowledge distillation (CKD)

CKD uses the teacher's representation as a supervision signal to train the student to learn another task [6]. It is helpful especially when the data or labels for the target modalities are hard to get. Based on the hypothesis that the emotional content of speech correlates with the facial muscular movement and facial expression of the speaker, Afouras et al. [7] transfer voice knowledge to train lip reading-based visual speech recognition models, while Nagrani et al. [8] transfer the visual knowledge to learn voice feature-based speech classification, both of which are without access to any form of human-labeled ground truth. Hoffman et al. [9] utilize the RGB information to teach a depth network, and fuse the information across modalities. Gupta et al. [10] learn a student model on unlabeled depth images and optical flow by transferring the knowledge of a teacher model trained on well-annotated RGB images. Zhao et al. [11] use radio data to guide human pose estimation on occluded images. Thoker and Gall [12] employ paired RGB videos and skeleton sequences for CKD. The knowledge learned on RGB videos is transferred to the student model for skeleton-based human action recognition. Garcia et al. [13] use additional depth images to generate a hallucination stream for RGB image modality and thereby improve the action recognition performance. Tian et al. [14] employ a contrastive loss to transfer relation-based knowledge across modalities. Roheda et al. [15] use generative adversarial networks (GAN) for distillation among the missing and available modalities. We see that most of the CKD methods are for target detection and action recognition. It is rarely explored in the area of CER.

2.2. CER Methods

The term "continuous" possesses two characteristics in our context. Spatially, it aims to place the emotional state as a continuous-valued point in the multi-dimensional space of the dimensional theory, instead of choosing categorical labels. Temporally, it continuously predicts the emotional state for a fixed time interval, constituting the emotional trace of the subject over a specified time span.

The CER has always been challenging due to the following causes. First, the emotion itself is highly subjective and subject-dependent. For example, the perception of emotion is influenced by individual experiences. Physically abused children are much quicker than other children to spot the signals of anger [16]. As a result, the data from the subjects and the ground truth from the annotators are prone to personal bias. Multimodality and Transfer learning among visual, audio, and physiological data are two promising techniques to alleviate this issue and develop reliable CER models. Second, by taking the facial muscular movement as actions, the complex emotion cues over a large time span are a composition of complex one-actions [17]. Typical one-actions can be defined by FACS [18] that codes the movements of individual facial muscles. However, as atomic as the FACS may be, human emotion, no matter from which modality it is observed, usually exhibits large variations in terms of intensity and order in their duration, and takes longer to unfold. Models which can learn the long-range temporal dependencies are in need to counter this issue.

Soleymani et al. [19] propose a multimodal method for continuous valence prediction based on facial landmark sequence and EEG signal. A long short-term memory (LSTM) network is used for feature learning. The features from the two modalities are fused using feature-level and decision-level fusion schemes before feeding to the fully-connected layers. Somandepalli et al. [20] propose a linear dynamical system method with a late fusion method. It models unimodal predictions as observations in a Kalman filter formulation. By leveraging the inter-correlations between arousal and valence, the predicted arousal is taken as an additional feature to improve valence predictions. Han et al. [21] propose strength modeling with two models being concatenated in a hierarchical framework. The strength information of the first model is joined with

the original features. It expands the feature space of the input for the successive model. By characterizing the perceived emotion as time-invariant responses to salient events, Wataraka Gamage et al. [22] model arousal and valence variation as the output of a parallel array of time-invariant filters, with each filter representing a salient event in the context. Chen et al. [23] combine a pretrained 2D-CNN and a TCN to learn deep spatiotemporal features from video frames and audio spectrograms, and use a spatiotemporal graph convolutional network to encode facial landmarks graph. Finally, a bidirectional LSTM network is employed for unimodal and multimodal predictions. Zhao et al. [24] employ adversarial domain adaption to overcome the domain shift caused by cultural differences. Typically, a person from an individualist culture tends to express higher arousal emotions than that from a collectivist culture. Given the culture-specific training and testing data, the proposed method achieves the generalization by using several interaction strategies for adversarial training among the visual, audio, and textual features. Deng et al. [25] address the issue of missing labels in multi-task learning by using the output of a teacher model as the soft labels. The latter and the ground truth are then used to train a student model.

We see that most CER methods are based on visual and audio modalities. To the best of our knowledge, out of all the publicly available CER databases, there is only a subset [19] of the MAHNOB-HCI database [26] where the facial video, EEG signal, and continuous valence label are available. Readers interested in the comprehensive review on CER databases can refer to [27]. Our work is based on this subset of the MAHNOB-HCI database.

Soleymani et al. [19] and Chen et al. [23] are the most relevant methods to ours. The differences are explained as follows. First, concerning the motivation, our work intends to investigate the CKD on visual and EEG modalities, while the two papers are for multimodal feature fusion. In the case where the visual information is not available, our model can still work and infer based on the EEG signal and the learned visual knowledge. Second, our visual model comprises a cascade 2DCNN-TCN architecture. The produced spatiotemporal features are directly fed to a linear layer to infer. Whereas in [23], the 2DCNN-TCN is first trained as a feature extractor. An independent bidirectional LSTM network is then trained on top of the extracted features to infer. Third, our EEG model use TCN to learn the temporal encoding of the EEG band power, while in [19] an LSTM network is used for the same purpose.

3. Physiological grounding

The section aims to provide a decent physiological foundation of physiology and neuroscience related to emotion, with which the reader can have a better understanding of our methodology and experimentation.

Our brain consists of the left and right hemispheres, each of which can be further divided into four lobes, as shown in Fig. 1. The frontal lobe lies behind the forehead. It involves speaking, muscle movements, judgment, and plan making. The parietal lobe lies at the top of the head toward the rear. It mainly processes the sensory input for touch and body position, and also integrates various sensory information. The occipital lobe which is at the back of the head processes visual information. And, the temporal lobe lying roughly above the ears processes auditory information from the opposite ear. About 25% of these four cortical areas either receive sensory input or direct muscular output, and the rest 75%, which are called association areas, are involved in higher mental functions (e.g., thinking, speaking, and learning) and make us human. These complex functions are not located in precise brain areas, but the result from synchronized activities of many [28].

Emotion is one of such complex functions. Given our context where a subject is watching short film clips during the data acquisition of a CER database, one possible pathway associated with emotion can be as follows. The visual and audio stimuli first go through the occipital and temporal lobes, the sensory information is then delivered to the parietal lobe for integration. After which the integrated information is delivered to the front lobe. In there, a variety of judgments and regulations are made. And finally, the frontal lobe directs the subject's facial muscular movement. The latter is then annotated by the expert, where a similar pathway would repeat again in the latter's brain. Two emotion theories are used to model human emotion. The basic theory labels emotions discretely as several categories. It holds that the basic emotions, i.e., happiness, anger, fear, sadness, disgust, and surprise, are the foundation of human emotion. Other emotions such as satisfaction, fatigue, and confusion are compounds of them. The dimensional theory models human emotion using a multi-dimensional space, within which each dimension is a perspective of emotion, such as valence, arousal, and dominance.

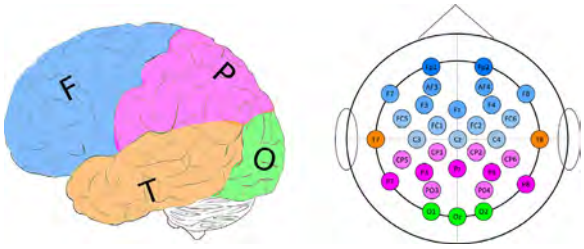

Fig. 1. The illustration of the four brain lobes (left) and the placement of the 32 EEG electrodes (right). Our brain consists of four lobes, i.e., the frontal (F), parietal (P), temporal (T), and occipital (O) lobes. By placing scalp electrodes on specific locations following the 10–20 system of electrode placement, the potential fluctuations of the underlying cerebral regions can be measured. The correspondences between the four lobes and 32 electrodes are indicated in color.

The EEG signals, which are potential fluctuations produced by the central nervous system, provide promising information to decode the emotion process. By placing scalp electrodes on specific locations following the 10–20 system of electrode placement, the potential fluctuations of the underlying cerebral regions can be measured, as shown in Fig. 1. As a direct reflection of brain activity, EEG can be divided into five frequency bands, each corresponding to different mental states. The δ wave $(0.3-5Hz)$ is associated with the unconscious mind. It appears when one is anesthetized or in a dreamless sleep. The θ wave $(5-8Hz)$ is associated with the subconscious mind and memory load. It appears when one is sleeping and dreaming, during which the working memory is being encoded to form the long-term memory. The α wave $(8-12Hz)$ is associated with a relaxed yet aware mental state. It can be reduced or disappeared when one is under external visual or auditory stimuli. β wave $(12-30Hz)$ is associated with an active state of mind. It can be observed when one is carrying out an intense focused mental activity, and is more obvious in the frontal lobe. Compared to the "fast idle" β_1 wave $(12-18Hz)$, the β_2 wave $(18-30Hz)$ is associated with complex thought, integrating new experiences, high anxiety, or excitement. The γ wave $(>30Hz)$ is associated with high-level cognitive brain activities or attention-intensive activities such as the perception, transmission, processing, integration, and feedback of information. It is also found in the process of multi-modal sensory processing. (Note that the actual γ band we consider in our work is within $30-45Hz$, as the higher band cannot be effectively measured using current EEG technology due to muscle contamination. Also, slight differences regarding the band partitioning may appear in relevant works.) The work by Solemani et al. [19] investigates the effect of EEG features on estimating valence given facial expressions and eye movements. The ANOVA test shows that the EEG band power adds information which is independent to the visual modality for valence prediction. It manifests

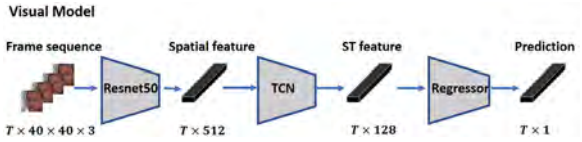

Fig. 2. The illustration of the visual model. The model takes T video frames sized at $40 \times 40 \times 3$ as the input. The Resnet50 plays the role of backbone and yields the per-frame spatial features. The latter is then fed to TCN producing the spatiotemporal features. And finally, the regressor maps each feature point onto the 1-D space. ST: spatiotemporal.

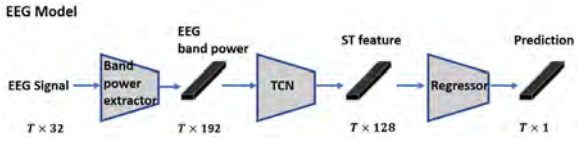

Fig. 3. The illustration of the EEG model. The model takes raw EEG signal as the input, from which the EEG band power features are extracted. The latter is then fed to TCN producing the spatiotemporal features. And finally, the regressor maps each feature point onto the 1-D space. ST: spatiotemporal.

the complementary nature between the visual and EEG modality, and further inspires us to explore the idea of using visual to teach EEG for CER.

4. Method

Our method seeks to i) train a teacher model in the visual modality, and then 2) use the knowledge from the teacher and the labels to supervise the student model in the EEG modality. In this section, we will first detail the models for the visual and EEG modalities, followed by explaining the interaction between the teacher and student.

4.1. Teacher: The visual model

The goal of the visual model is to predict the emotional state given the video frames as the inputs. Generally, there are two fundamental frameworks of neural networks for visual-based emotion recognition: (i) the cascade spatiotemporal architecture and (ii) the standalone architecture. Type (i) usually contains a CNN to extract spatial information, from which the temporal information is obtained by using temporal models such as Time-delay, recurrent neural networks (RNN), long short-term memory networks (LSTM), or TCN. Type (ii) combines the two separated steps into one and extracts the spatiotemporal feature using a unified model like the 3D-CNNs.

We choose Type (i) due to the following facts. First, a 3D-CNN model [29] usually has considerably more parameters than 2D-CNNs due to the extra kernel dimension, and therefore requires more data and longer time to train. However, 3D-based emotion recognition databases [30] are typically based on posed behavior with a few subjects, little diversity, and limited continuous labels. By contrast, there are a large amount of 2D-based facial image or emotion databases, such as MS-CELEB-1M [31], VGGFace2 [32], which are more diverse and determinant. Though there are abundant 3D video understanding databases that might be available for self-supervised or semi-supervised pretraining of a potential 3D-CNN-based emotion recognition model, the techniques involved are still a hot research topic. Second, 3D-CNNs alone may not be suitable to capture long-range temporal dependencies. As we mentioned before, CER requires to map a composition of complex one-actions with varied intensity and order to a sequence of continuous labels. However, most 3D-CNN-based networks are designed for at most 128 time steps [33], whereas an exclusive temporal model can easily exceed this limit.

Our visual model is illustrated in Fig. 2. It consists of a pretrained Resnet50, a TCN, and a regressor (i.e., a fully connected layer). Fed by T consecutive video frames, the Resnet50 produces T 512-D spatial features. The latter is then fed to the TCN producing T 128-D spatiotemporal features. Finally, the regressor maps the features onto the 1-D.

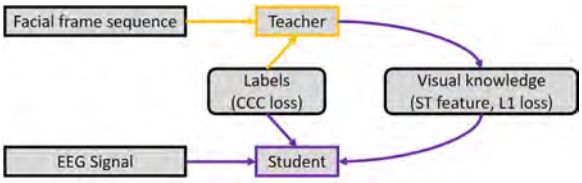

Fig. 4. The illustration of the teacher-student interaction. ST feature denotes the spatiotemporal features. The training of the teacher and student models are colored in yellow and purple, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.2. Student: The EEG model

The goal of the EEG model is to predict the emotional state given the EEG signal as the input. It consists of a band power extractor, a TCN, and a regressor. The EEG signal which is synchronous to the video frame sequence is firstly used for band power calculation. Six bands, i.e., the δ, θ, α, β_1, β_2, γ we introduced in Section 3, are chosen on top of the 32 channels, leading to T 192-D EEG band power features. The latter are then fed to a TCN and a regressor is used to infer. Our EEG model is illustrated in Fig. 3.

The average band power computes a single scalar that summarizes the contribution of a given frequency band to the overall power of the signal. Given a windowed EEG discrete signal $X(n)$ with N samples from one EEG electrode in the time domain, the fast Fourier transform (FFT) returns N complex number whose real and imaginary parts represent the amplitude and phase of the signal in the frequency domain. The magnitude-squared of the FFT can be used to obtain an estimate of the power spectral density

$$S_X(f) = \frac{1}{N}\left|\sum_{n=1}^{N} X(n)e^{-i2\pi f n\Delta t}\Delta t\right|^2 \tag{1}$$

at f, based upon which the average band power in the frequency band $[f_1, f_2]$ is defined as

$$P_{[f_1,f_2]} = \int_{f_1}^{f_2} S_X(f)df. \tag{2}$$

In our work, $S_X(f)$ is obtained using Welch's method from the *scipy* library.

4.3. Visual-to-EEG KD

The goal of the visual-to-EEG KD is to use the visual knowledge (i.e., the spatiotemporal visual features produced by the TCN of the visual model) alone with the labels to train an improved EEG model. The interaction between the teacher and student is illustrated in Fig. 4.

Two stages are involved in the teacher-student interaction. In the first stage, the teacher model is trained by minimizing the CCC loss function between the frame sequences and the corresponding labels. In the second stage, the trained teacher model is used to extract the spatiotemporal features of the visual modality, and

the student model is then trained using the EEG Signal, the corresponding labels, and visual spatiotemporal features.

4.4. Loss function

Two types of loss functions are employed. The CCC loss is used for teacher training, and the weighted sum of CCC and L1 loss is used for student training.

CCC has been widely used to evaluate the performance of CER methods. Given two time sequences, CCC evaluates the agreement in-between by scaling the correlation coefficient with their mean square difference. As a result, a prediction that is well correlated with the labels can still be penalized in proportion to the deviation, should the two have a shift in the mean. CCC is valued in $[-1, +1]$, a higher CCC indicates greater concordance. Given two sequences $\mathbf{U} \in \mathbb{R}^{T\times 1}$ and $\mathbf{V} \in \mathbb{R}^{T\times 1}$, their CCC is defined as

$$\rho_c(\mathbf{U}, \mathbf{V}) = \frac{2\sigma_{UV}}{\sigma_U^2 + \sigma_V^2 + (\mu_U - \mu_V)}, \tag{3}$$

where σ_U^2 and σ_V^2 are the variances, σ_{UV} is the covariance of the two sequences, and μ_U and μ_V are the means. The CCC loss is defined by converting CCC to the dissimilarity measure $1 - \rho_c(\mathbf{U}, \mathbf{V})$ for model training.

Inspired by Romero et al. [34] which distills the knowledge by enforcing the proximity of intermediate feature maps using the L2 loss, we further use the sparser L1 loss as the KD loss, in order to produce a more reasonable magnitude relevant to the CCC loss and makes the training more controllable. Given two feature vectors $\mathbf{U} \in \mathbb{R}^{T\times F}$ and $V \in \mathbb{R}^{T\times F}$, the L1 loss is defined as

$$L_1(\mathbf{U}, \mathbf{V}) = \frac{1}{TF}\sum_{i=1}^{T} |u_i - v_i| \tag{4}$$

where $u_i \in \mathbb{R}^F$ and $v_i \in \mathbb{R}^F$ are the feature points in each time step.

The weighted sum of CCC and L1 loss for student training is defined as

$$\ell(\mathbf{X}, \mathbf{Y}, \mathbf{V}_t, \mathbf{V}_s) = 1 - \rho_c(\mathbf{X}, \mathbf{Y}) + w \cdot L_1(\mathbf{V}_t, \mathbf{V}_s) \tag{5}$$

where $\mathbf{X}$ and $\mathbf{Y}$ denote the predictions and the labels, and $\mathbf{V}_t$ and $\mathbf{V}_s$ denote the spatiotemporal features of the teacher and student model, respectively, with the constant w being the trade-off. The grid searching is employed to find the optimal w.

5. Implementation detail

5.1. Database

MAHNOB-HCI is a multimodal database recorded in response to affective stimuli with the goal of emotion recognition and implicit tagging research [26]. It provides the synchronized recording of facial videos, audio signals, eye gaze data, EEG signals, and other physiological signals from 30 subjects. The subjects are asked to watch 20 emotional video clips, resulting in 440 trials. The video clips are between 35 and 117 s long. The EEG signals are acquired from 32 electrodes on the $10-20$ international system. The sampling frequency is 256 Hz. The facial videos are captured at 60 fps and 780×580 resolution. For each trial, four integers ranging from 1 to 9 and self-reported by the subjects are used to label the valence, arousal, dominance, and emotional keywords, respectively.

A subset [19] of the original MAHNOB-HCI database is chosen to be continuously labeled. It contains 239 trials from 24 subjects with obvious facial expressions. The trial number for each subject is not even. Five experts are employed for the annotation using FEELTRACE and a joystick. Only the valence is continuously labeled. The reason is that the subjects are quiet and passively watching videos, which makes the annotation of arousal, power, or expectation unavailable [19]. The continuous valence label is determined by the average of the five experts' labels. Our work is based on this subset.

5.2. Data preprocessing

5.2.1. Facial video

Given the facial video of a trial, it contains the facial expression of the subject during the stimuli watching and self-reporting. The latter is excluded by trimming the facial video according to the time stamp information. The video is then changed to 64 fps for more convenient synchronization with the continuous valence label which is at 4 fps, i.e., every 16 consecutive frames correspond to 1 valence label point. Finally, the video frames are resized to $48 \times 48 \times 3$.

5.2.2. EEG Signal

Given the EEG signal of a trial, the first and last 30s of the recording which do not correspond to stimuli watching are excluded according to the database manual.[1] The signals from the 32 electrodes are then re-referenced to the average reference to enhance the signal-to-noise ratio. The default API *set_eeg_reference* from MNE toolkit[2] is used for the average reference. After which, the average band power on the six bands is calculated. The physiological motivation for the six-band division is elaborate in Section 6.3. The window size and hop size for band power calculation are 2s and 0.25s, respectively. The resulted $6 \times 32 = 192$-D band power features at the frequency of $4Hz$ are therefore synchronized with the continuous valence labels. Note that the EEG preprocessing was carried out following the baseline method [19] which employed only the average reference and band-pass filtering. We did not employ other techniques to deal with the artifacts caused by motion and respiratory. Theoretically, the delta band $(0.3 - 5Hz)$ could contain such artifacts. The visualization using saliency maps, as shown in Fig. 6, manifests small active regions in the frontal lobe of several subjects, which are possibly caused by eye blinking or rolling.

5.3. Data partitioning

Two data partitioning schemes are used: (i) trial-level random shuffling (TRS, 10-fold) [19] and (ii) leave-one-subject-out (LOSO, 24-fold). TRS focuses on the trial-level and overlooks from which subject the trial comes. It first randomly shuffles the 239 trials, and then splits the 239 trials into 129, 86, and 24 trials for training, validation, and test, so that the test set contains 10% of the data and the training and validation sets contain the 60% and 40% of the remaining data. LOSO focuses on the subject-level. For the ith fold, trials from the ith subject are taken as the test set. All the trials from the remaining 23 subjects are randomly shuffled, with 80% and 20% being the training and validation sets, respectively.

TRS may lead to data leakage. The random shuffling would split the data from the same subject to training, validation, and test sets. Compared to the data from different subjects, the data from the same subject has greater consistency. The model trained in this manner has actually seen the test data to some extent and would inflate the test performance. TRS is widely used in fields like computer vision and natural language processing, where the data usually are vastly greater in diversity and therefore invulnerable to the overfitting problem. However, the negative influence becomes nontrivial for fields with limited training data. In AI-based emotion recognition, both the TRS and LOSO are widely used. In our experiment, we choose to employ both schemes and objectively report the results.

[1] https://mahnob-db.eu/hci-tagging/media/uploads/manual.pdf.
[2] https://mne.tools/stable/index.html.

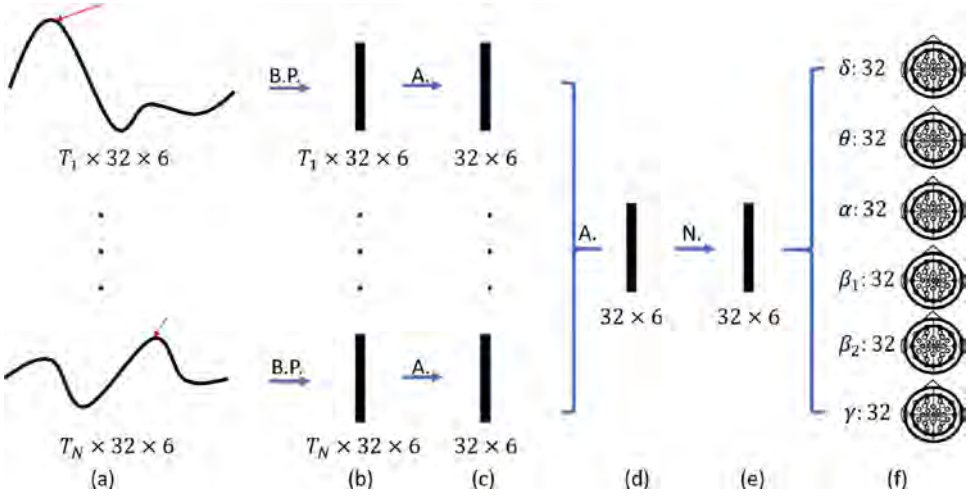

Fig. 5. The illustration of generating the peak response mapping for interpretability [42] investigation. Given the trained EEG model for the ith subject, (a) N $T_j \times 32 \times 6$ valence predictions for N trials are obtained. By selectively back-propagate the peaks, (b) N $T_j \times 32 \times 6$ gradient vectors for the N trials are obtained. By averaging on the temporal dimension, (c) N 32×6 gradient vectors are obtained. After which, the average over the trial dimension is conducted producing (d) the 32×6 gradient vector of the ith subject. (e) The normalized version of the latter is finally used to plot (f) the heatmap on the six bands using the MNE toolkit. B.P.: backward propagation. A.: average. N.: normalization. The red arrow points to the peak value for the backpropagation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5.4. Feature synchronization

Recall that after the data preprocessing, we have facial frames, EEG band power features, and continuous valence labels at the frequency of 64Hz, 4Hz, and 4Hz, respectively. In order to synchronize the facial frames with the other two features, downsampling is employed. A consecutive 16 frames are taken as one group, corresponding to one valence label point. During the teacher training, the nth frame for each group is loaded in sequential and fed to the teacher model. The integer n is randomly chosen from 0 to 15 for each epoch. For the inference, only the 0th frames of each group are loaded.

The visual knowledge is generated by feeding the facial frames without downsampling to the trained teacher model. The generated visual knowledge is at the frequency of 64Hz. During the student training, the same downsampling scheme used on the video frames is applied to the visual knowledge for the synchronization with the EEG band power features and continuous valence labels.

5.5. Model training

The teacher model is trained as follows. A Resnet50 is used as the visual backbone. It is pre-trained on the MS-CELEB-1M dataset[3] [31] as a facial recognition task, it is then fine-tuned on the FER+ [35] dataset as a facial expression recognition task.

The training settings of our teacher model are summarized in Table 1. To fine-tune the teacher model on the MAHNOB-HCI dataset, two groups (i.e., the output layer and the whole layer4, according to the Pytorch official implementation) of the Resnet50 backbone are selected. The backbone is initially frozen. When the minimum learning rate is reached, unfreeze one group (starting from the output layer) and reset the scheduler. At the end of each epoch, the best model parameters are loaded. The training would stop if i) there is no remaining backbone layer group, ii) the early stopping counter reaches 20, or iii) the epoch reaches 30.

The training of our student model is much the same as the teacher training except the following. First, since no images are involved, data augmentation and normalization are not employed. Second, the maximal epoch number and early stopping counter are both set to 15 to prevent gradient explosion. Finally, since the student model does not contain a Resnet backbone, it is no need to reset the scheduler.

[3] https://github.com/TreB1eN/InsightFace_Pytorch#2-pretrained-models-performance.

6. Experiment result

The experiment is conducted in three stages. The first stage examines that the standalone teacher and student models can produce results no worse than the baseline method on the valence regression task. The second stage investigates to what extent can the EEG modality be improved by the visual modality under different w from Eq. (5). The last stage explores the contributions of the band frequencies and brain regions towards the emotion process.

The best model from Soleymani et al. [19], i.e., a two layers LSTM network is adopted as the baseline for the valence regression task. To make a fair comparison, the baseline model is implemented and incorporated into our pipeline. The results for the valence regression and CKD experiments are obtained in two steps. First, the model outputs for all the trials of a partition are concatenated along the temporal dimension. Recall that our resampling windows have 66.7% overlap. A direct concatenation is not welcomed as it would produce an over-lengthy prediction vector and further inflates the metrics. Instead, the concatenation is done by placing each output segment according to their windowing indexes. The obtained prediction vector is therefore temporally restored to the original form, which is N-to-N corresponded to the labels. The mean values are taken for the overlapped steps. Second, the RMSE, PCC, and CCC are calculated based on the concatenated prediction vectors and the continuous labels. The results are averaged over the N-fold. Specifically, for the TRS and LOSO partitioning, we have 10 and 24 groups of evaluation results, and the final results are the average across the groups, respectively. This evaluation protocol has been widely used in many CER contests [36–40]. In addition, to obtain the p-value for the CKD experiments, the one-tailed paired t-test is conducted. For example, for the 10-fold TRS partitioning, results of the ith fold from the student with and without CKD are paired. The t-test is conducted using the 10-pair results for the three metrics.

The three metrics are RMSE, PCC, and CCC. Given the prediction $\mathbf{X} \in \mathbb{R}^{T_N \times 1}$ and the continuous label $\mathbf{Y} \in \mathbb{R}^{T_N \times 1}$, where the constant

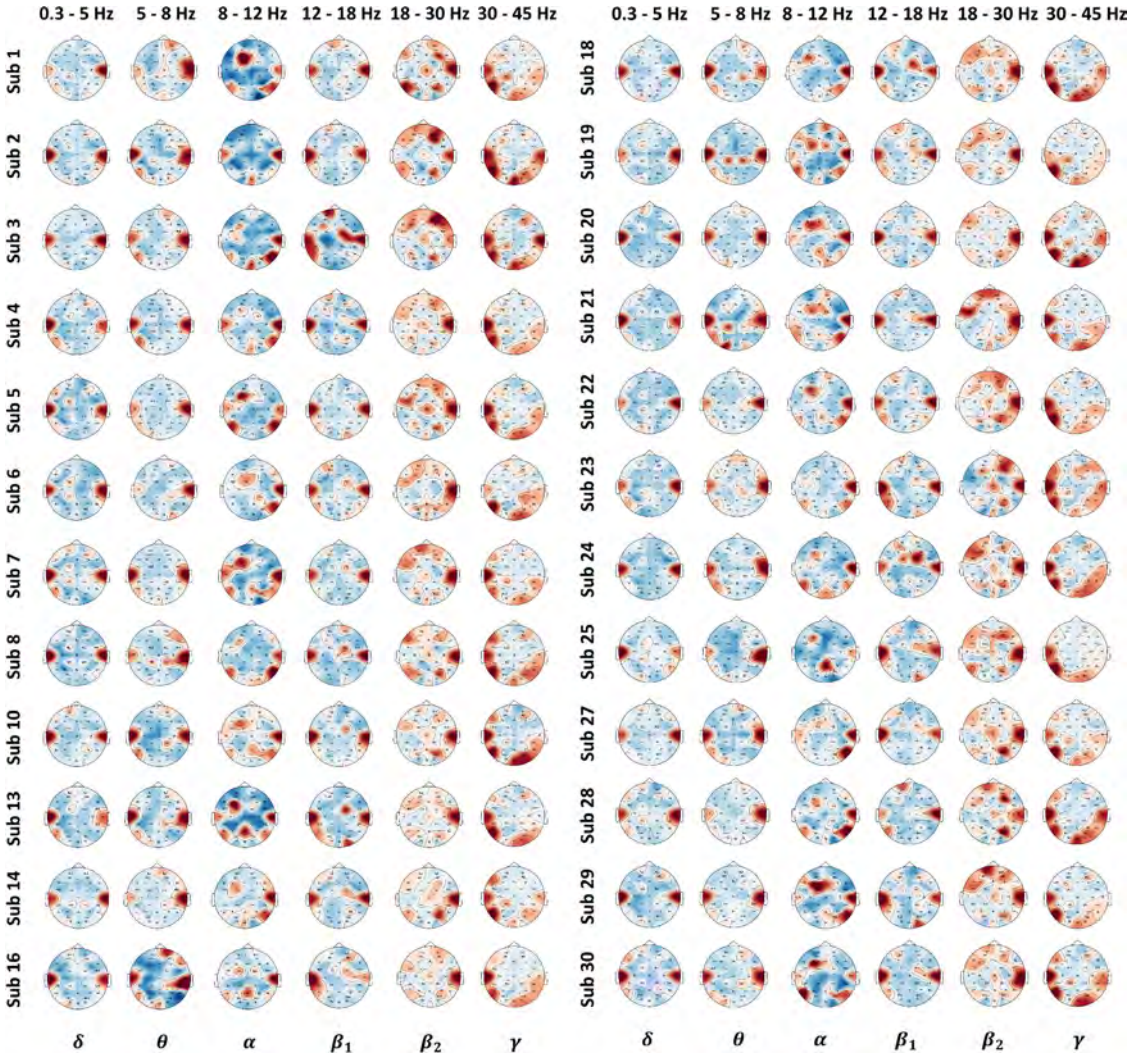

Fig. 6. The topographic saliency maps for the 24 subjects. The gradients of the EEG band power over the 32 electrodes are calculated following the procedure shown in Fig. 5. The warmer color means a higher gradient. A region having warmer color implies that it contributes more to the valence prediction. Therefore, the red regions tend to be more informative for the neural networks to infer the valence compared to the blue counterparts. Sub: subject. The subject numbering is determined by the MAHNOB-HCI database [26]. The missing subjects are not included in the subset [19] since they are not continuously labeled in valence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
The training settings for the teacher model. The Adam optimizer and ReduceLROnPlateau are from the PyTorch library.

Optimizer		Scheduler	
Adam with the CCC loss		ReduceLROnPlateau	
Learning rate	$1e-5$	Patience	5
Weight decay	$1e-4$	Factor	0.5
		Others	
Maximum epoch	30	batch size	2
Early stopping counter	20	Window length (s)	24, equal to 96 data points
Minimum learning rate	$1e-6$	Hop length (s)	8, equal to 32 data points
		Random flip (0.5) + random crop (40) for training	
		Only center crop (40) for validation	
		Normalization of video frames: mean = std = 0.5	

T_N denotes the length sum of all the trials from a partition, the RMSE is formulated as:

$$\epsilon(\mathbf{X},\mathbf{Y}) = \|\frac{\mathbf{X}-\mathbf{Y}}{T_N}\|, \tag{6}$$

and the PCC is formulated as:

$$r(\mathbf{X},\mathbf{Y}) = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}. \tag{7}$$

The CCC is formulated by Eq. (3), and by substituting r we can reformulate it as a factor of PCC:

$$\rho_c(\mathbf{X},\mathbf{Y}) = \frac{2r(\mathbf{X},\mathbf{Y})\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2+(\mu_X-\mu_Y)}. \tag{8}$$

It is worth noting that the strictness of the three metrics is varied. RMSE takes only the overall mean difference into account. A straight vector, once placed soundly, can have a very low RMSE to-

Table 2
The result of our visual model against the baseline using the TRS and LOSO data partitioning. The mean and standard deviation are reported. TRS: trial-wise random shuffling. LOSO: leave-one-subject-out. ↑: the higher the better. ↓: the lower the better. Bold fonts indicate the best results.

Visual modality	Ours with TRS		Soleymani et al. with TRS		Ours with LOSO		Soleymani et al. with LOSO	
	Validation	Test	Validation	Test	Validation	Test	Validation	Test
RMSE ↓	0.054±0.006	0.054±0.006	0.060±0.006	0.058±0.009	0.053±0.005	0.049±0.015	0.060±0.007	0.055±0.020
PCC ↑	0.697±0.054	0.686±0.079	0.623±0.079	0.611±0.150	0.699±0.068	0.684±0.203	0.611±0.105	0.602±0.264
CCC ↑	0.690±0.057	0.674±0.085	0.606±0.086	0.589±0.163	0.695±0.068	0.602±0.228	0.596±0.107	0.533±0.251

Table 3
The result of our EEG model against the baseline using the TRS and LOSO data partitioning. The mean and standard deviation are reported. TRS: trial-wise random shuffling. LOSO: leave-one-subject-out. ↑: the higher the better. ↓: the lower the better. Bold fonts indicate the best results.

EEG modality	Ours with TRS		Soleymani et al. with TRS		Ours with LOSO		Soleymani et al. with LOSO	
	Validation	Test	Validation	Test	Validation	Test	Validation	Test
RMSE ↓	0.067±0.005	0.066±0.009	0.083±0.006	0.080±0.012	0.068±0.007	0.066±0.025	0.087±0.020	0.081±0.034
PCC ↑	0.463±0.103	0.435±0.205	0.347±0.091	0.353±0.228	0.467±0.116	0.474±0.267	0.360±0.147	0.427±0.267
CCC ↑	0.444±0.109	0.415±0.201	0.331±0.092	0.333±0.214	0.445±0.118	0.377±0.250	0.348±0.129	0.306±0.257

wards a twisting vector. PCC and CCC evaluate the linearity and agreement. Intuitively, PCC looks into the "shape" of the two given vectors, and discards the "distance" in-between. And both the shape and distance are evaluated by CCC since it contains an extra term of mean difference in the denominator. Therefore, CCC can be taken as the most strict metric out of the three. Also note that during the training, the sequence length for CCC computing is determined by the resampling window length of a batch. And during the evaluation, it is determined by the length sum of all the trials in a partition.

6.1. Valence regression result

The result of our visual model against the baseline using the TRS and LOSO data partitioning are reported in Table 2. For the TRS partitioning, it can be observed that results from the test set are more consistent with those from the validation set. Only a slight drop on PCC and CCC when it goes from validation to test on both of the two methods. As we mentioned before, the reasons are two-fold: i) the visual modality is highly determined, and ii) data from one subject are already seen by the model during the training stage. For the LOSO partitioning, the gap between the validation and test results is relatively larger. Up to 13.38% and 10.57% drop on CCC can be observed from ours and the baseline, respectively. Overall, our method produces superior results to that from the baseline methods on RMSE, PCC, and CCC.

The result of our EEG model against the baseline using the TRS and LOSO data partitioning are reported in Table 3. For the TRS partitioning, consistent results between the validation and test are also observed for the two methods. For the LOSO partitioning, up to 15.28% and 12.07% drop on CCC are observed from ours and the baseline, respectively, which is larger than their visual counterpart. In this modality, our EEG model also produces better results against the baseline in terms of RMSE, PCC, and CCC.

6.2. Knowledge distillation result

Based on Eq. (5), the grid search is employed for w ranging from 0.2 to 2.0, with a step of 0.2. All the other settings remain the same.

The results of the student with and without CKD using TRS and LOSO partitioning are reported in Tables 4 and 5, respectively. In the interval of $0.2 \leq w \leq 1.0$, the best validation results are found when $w = 1.0$ and $w = 0.4$ for TRS and LOSO, respectively. They also lead to the best test results with statistical significance (p-value ≤ 0.01 and p-value < 0.05 on the three metrics for TRS and LOSO partitioning, respectively). When $1.0 < w \leq 2.0$, though better validation results are yield for LOSO, the corresponding test results are without statistical significance.

Comparing the results from TRS partitioning against LOSO partitioning, we can see that the former tends to have more stars (i.e., smaller p-values) and more consistent metrics between the validation and test set. LOSO is prone to over-fitting when $w > 1.0$. We can therefore infer that the CKD using TRS partitioning is more effective. Indeed, it can be explained that both the teacher and student have seen examples similar to the test examples, and therefore produce joint embeddings that are of greater representability during the testing.

6.3. Interpretation

In our work, we are also particularly interested in revealing the contribution of each brain lobe and the band of EEG towards the emotion process. To this end, we visualize the skull saliency map for each band and subject, based on the trained student model and the peak response mapping [41] (PRM). The PRM is based on an observation that the backward propagation of the peak logit usually leads to informative regions of an image corresponding to the class.

Given the EEG band power calculated in the 6 bands ($0.3-5Hz$, $5-8Hz$, $8-12Hz$, $12-18Hz$, $18-30Hz$ and $30-45Hz$) from 32 electrodes (note that the β band is split to two sub-bands β_1 and β_2), the PRM is adopted as follows, with Fig. 5 illustrating the pipeline. The EEG band power from the LOSO partitioning is fed to the trained student model that is corresponding to the ith subject. For each trial of the test set from the ith subject, the peak scalar from the prediction is used to carry out the backward propagation, producing T_j 32×6 gradient vectors, where T_j denotes the time steps of the jth trial. The temporally averaged gradient vector ($N \times 32 \times 6$ where N denotes the trial number) of this trial together with those from other trials are averaged again to obtain the gradient vector (32×6) for the ith subject. Normalization of the gradient vectors over the 6 bands is employed, so that the inter-band information is preserved. Note that a per-band normalization rescaling each band independently is inappropriate, since the visualization would all look similar in terms of color intensity. After which, for the kth out of the 6 bands, 32 scalars corresponding to the 32 electrodes of the 10–20 system are used to generate the 6 skull saliency maps for the ith subject. The same process is conducted on all the 24 subjects obtaining 24×6 topographic saliency maps. Finally, the results from all the subjects are averaged and visualized as well. Theoretically, we expect that

Table 4
The result of our EEG model taught by visual knowledge against the standalone counterpart using the TRS partitioning. The mean, standard deviation, and p-value are reported. The p-value is obtained using the one-tailed paired t-test over the 10-fold TRS partitioning. TRS: trial-wise random shuffling. ↑: the higher the better. ↓: the lower the better. ⋆: $0.01 < p\text{-value} \leq 0.05$. ⋆⋆: $0.001 < p\text{-value} \leq 0.01$. ⋆⋆⋆: $p\text{-value} \leq 0.001$. Bold fonts indicate the best results.

TRS		Without KD	$w = 0.2$		$w = 0.4$	
		mean±std	mean±std	p-value	mean±std	p-value
Validation	RMSE ↓	0.067±0.005	0.066±0.005	0.001 (⋆⋆⋆)	0.066±0.005	0.004 (⋆⋆)
	PCC ↑	0.463±0.103	0.469±0.103	0.019 (⋆)	0.469±0.105	0.003 (⋆⋆)
	CCC ↑	0.444±0.109	0.450±0.110	0.014 (⋆)	0.449±0.111	0.030 (⋆)
Test	RMSE ↓	0.066±0.009	0.066±0.010	0.013 (⋆)	0.065±0.010	0.010 (⋆⋆)
	PCC ↑	0.435±0.205	0.442±0.205	0.013 (⋆)	0.442±0.204	0.011 (⋆)
	CCC ↑	0.415±0.201	0.422±0.202	0.022 (⋆)	0.422±0.201	0.015 (⋆)
TRS		$w = 0.6$		$w = 0.8$		$w = 1.0$
		mean±std	p-value	mean±std	p-value	mean±std	p-value
Validation	RMSE ↓	0.065±0.005	0.004 (⋆⋆)	0.065±0.005	0.001 (⋆⋆⋆)	**0.065 ± 0.005**	**< 0.001** (⋆⋆⋆)
	PCC ↑	0.473±0.107	0.001 (⋆⋆⋆)	0.477±0.108	0.001 (⋆⋆⋆)	**0.478 ± 0.107**	**< 0.001** (⋆⋆⋆)
	CCC ↑	0.453±0.114	0.011 (⋆)	0.457±0.115	0.005 (⋆⋆)	**0.458 ± 0.113**	**0.002** (⋆⋆)
Test	RMSE ↓	0.065±0.010	0.002 (⋆⋆)	0.065±0.010	0.003 (⋆⋆)	**0.065 ± 0.009**	**< 0.001** (⋆⋆⋆)
	PCC ↑	0.450±0.202	0.001 (⋆⋆⋆)	0.457±0.204	0.008 (⋆⋆)	**0.454 ± 0.207**	**0.010** (⋆⋆)
	CCC ↑	0.428±0.199	0.002 (⋆⋆)	0.436±0.202	0.011 (⋆)	**0.433 ± 0.205**	**0.010** (⋆⋆)
TRS		$w = 1.2$		$w = 1.4$		$w = 1.6$	
		mean±std	p-value	mean±std	p-value	mean±std	p-value
Validation	RMSE ↓	0.065±0.005	< 0.001 (⋆⋆⋆)	0.064±0.005	< 0.001 (⋆⋆⋆)	0.064±0.005	< 0.001 (⋆⋆⋆)
	PCC ↑	0.479±0.108	< 0.001 (⋆⋆⋆)	0.480±0.109	0.002 (⋆⋆)	0.481±0.108	< 0.001 (⋆⋆⋆)
	CCC ↑	0.459±0.114	0.004 (⋆⋆)	0.461±0.116	0.009 (⋆⋆)	0.461±0.115	0.006 (⋆⋆)
Test	RMSE ↓	0.064±0.009	< 0.001 (⋆⋆⋆)	0.064±0.009	< 0.001 (⋆⋆⋆)	0.064±0.009	< 0.001 (⋆⋆⋆)
	PCC ↑	0.456±0.206	0.001 (⋆⋆⋆)	0.459±0.205	0.004 (⋆⋆)	0.460±0.205	0.002 (⋆⋆)
	CCC ↑	0.436±0.204	0.005 (⋆⋆)	0.437±0.204	0.007 (⋆⋆)	0.438±0.204	0.006 (⋆⋆)
TRS		$w = 1.8$		$w = 2.0$			
		mean±std	p-value	mean±std	p-value		
Validation	RMSE ↓	0.064±0.005	0.003 (⋆⋆)	0.064±0.005	0.005 (⋆⋆)		
	PCC ↑	0.476±0.103	0.084	0.475±0.105	0.140		
	CCC ↑	0.453±0.108	0.317	0.452±0.110	0.415		
Test	RMSE ↓	0.064±0.009	0.028(⋆)	0.064±0.009	0.024 (⋆)		
	PCC ↑	0.450±0.207	0.252	0.450±0.207	0.249		
	CCC ↑	0.427±0.206	0.369	0.427±0.206	0.377		

Table 5
The result of our EEG model taught by visual knowledge against the standalone counterpart using the LOSO partitioning. The mean, standard deviation, and p-value are reported. The p-value is obtained by using the one-tailed paired t-test over the 24-fold LOSO partitioning. LOSO: leave-one-subject-out. ↑: the higher the better. ↓: the lower the better. ⋆: $0.01 < p\text{-value} \leq 0.05$. ⋆⋆: $0.001 < p\text{-value} \leq 0.01$. ⋆⋆⋆: $p\text{-value} \leq 0.001$. Bold fonts indicate the best results.

LOSO		Without KD	$w = 0.2$		$w = 0.4$	
		mean±std	mean±std	p-value	mean±std	p-value
Validation	RMSE ↓	0.068±0.007	0.067±0.006	0.096	0.067±0.006	0.002 (⋆⋆)
	PCC ↑	0.467±0.116	0.475±0.110	0.225	0.480±0.111	0.039 (⋆)
	CCC ↑	0.445±0.118	0.451±0.115	0.259	0.454±0.115	0.050 (⋆)
Test	RMSE ↓	0.066±0.025	0.065±0.025	0.059	**0.063 ± 0.025**	**0.001** (⋆⋆⋆)
	PCC ↑	0.474±0.267	0.480±0.269	0.034 (⋆)	**0.482 ± 0.269**	**0.014** (⋆)
	CCC ↑	0.377±0.250	0.382±0.250	0.319	**0.387 ± 0.253**	**0.033** (⋆)
LOSO		$w = 0.6$		$w = 0.8$		$w = 1.0$
		mean±std	p-value	mean±std	p-value	mean±std	p-value
Validation	RMSE ↓	0.067±0.006	0.004 (⋆⋆)	0.067±0.006	0.004 (⋆⋆)	0.066±0.006	0.004 (⋆⋆)
	PCC ↑	0.477±0.111	0.107	0.477±0.111	0.107 (⋆)	0.479±0.112	0.067
	CCC ↑	0.452±0.115	0.149	0.452±0.115	0.149	0.454±0.115	0.084
Test	RMSE ↓	0.064±0.025	0.003 (⋆⋆)	0.064±0.025	0.003 (⋆⋆)	0.063±0.024	< 0.001 (⋆⋆⋆)
	PCC ↑	0.482±0.268	0.022 (⋆)	0.482±0.268	0.022 (⋆)	0.481±0.270	0.030 (⋆⋆)
	CCC ↑	0.383±0.251	0.177	0.383±0.251	0.177	0.385±0.254	0.084
LOSO		$w = 1.2$		$w = 1.4$		$w = 1.6$	
		mean±std	p-value	mean±std	p-value	mean±std	p-value
Validation	RMSE ↓	0.066±0.006	< 0.001 (⋆⋆⋆)	0.066±0.006	< 0.001 (⋆⋆⋆)	0.066±0.006	< 0.001 (⋆⋆⋆)
	PCC ↑	0.484±0.111	0.015 (⋆)	0.484±0.115	0.016 (⋆)	0.486±0.114	0.011 (⋆)
	CCC ↑	0.458±0.116	0.012 (⋆)	0.458±0.118	0.015 (⋆)	0.459±0.118	0.012 (⋆)
Test	RMSE ↓	0.064±0.024	0.003 (⋆⋆)	0.064±0.024	0.001 (⋆⋆⋆)	0.063±0.024	< 0.001 (⋆⋆⋆)
	PCC ↑	0.481±0.271	0.037 (⋆)	0.480±0.270	0.200	0.481±0.270	0.150
	CCC ↑	0.381±0.253	0.324	0.378±0.250	0.798	0.380±0.251	0.504
LOSO		$w = 1.8$		$w = 2.0$			
		mean±std	p-value	mean±std	p-value		
Validation	RMSE ↓	**0.065 ± 0.006**	**< 0.001** (⋆⋆⋆)	0.065±0.006	< 0.001 (⋆⋆⋆)		
	PCC ↑	**0.487 ± 0.116**	**0.009** (⋆⋆)	0.485±0.116	0.017		
	CCC ↑	**0.460 ± 0.120**	**0.011** (⋆)	0.457±0.118	0.035		
Test	RMSE ↓	0.063±0.024	< 0.001 (⋆⋆⋆)	0.063±0.024	< 0.001 (⋆⋆⋆)		
	PCC ↑	0.481±0.270	0.140	0.479±0.271	0.220		
	CCC ↑	0.381±0.251	0.483	0.382±0.253	0.437		

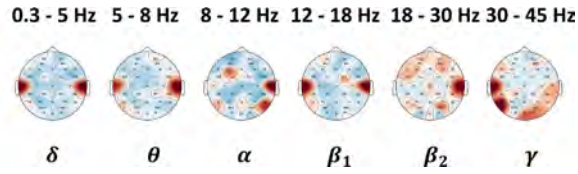

Fig. 7. The overall topographic saliency maps for all the 24 subjects. The warmer color means a higher gradient, and further implies more contribution to the valence prediction.

the saliency map of the β and γ bands should manifest a warmer color compared to those from other bands.

The per-subject topographic saliency maps over the six bands are illustrated in Fig. 6. Note that by referencing the saliency maps to the brain division and 32-electrode placement shown in Fig. 1, we can locate the active and less active brain regions for the state of high valence. We see that the last two columns, corresponding to the β_2 and γ bands, are apparently warmer than the rest four columns. Specifically, in the β_2 band, an active frontal lobes is observed on all subjects, while in the γ band, the occipital and parietal lobes take the place. We could explain that during the experiment, subjects should be focused on watching the movie clips, with their occipital and temporal lobes perceiving the visual and audio stimuli, the parietal lobe integrating the perceived, and the frontal lobe making the decision and directing the facial expression.

Moreover, active temporal lobes are observed over all the subjects and six bands. In the δ band ($0.3-5Hz$), no lobe is comparably active against the temporal lobe. In the θ band ($5-8Hz$), mild activation of the frontal lobe can be observed from Subject 1, 2, 3, 10, 14, 16, 18, 20, 23, and 27. And so are the active parietal lobe observed from Subject 8, 10, 13, 14, 19, and 21. And the active occipital lobes are observed from Subject 1, 5, 16, 19, 20, 21, 25, 27, and 30. In the α band ($8-12Hz$), more active frontal lobes can be observed from Subject 1, 5, 6, 7, 13, 19, 20, 21, 22, 29, and 30. Highly active parietal lobes are observed for all the subjects except for Subject 2. And, Subject 2, 3, 4, and 24 have active occipital lobes.

To summarize, we focus on the overall saliency map shown in Fig. 7. In terms of brain lobes, all the four lobes can be active on the six bands, which conforms to the fact that complex mental functions do not reside in any one place [16], instead of locating complex functions in precise brain areas [43–45]. In terms of bands, the β_2 and γ, with the frequency of $18-30Hz$ and $30-45Hz$, contribute the most to the human emotion process compared to other bands. The observation complies with the knowledge we discussed before, that i) the β band corresponds to a focused state of mind, and is more obvious in the frontal lobe, and ii) the γ band corresponds to the high-level cognition process such as the perception, transmission, and integration of the visual and audio stimuli from the occipital, temporal, and parietal lobes.

7. Conclusion

The goal of CER is to continuously predict the emotional trace in the multi-dimensional space over a specified time span. However, the recognition of emotion, if driven by data, suffers from the subject bias that exhibits in multiple stages of the emotion process. The issue is escalated for physiological signal, compared to the more objective and determinant cues from visual or audio modalities. Also, emotion cues over a large time span are a composition of complex one-actions, manifesting large variations on intensity and order in their duration. A model that is capable of capturing long-range dependencies is crucial in this area.

To improve the CER performance of the EEG modality, we explore the idea of teaching an EEG-based CER model using the visual knowledge from a visual-based CER model. The teacher model features a cascade CNN-TCN architecture and is fed by video frames. A subset [19] of the MAHNOB-HCI database [26] that includes facial videos, EEG signals, and continuous valence labels of 24 subjects are employed for the experiment. Two data partitioning schemes, i.e., the TRS and LOSO are employed. The experiment first validates the performance of the standalone teacher and student models in visual and EEG modalities, and obtained promising results compared to the baseline. After which, the spatiotemporal feature of the trained teacher is taken as the dark knowledge. The latter, together with the continuous label, is used to teach the student model. The experiment using the TRS and LOSO partitioning schemes both show an increase with statistical significance, i.e., p-value < 0.01 for TSR and p-value < 0.05 for LOSO partitioning on RMSE, PCC, and CCC.

Moreover, we also explore the interpretability of the student model from the physiological perspective. Specifically, the contribution of different brain regions and EEG frequency bands towards the emotion process is visualized using the PRM [41]. The result shows that all four brain lobes play synergistic roles in the emotion process. Complex brain function does not reside in any one location. And, the β_2 and γ bands, corresponding to the frequency band of $18-30Hz$ and $30-45Hz$, contribute more to the emotion process compared to other bands.

To the best of the authors' knowledge, this is the first visual-to-EEG CKD method on CER. Compared to other audio-visual CER works where a larger amount of data and continuous labels are available, our work involving EEG has only a subset of MAHNOB-HCI with continuous labels on valence to employ. More data are in need for more extensive experimentation and compelling results. Inspired by mental chronometry, which is a subject that studies the reaction time in perceptual-motor tasks to infer the content, duration, and temporal sequencing of mental operations, the future work may aim to model the pattern of visual and EEG cues, as well as their spatiotemporal connection in-between.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the RIE2020 AME Programmatic Fund, Singapore (No. A20G8b0102).

References

[1] R. Jiang, A.T. Ho, I. Cheheb, N. Al-Maadeed, S. Al-Maadeed, A. Bouridane, Emotion recognition from scrambled facial images via many graph embedding, Pattern Recognit. 67 (2017) 245–251.
[2] M. Faraki, X. Yu, Y.-H. Tsai, Y. Suh, M. Chandraker, Cross-domain similarity learning for face recognition in unseen domains, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15292–15301.
[3] P. Ekman, E.L. Rosenberg, What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS), Oxford University Press, USA, 1997.
[4] Y. Wang, S. Qiu, X. Ma, H. He, A prototype-based SPD matrix network for domain adaptation eeg emotion recognition, Pattern Recognit. 110 (2021) 107626.
[5] S.M. Alarcao, M.J. Fonseca, Emotions recognition using eeg signals: a survey, IEEE Trans. Affect. Comput. 10 (3) (2017) 374–393.
[6] L. Wang, K.-J. Yoon, Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks, IEEE Trans. Pattern Anal. Mach.Intell. (01) (2021) 1.
[7] T. Afouras, J.S. Chung, A. Zisserman, ASR is all you need: cross-modal distillation for lip reading, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 2143–2147.

[8] A. Nagrani, S. Albanie, A. Zisserman, Seeing voices and hearing faces: cross–modal biometric matching, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 8427–8436.
[9] J. Hoffman, S. Gupta, J. Leong, S. Guadarrama, T. Darrell, Cross-modal adaptation for RGB-D detection, in: 2016 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2016, pp. 5032–5039.
[10] S. Gupta, J. Hoffman, J. Malik, Cross modal distillation for supervision transfer, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016, pp. 2827–2836.
[11] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, D. Katabi, Through-wall human pose estimation using radio signals, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7356–7365.
[12] F.M. Thoker, J. Gall, Cross-modal knowledge distillation for action recognition, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 6–10.
[13] N.C. Garcia, P. Morerio, V. Murino, Modality distillation with multiple stream networks for action recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 103–118.
[14] Y. Tian, D. Krishnan, P. Isola, Contrastive representation distillation, in: International Conference on Learning Representations, 2020.
[15] S. Roheda, B.S. Riggan, H. Krim, L. Dai, Cross-modality distillation: a case for conditional generative adversarial networks, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 2926–2930.
[16] D.G. Myers, Psychology, 2004.
[17] N. Hussein, E. Gavves, A.W. Smeulders, Timeception for complex action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 254–263.
[18] N. Sankaran, D.D. Mohan, N.N. Lakshminarayana, S. Setlur, V. Govindaraju, Domain adaptive representation learning for facial action unit recognition, Pattern Recognit. 102 (2020) 107127.
[19] M. Soleymani, S. Asghari-Esfeden, Y. Fu, M. Pantic, Analysis of eeg signals and facial expressions for continuous emotion detection, IEEE Trans. Affect. Comput. 7 (1) (2015) 17–28.
[20] K. Somandepalli, R. Gupta, M. Nasir, B.M. Booth, S. Lee, S.S. Narayanan, Online affect tracking with multimodal kalman filters, in: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, 2016, pp. 59–66.
[21] J. Han, Z. Zhang, N. Cummins, F. Ringeval, B. Schuller, Strength modelling for real-worldautomatic continuous affect recognition from audiovisual signals, Image Vis. Comput. 65 (2017) 76–86.
[22] K. Wataraka Gamage, T. Dang, V. Sethu, J. Epps, E. Ambikairajah, Speech-based continuous emotion prediction by learning perception responses related to salient events: a study based on vocal affect bursts and cross-cultural affect in AVEC 2018, in: Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, 2018, pp. 47–55.
[23] H. Chen, Y. Deng, S. Cheng, Y. Wang, D. Jiang, H. Sahli, Efficient spatial temporal convolutional features for audiovisual continuous affect recognition, in: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, pp. 19–26.
[24] J. Zhao, R. Li, J. Liang, S. Chen, Q. Jin, Adversarial domain adaption for multi--cultural dimensional emotion recognition in dyadic interactions, in: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, pp. 37–45.
[25] D. Deng, Z. Chen, B.E. Shi, Multitask emotion recognition with incomplete labels, in: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, 2020, pp. 592–599.
[26] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE Trans. Affect Comput. 3 (1) (2011) 42–55.
[27] M. Jaiswal, C.-P. Bara, Y. Luo, M. Burzo, R. Mihalcea, E.M. Provost, Muse: a multimodal dataset of stressed emotion, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 1499–1510.
[28] N. Müller, R. Knight, The functional neuroanatomy of working memory: contributions of human brain lesion studies, Neuroscience 139 (1) (2006) 51–58.
[29] G.R. Alexandre, J.M. Soares, G.A.P. Thé, Systematic review of 3D facial expression recognition methods, Pattern Recognit. 100 (2020) 107108.
[30] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, L. Van Gool, A 3-D audio-visual corpus of affective communication, IEEE Trans. Multimedia 12 (6) (2010) 591–598.
[31] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, MS-Celeb-1M: a dataset and benchmark for large-scale face recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 87–102.
[32] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, VGGFace2: A dataset for recognising faces across pose and age, in: 2018 13th IEEE international conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 67–74.
[33] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[34] A. Romero, N. Ballas, S. Ebrahimi Kahou, A. Chassang, C. Gatta, Y. Bengio, FitNets: hints for thin deep nets, arXiv e-prints (2014). arXiv–1412
[35] E. Barsoum, C. Zhang, C.C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 279–283.
[36] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic, AVEC 2016: depression, mood, and emotion recognition workshop and challenge, in: Proceedings of the 6th International on Audio/Visual Emotion Challenge and Workshop, 2016, pp. 3–10.
[37] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, M. Pantic, AVEC 2017: real-life depression, and affect recognition workshop and challenge, in: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 2017, pp. 3–9.
[38] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, et al., AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition, in: Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, 2018, pp. 3–13.
[39] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, et al., AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition, in: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, pp. 3–12.
[40] D. Kollias, A. Schulc, E. Hajiyev, S. Zafeiriou, Analysing affective behavior in the first ABAW 2020 competition, in: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG), 794–800.
[41] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, J. Jiao, Weakly supervised instance segmentation using class peak response, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3791–3800.
[42] X. Bai, X. Wang, X. Liu, Q. Liu, J. Song, N. Sebe, B. Kim, Explainable deep learning for efficient and robust pattern recognition: a survey of recent developments, Pattern Recognit. 120 (2021) 108102.
[43] D.M. Beck, The appeal of the brain in the popular press, Perspect. Psychol. Sci. 5 (6) (2010) 762–766.
[44] A.P. Shimamura, Bridging psychological and biological science: the good, bad, and ugly, Perspect. Psychol. Sci. 5 (6) (2010) 772–775.
[45] W.R. Uttal, The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain, The MIT press, 2001.

Su Zhang received his bachelor degree from Xiamen University, China in 2013, and master degree from Yunnan Normal University in 2018. He is currently working towards the PhD degree in the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His current research interests include machine learning, deep learning and emotion recognition.

Chuangao Tang received the BS degree in electronic and information engineering from Northwest A&F University, Yangling, China, in 2013, and the MS degree in signal and information processing form Beijing Normal University, Beijing, China, in 2016. He is currently pursuing the PhD degree with the Key Laboratory of Child Development and Learning Science, Ministry of Education, in the School of Biological Science and Medical Engineering, Southeast University, Nanjing, China. He was a visiting student in Nanyang Technological University from 2020 to 2021. His research interests include affective computing, pattern recognition, and deep learning.

Cuntai Guan received his PhD degree from Southeast University, China, in 1993. He is currently a President's Chair Professor in the School of Computer Science and Engineering, Director of Artificial Intelligence Research Institute, Director of Centre for Brain-Computing Research, and Co-Director of S-Lab for Advanced Intelligence, at the Nanyang Technological University, Singapore. His research interests include brain-computer interfaces, machine learning, neural signal and image processing, and artificial intelligence. He is a recipient of the Annual BCI Research Award, the IES Prestigious Engineering Achievement Award, the Achiever of the Year (Research) Award, and the Finalist of President Technology Award. He is an Associate Editor of the IEEE Transactions on Biomedical Engineering, IEEE Transactions on Artificial Intelligence, Brain-Computer Interfaces, etc. He is a Fellow of IEEE and AIMBE.