Mental Disorder Detection for Schizophrenia Patients via Deep Visual Perception

Bing-Jhang Lin, and Li-Chen Fu, *Fellow, IEEE*

Abstract—Schizophrenia is a mental illness that will progressively change a person's mental state and cause serious social problems. In fact, schizophrenia involves a range of problems with thinking, behavior, and emotions, and its symptoms are highly correlated to emotion and depression. Importantly, doctors often refer to the mood aspect to estimate the severity of schizophrenia patients. We are thus motivated to design a mental disorder detection system for schizophrenia patients in order to provide an assessment for doctors during psychological counseling. Specifically, our system consists of two phases, including learning and detection. For the learning phase, we propose a multi-task learning framework to infer the patient's mental state, including emotion and depression. Observe that previous studies mainly focus on facial analysis to recognize human emotion and depression. Thus, they frequently fail to provide satisfactory performance due to the ambiguous emotional signals from the face. To overcome the limitation of facial analysis, we propose Cross-Modality Graph Convolutional Network (CMGCN) to effectively integrate visual features from different modalities, including the face and context. In addition to this effort, we design task-aware objective functions to realize better model convergence for multi-task learning, *i.e.*, emotion recognition and depression estimation. Further, we follow the correlation between depression and emotion to design a knowledge transfer approach, namely, Emotion Passer, to effectively transfer the prior knowledge on emotion to the depression model. On the other hand, for the detection phase, we draw on characteristics of schizophrenia to detect the mental disorder pattern based on the predicted mental state. In the experiments, we perform a series of experiments on several benchmark datasets, including CAER and AVEC 14. The proposed learning framework can achieve 87.23% in accuracy on CAER and 6.82/8.50 in MAE/RMSE on AVEC 14, which outperforms all advanced state-of-the-art (SOTA) methods. In addition, our system can achieve 73.38 in mAP on the collected schizophrenia patients.

Index Terms—Graph Convolutional Networks, Emotion Recognition, Depression Estimation, Transfer Learning.

1 Introduction

Schizophrenia [1] is a mental illness that will progressively change a person's mental state and cause serious social problems. People with schizophrenia are unable to express their real thinking ordinarily, and their behaviors are often different from normal ones. According to the medical literature about mental illness [2], this mental illness includes a range of issues with thinking, behaviors. Even though the emotions are not discriminative enough due to the nature of schizophrenia, doctors can still percept the mood of the patient during psychological counseling and infer the severity of schizophrenia further. Thus, in this paper, our main goal aims to provide an assessment for schizophrenia patients rather than diagnosis. Particularly, we here focus on detecting mental disorders about the mood aspect for patients during the counseling because the patient's mood is often unstable. The mental disorders about the mood aspect can be described by *mania* and *depression*. The former is a period of extremely high energy or mood and may cause schizophrenia patients with more severe psychotic symptoms, especially for hallucinations or disorganized speech. While the latter is a low emotional status, and the patients often stay in pervasive sadness and depression. Since the mental disorder about the mood aspect is highly correlated with emotion and depression, we can naturally employ techniques of emotion recognition and depression estimation via visual perception to infer mental states of patients, further realizing a mental disorder detection system. Our mental disorder detection system consists of two phases, including learning and detection. For the learning part, we propose a multi-task learning framework to learn a robust model to solve the limitation of the conventional FER systems in inferring the emotional status and depressive level of humans. On the other hand, for the detection phase, we first employ the learned robust multi-task model to infer the mental state of the patient, and then follow the observation about human emotion in cognitive science [3], [4] and the nature of schizophrenia [1] to design an algorithm to detect the mental disorders, including *mania* and *depression*.

Previous studies [5], [6], [7], [8], [9] assume that the human face can provide the most discriminative emotional signals, and hence have already done extensive discussions based on facial analysis. However, the conventional facial expression recognition (FER) systems frequently fail to precisely infer the mental state of people, even schizophrenia patients, due to lacking trustable emotional signals. Typically, human's facial expressions are extremely unstable emotional signals. Because of the facial muscle movements, such as blinking the eye or opening the mouth, facial expressions may yield some emotional signals conflicting to those which might differ from the total content in the associated video, and consequently leading to incorrect and inconsistent predictions. In addition to the above, due to

- *Bing-Jhang Lin was with the Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan (R.O.C.).*
  E-mail: r07921114@ntu.edu.tw
- *Li-Chen Fu is with the Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan (R.O.C.).*
  E-mail: lichen@ntu.edu.tw

*Manuscript received ***.*

the nature of schizophrenia and the effects of medicine, patients particularly tend to express fewer emotional signals [1]. Therefore, facial analysis alone may not be suitable for detecting the emotional status of patients. Moreover, in cognitive science, some studies [3], [4] have shown that people recognize the emotions of others not only from their faces but also from the surrounding context, such as interaction with others, and the overall behavior of human appearance.

To overcome the shortcomings of facial analysis, in our multi-task learning framework, we design *cross-modality graph convolutional networks* (CMGCN) to effectively integrate the visual cues from different modalities, including the face and context. Specifically, We draw on an affinity graph to encode the pixel-wise correlations between different modalities, and then a sampling scheme is employed to link those pixel pairs (cross-modality) with similar semantic meaning, that is emotional class or depressive level, and to drop other semantic irrelevant pairs, such as background pixel pairs. By doing so, we can thus yield a robust representation for both tasks, including emotion recognition and depression estimation. In addition to this effort, we also design task-aware objective functions to realize better model convergence. For emotion recognition, which is a classification task, we propose *density loss* for metric learning to form a powerful regularization for embedding learning. In particular, we consider comprehensive criteria related to the embedding density (the density of each class in the embedding space) and the orthogonal relation (a regularization criterion for similarity pairs) to design the metric function, which thus allows us to accomplish a discriminative embedding with high intra-class compactness and inter-class separability. On the other hand, for depression estimation, we take the classification viewpoint to form *distributed loss* for the regression task. By dividing the regression level into several bins, it can emphasize the influence of the loss in a meticulous way. Further, as depression is a disorder characterized by a low emotional status, we present a knowledge transfer strategy, namely, *emotion passer* (EP), to effectively pass the prior knowledge on emotion to the depression model. In each training iteration, our scheme takes Exponential Moving Average (EMA) to smoothly transfer the knowledge from the emotion model to the depression one. As knowledge is progressively transferred, it can achieve a robust model for unseen data in depression estimation. Having obtained the well learned multi-task model, we then follow the observation about human emotion in cognitive science [3], [4] and the nature of schizophrenia [1] to design an algorithm to detect the mental disorders, including *mania* and *depression*.

The main contributions of our system can be characterized as follows: (1) We introduce a novel mental disorder system for schizophrenia patients in order to provide an assessment for doctors during psychological counseling. (2) We design an multi-task learning framework with comprehensive and novel modules, and thus it can yield a robust multi-task model and achieve the state-of-the-art (SOTA) performance (3) We follow the medical literature and the cognitive science of human emotion to design an rule-based algorithm in detecting the mental disorders about the mood aspect of schizophrenia patients. In the collected cases from National Taiwan University Hospital (NTUH), the proposed detection system can achieve 73.38 performance in mAP (mean Average Precision), which denotes our system can detect the mental disorders about mood aspect to a certain degree.

2 Related Work

2.1 Emotion Recognition

Emotion recognition is a process of identifying the internal state of a given person. In the computer vision area, human emotion is often defined as one out of a set of discrete labels, including happiness, anger, sadness, surprise, disgust, and fear. To recognize the correct labels, previous studies [5], [6], [7], [8], [9], [10], [11] mainly rely on facial analysis, which unfortunately will experience the limited ability to precisely infer the mental state of the human for the reason as mentioned earlier. To overcome these limitations, some methods adopt other visual cues, such as the context information, to boost model robustness for the real-world scenario. By involving the context information, this kind of emotion recognition can also be called as Context-Aware Emotion Recognition (CAER). Schindler *et al.* [12] adopted the body pose to identify six emotion categories. Chen *et al.* [13] proposed a context fusion network to recognize human emotion by integrating events, objects, and scenes. Kosti *et al.* [14] presented an end-to-end model for emotion recognition in context by jointly encoding the face and body information. However, these approaches are in lack of practical solutions to encode the salient context information for emotion recognition in the context. To better model the information from different modalities, Lee *et al.* [15] presented a two-stream architecture followed by a fusion network for CAER. One stream focuses on the face modality, and the other focuses on context modality. Instead of directly feeding the context image into the context stream, they particularly mask the human face to explore more emotion relevant features from the context image. Mittal *et al.* [16] proposed a multi-model approach for the CAER task. Specifically, they exploit the depth images to model socio-dynamic interactions and employ several sub-networks to infer the perceived emotion.

2.2 Depression Estimation

According to the machine learning perspective, depression estimation is essentially a regression problem. In the public benchmark datasets, such as AVEC 2013 [17] and AVEC 2014 [18], the depression level is categorized based on Beck Depression Inventory-II (BDI-II) [19]. The mainstream can be regarded as a pipeline, including two steps. The first step is to extract the visual features from the given video. After that, the second step is to predict the score based on the extracted visual features. In the earlier stage, many

BDI-II Score	Depression Severity
0 - 13	None
14 - 19	Mild
20 - 28	Moderate
29 - 63	Severe

TABLE 1
Beck Depression Inventory-II (BDI-II) score and Depression Severity.

studies employed methods to extract hand-crafted visual features. The baseline approach in AVEC 2013 [17] adopted Local Phase Quantization (LPQ) followed by a regressor for learning and prediction. Unlike the single hand-crafted descriptor, Cummins *et al.* [20] exploited a bag-of-words scheme to model the visual feature with a complex way to realize better performance. However, these approaches basically require many assumptions and prior knowledge to define visual descriptors. Thus, it is not easy to be generalized to the real-world application. To overcome these limitations, recent studies start to employ an end-to-end learning manner to improve performance. Yang *et al.* [21] designed a multi-model framework to encode more meaningful features for depression estimation. They proposed a multi-model framework to jointly encode the visual and audio information via Convolutional Neural Networks (CNNs) and Long-Short Term Memory (LSTM). Instead of combining the audio information, Zhu *et al.* [22] proposed a CNNs framework to model the visual information, including facial appearance and dynamics. Specifically, they designed a two-stream architecture to encode both special and temporal information for learning and prediction. Ding *et al.* [23] proposed a deep learning framework, called DeepInsight, to quickly diagnose the Autism Spectrum Disorder (ASD) and Major Depressive Disorder (MDD) based on CNNs. They designed a multi-task learning framework for the diagnosis of ASD and MDD. By a share-weight backbone with a multi-scale design, they showed an impressive performance on their own datasets.

3 System Overview

Our mental disorder detection system is illustrated in Figure 1. The system consists of two parts, including learning and detection. For the learning part, we propose a multi-task learning framework to learn a robust model to solve the limitation of the conventional FER systems in inferring the emotional status and depressive level of humans. For the detection part, we first apply the learned robust multi-task model to infer the mental state of the patient, and then follow the observation about human emotion in cognitive science and the nature of schizophrenia to design an algorithm to detect the mental disorders, including *mania* and *depression*. In what follows, we first introduce an overview of our multi-task learning framework in Section 3.1, and then, in Section 3.2, we illustrate an algorithm to detect the mental disorders about the mood aspect.

3.1 Multi-task Learning Framework

First of all, we introduce our multi-task learning framework, which is shown on the learning part in Figure 1. Our design mainly consists of four main components: 1) *cross-modality graph convolution networks* (CMGCN), 2) *task-aware objective functions* (Subsection 3.1.2), and 3) *emotion passer* (Subsection 3.1.3). For the backbone network of each task, we here exploit two-stream architecture, including 2 CNNs with 5 layers, to encode high-level features from different modalities, including the face and context. Following the last layer of backbone network, the extracted high-level feature maps can be expressed as $X_f \in \mathbb{R}^{N \times h \times w \times D}$ and $X_c \in \mathbb{R}^{N \times h \times w \times D}$, where f and c respectively symbolize the face and context modalities; N, h, w, and D are the batch size, height, width, and the embedding size, respectively. We then employ the proposed CMGCN to integrate these high-level features to yield a comprehensive representation for the following processing. The details of CMGCN will be introduced in Subsection 3.1.1. As the nature of each task is completely different, one for the classification and another one for the regression, we develop the task-aware objective functions for each task to realize a better model convergence, which will be cleared in SubSection Subsection 3.1.2. Finally, for cross-task knowledge transfer, we present emotion passer in Subsection 3.1.3 to effectively transfer the prior knowledge from the emotion model to the depression model via Exponential Moving Average (EMA) mechanism. With our well designed multi-task learning framework, our framework can successfully outperform the other state-of-the-art algorithms.

3.1.1 Cross-modality Graph Convolution Networks

An intuitive idea [15] is to weight the features from different modalities to emphasize the importance of each separately. Observe that the previous study adopts Global Average Pooling (GAP) to generate the representation of each modality in isolation, and then separately learns different weights via several isolated Multi-Layer Perceptron (MLP) layers. However, this fusion mechanism will include many irrelevant emotional regions, such as background pixels. Instead of the above separate fusion scheme, we propose a Cross-Modality Graph Convolutional Networks (CMGCN) to effectively integrate the visual features from the face and context modality, as shown in Figure 2. Particularly, we here exploit the graph viewpoint to model the pixel-wise correlations between different modalities so that we can learn a joint representation comprehensively. Since too many irrelevant regions are included in the context image, we thus introduce a sampling scheme to build a sparse graph, which allows us to seek those pixel pairs with similar semantic meaning and discard other dissimilar ones. Finally, we employ GCN to integrate features from different modalities via the constructed sparse graph to yield a robust representation. In what follows, we elaborate on our CMGCN step by step. We begin with the cross-modality graph construction to present how we model the correlations between different modalities. Then, we describe the core module, sampling scheme and GCN embedding, which are used to integrate features from different modalities. Finally, we conclude the final representation via a bidirectional fusion mechanism.

Cross-modality graph construction: To encode the correlation between different modalities, we follow a similar idea to model the pairwise relationship via an affinity graph [24]. Note that we here consider using the graph to model the pairwise relationship about the pixel pairs across different modalities, which is different from the previous study that adopts the graph to model the pairwise relationships about the inter-identity relationships. Given a pair of the face and context features maps, the size of both of their tensors is shown as $h \times w \times D$, and we then reshape the feature maps into matrices with dimension $hw \times D$ for convenient processing. To construct the cross-modality

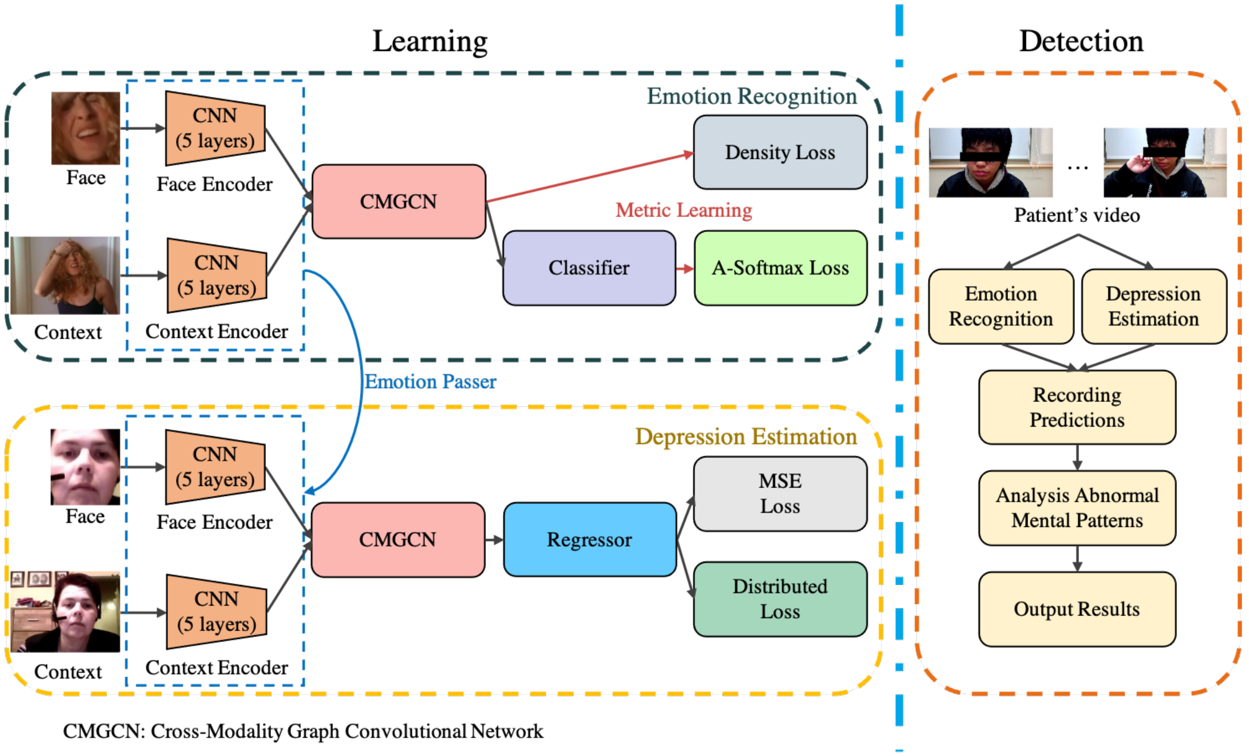

Fig. 1. Our mental disorder detection system.

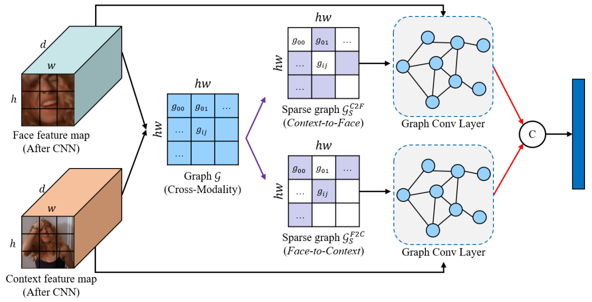
d
w
h
Face feature map (After CNN)
hw
hw
Graph $\mathcal{G}$ (Cross-Modality)
hw
hw
Sparse graph $\mathcal{G}_S^{C2F}$ (Context-to-Face)
Graph Conv Layer
hw
hw
Sparse graph $\mathcal{G}_S^{F2C}$ (Face-to-Context)
Graph Conv Layer
C
d
w
h
Context feature map (After CNN)

Fig. 2. An overview of our CMGCN.

graph $\mathcal{G} \in R^{hw \times hw}$ from the given face and context feature maps, we here regard each pixel of each feature map as a vertex, and the edge between each pair of vertices across different modalities is initialized via the cosine similarity. The edge can be formulated as

$$g_{ij} = < \frac{\mathbf{x}_i^f}{\|\mathbf{x}_i^f\|}, \frac{\mathbf{x}_j^c}{\|\mathbf{x}_j^c\|} > \tag{1}$$

where $\mathbf{x}_i^f$ and $\mathbf{x}_j^c$ denote the pixel from face and context feature maps, respectively. i and j respectively indicate the indexes of i-th and j-th pixel in each modality.

As the value range of g_{ij} is in $[-1, 1]$, the negative values may countervail other features. Here, we apply a kernel function to transform the graph G, say the range of $g_{ij}$from $[-1, 1] \rightarrow [0, 1]$. Moreover, it can further intensify the difference of inter-entry in the given graph to achieve more useful weights for graph embedding. The kernel function can be expressed as follows:

$$g_{ij} = \exp g_{ij} - 1^p \tag{2}$$

where the range of g_{ij} will be transformed from $[-1, 1] \rightarrow [0, 1]$. p is the power factor.

Sampling scheme and GCN embedding: Observe that the cross-modality graph $\mathcal{G}$ is a fully-connected graph, which links the pairwise correlation of pixels among different modalities, *i.e.*, from face to context or the other way around. As we mentioned above, only a few regions provide the discriminative emotional signals in the context image. Thus, if we directly apply this graph for the following GCN embedding, the resulting graph feature may easily be dominated by the information that denotes the dissimilar pairs and irrelevant to semantic meaning, such as background pixel pairs, as shown in Figure 3. To this end, we come up with a sampling scheme to enhance the sparsity of the graph to reduce the influence of other irrelevant information to tackle the above issue.

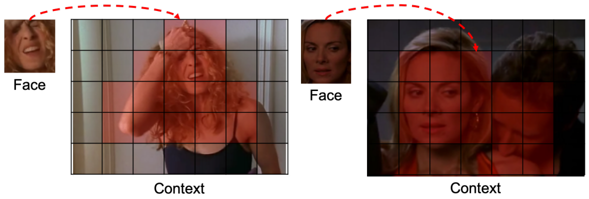

Fig. 3. The correlations between the face and the context modalities. As we can see, in the context image, only a few regions (red pixels) provide discriminative emotional signals that are highly related to semantic meaning (ground truth label).

Having obtained the graph $\mathcal{G} \in \mathbb{R}^{hw \times hw}$, we can now enhance its sparsity to reduce the influence of those uncorrelated pixel pair. Note that these sampling strategies will generate a mask $\in \mathbb{R}^{hw \times hw}$ with the same shape as the given graph $\mathcal{G}$, where an entry $m_{ij} = 1/0$ means that its counterpart in $\mathcal{G}$ would be kept/dropped. Our core idea is to make the model learn to pay more attention to relevant visual features accross different modalities, *i.e.*, connecting those pixels with similar semantic meaning in different modalities. Specifically, it aims to encourage relevant features with higher similarity (probability) values be selected in the sampling process so that the GCN embedding could result from the legitimate weighting of combining relevant features. To this effect, we consider using the Bernoulli sampling, which responses to those elements with higher similarity (probability) values in the given graph $\mathcal{G}$. Notably, because of the random property of Bernoulli sampling, lower similarity elements will not be completely ignored. Thus, the model can learn with more kinds of graph features (messy or high-quality) and understand how to intensify the connections between relevant features and suppress other irrelevant ones. By doing so, the evolution of GCN embedding is expected to be smooth rather than protruding and thus achieve better performance. The resulting sparse graph $\mathcal{G}_{sparse}$ can be expressed as follows:

$$P(m_{ij} = 1) = g_{ij}; P(m_{ij} = 0) = 1 - g_{ij} \tag{3}$$

$$g_{ij}^{sparse} = g_{ij} \times m_{ij} \tag{4}$$

where $g_{ij} \in [0, 1]$ is the similarity between different modalities and can be regarded as a probability for the Bernoulli sampling. As the m_{ij} is a random variable in the Bernoulli processing, it can be expressed as a form of probability.

After building the sparse cross-modality graph $\mathcal{G}_{sparse}$, we then introduce a general GCN [25] to construct an embedding to integrate features. Comparing with the typical graph convolution operation, the obtained graph $\mathcal{G}_{sparse}$ is sparse and exhibits essential correlations of different modalities. It can consequently result in a robust representation more relevant to the following emotion recognition and depression estimation. Formally, GCN embedding can be expressed as follows:

$$\hat{A} = \tilde{D}^{-1}\tilde{A} \tag{5}$$

$$F = \sigma(\hat{A}XW) \tag{6}$$

where $\tilde{A} = \mathcal{G}_{sparse} + I_n$ is an adjacency matrix describing the critical correlations of different modalities, and $\tilde{D}$ is a degree matrix of $\tilde{A}$, where $\tilde{D}_{ii} = \sum_{j=1}^{N} \tilde{A}_{ij}$ is a normalization diagonal degree matrix. σ represents an activation function for non-linear mapping, and X and W are the input feature map and the embedding of GCN, respectively.

Bidirectional fusion: To acquire a comprehensive representation, we here consider a bidirectional way to explore critical regions among multi-modalities. Specifically, we seek the highly correlated regions not only from face to context but also from context to face. Further, for the graph feature of each modality, we execute a residual connection to prevent the overfitting problem. Finally, we employ GAP to the graph feature of each modality and adopt concatenation operation to yield the final representation. The final representation of our CMGCN can be expressed as follows:

$$X_{final} = [GAP(X_G^f + X^f), GAP(X_G^c + X^c)] \tag{7}$$

where X^f and X^c are face and context features, X_G^f and X_G^c indicate the graph features based on Equ. (6), $[\cdot, \cdot]$ denotes the concatenation operation.

3.1.2 Task-aware Objective Functions

In this section, we introduce the proposed task-aware objective functions. Since the nature of each task is completely different, *i.e.*, emotion recognition is a classification task while depression estimation is a regression task, it is necessary to design task-aware objective functions to realize a better model convergence. In the following, we first elaborate on the proposed *density loss* for emotion recognition and then describe the *distributed loss* for depression estimation.

Density loss for emotion recognition: Emotion recognition is essentially a classification task that focuses on learning a discriminative embedding to better recognize the emotion from the given human images. After obtaining the representations via the proposed CMGCN, we then exploit the metric learning techniques to learn an embedding space with high intra-class compactness (samples with the same class label can be aggregated together) and inter-class separability (samples with the different classes are far apart).

We observe that one essential property, embedding density which denotes the density of each class in the deep embedding space, is often neglected in the previous metric learning techniques. Because of the intention of metric learning, we can assume that the metric learning can congregate samples of each class in an embedding space to certain degree. Nevertheless, the distribution of each class may still be sparse and with varied density, even when we apply some specific mining and weighting strategies to emphasize the influence of informative samples that play the critical role in metric learning. To solve the issue, we here consider enforcing the density prior of each class to form a useful regularization for embedding learning.

Given a batch of representation $X \in \mathbb{R}^{n \times D}$ and its corresponding label $Y \in \mathbb{R}^{n}$, we first impose the cosine-similarity to build a similarity matrix $S \in \mathbb{R}^{n \times n}$ to describe the pairwise relationships between samples. For the convenience of presentation, we here form two types of similarity pairs based on the label relation of the given pair. s_{ij}^{p} denotes an intra-class similarity pair, where the sample x_i and the sample x_j are with the same class ($y_i = y_j$), and s_{ij}^{n} denotes an inter-class similarity pair, where the sample x_i and the sample x_j are with the different class ($y_i \neq y_j$).

A natural way to measure the density is to average all intra-class/inter-class similarity pairs according to the given anchor i. The measured density of i-th anchor can be expressed as:

$$\mu_i^p = \frac{1}{N_p} \sum_{j=1}^{N} s_{ij}, where y_i = y_j \tag{8}$$

$$\mu_i^n = \frac{1}{N_n} \sum_{j=1}^{N} s_{ij}, where y_i \neq y_j \tag{9}$$

where μ_i^p and μ_i^n represent the density of intra-class similarity pair and inter-class similarity pair, respectively. s_{ij} is the similarity between x_i and x_j, p and n symbolize the intra-class and inter-class based on the label y_i and y_j. N_* denotes the number of satisfied pairs.

By averaging the similarity pairs, we can clearly identify which sample degrades the density so that we can emphasize its influence by enlarging its weight to let the model pay more effort on it to tune the parameters. Since always at least one sample is not satisfied (*i.e.*, less or higher than the density μ_i^p or μ_i^n), can adaptively and continuously emphasize the influence of outlier pairs based on the density, no matter how close an underlying sample is. The emphasizing terms of the proposed density loss are defined as follows:

$$w_{ij} = \begin{cases} \exp\left([\mu_i^p - s_{ij}]_+\right), & \text{if } y_i = y_j, \\ \exp\left([s_{ij} - \mu_i^n]_+\right), & \text{otherwise}, \end{cases} \tag{10}$$

where $[\cdot]_+$ denotes the hinge function in order to drop satisfied samples (larger/less than intra-class/inter-class field). We here consider using the exponential function to keep the weight of satisfied samples as $e^0 = 1$ and enlarge the penalty of unsatisfied samples.

Unlike learning to reduce the disparity between positive and negative samples ($s_n - s_p$) [26], we instead learn to regularize the feature distribution by minimizing the disparity between the similarity pair and its optimal boundary ($b_p - s_p$ and $s_n - b_n$). Specifically, we consider two boundaries, one for intra-class b_p and the other for inter-class b_n, and try to minimize the disparity between pairs of respective class type. Note that because we consider the angular measurement (cosine-similarity) for learning, the boundary b_p or b_n is a fix similarity value, for example, $b_p = 0.7$ or $b_n = 0.3$. We expect each intra-class similarity pair s_{ij}^p to be greater than the intra-class boundary b_p and each inter-class similarity pair s_{ij}^n to be less than the inter-class boundary b_n. Further, to better regularize the distribution of each class, we here design the boundary with an orthogonal relation, $b_n = 1 - b_p$. Because the cosine-similarity encodes each data point on a hyper-sphere, the intra-class boundary b_p can be regarded as a tolerated area of each class, and b_n denotes the distributed region of other classes to realize better regularization.

$$L_{ij} = \begin{cases} [b_p - s_{ij}]_+, & \text{if } y_i = y_j \\ [s_{ij} - b_n]_+, & \text{otherwise} \end{cases} \tag{11}$$

where$[\cdot]_+$ denotes the hinge function in order to clear the loss of satisfying samples.

Finally, we multiply the loss l_{ij} to our emphasizing terms w_{ij} to yield the final penalty for learning. Particularly, we here adopt L_p-norm among the entire mini-batch to minimize the loss of each pair to further emphasize the penalty. Our Density loss function can be cast as follows:

$$\mathcal{L}_{density} = \frac{1}{n}\left(\sum_{i=1}^{n}(w_{ij}L_{ij})^p\right)^{\frac{1}{p}} \tag{12}$$

where L_{ij} denotes the loss value that defines in (11) and $p > 1$ specifies the underlying norm function.

To yield better performance for emotion recognition, a classification task, we here consider learning the embedding space by pairwise and classwise metric learning techniques. For the pairwise metric learning, we adopt the proposed Density loss to intensify discrimination of the learned embedding space by the formed regularization in Equ. (12). For the classwise metric learning, we draw on an Angular softmax (A-softmax) [27] to regularize the distribution of each class. As a matter of fact, for classwise metric learning, we will approximate the center of each class and exploit the log-likelihood to maximize the score of the target class and minimize other non-target ones.

Distributed loss for depression estimation: Depression estimation is a regression task, which focuses on learning a model to precisely predict the depression level from the given human images. Commonly, for a regression task, a fundamental strategy is applying the Mean Square Error (MSE) loss to minimize the regressive value and the ground truth. However, MSE loss only exploits the square operation to extend the loss value; thus, it cannot take the meticulous way to adjust the strength of the loss. Specifically, we first adopt an MLP layer with 1 neural output to yield the regressive score $\hat{y}_i$, and then impose the MSE loss to minimize the difference between the predicted score and the target level y_i (ground-truth). The MSE loss can be expressed as:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{13}$$

where $\hat{y}_i$ and y_i indicate the predicted score and the target regressive level, respectively, i denotes the index of i-th sample in the give mini-batch.

To better refine the loss value, we here take the classification viewpoint to formulate a new loss function for the regression task. According to the AVEC datasets [17], [18], the level of depression is labeled with a single integer value, and the learning target can be expressed as a set with n continuous integer values, *i.e.*, 1 to n. Thus, we can divide the learning target into $n + 1$ bins for learning, as shown in Figure 4.

After formulating the loss function via a classification perspective, the learning target (regression value) becomes clear so that we can adopt the log-likelihood to maximize the score of the target bins. Specifically, the target y_i will be wrapped by y_i and y_{i+1} bin; therefore, we can jointly maximize the score of y_i and $y_i + 1$ bins to realize the

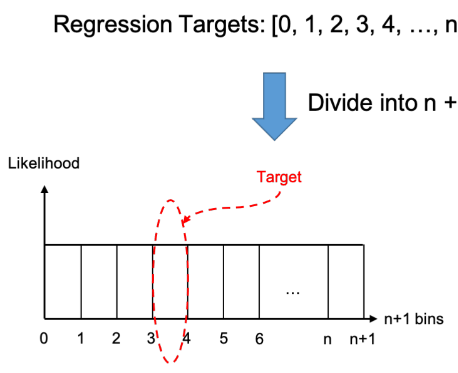

Fig. 4. The illustration of the concept of our Distributed loss. We introduce the classification viewpoint into the regression task.

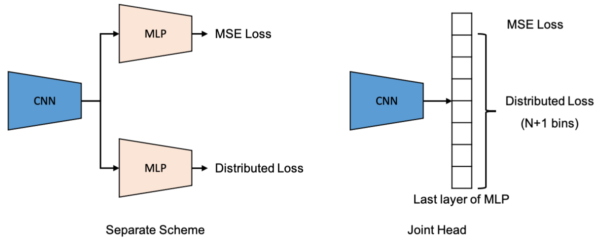

Fig. 5. The difference between separate scheme and our joint head.

objective. Finally, our Distributed loss can be expressed as follows:

$$\mathcal{L}_{distributed} = \frac{1}{N}\sum_{i=1}^{N} - \log(s_{y_i}) - \log(s_{y_i+1}) \tag{14}$$

where s_{y_i} and s_{y_i} indicate likelihood scores of target bins.

Finally, we jointly adopt MSE loss and the proposed Distributed loss to realize better performance. A common strategy is learning with an isolated linear layer (MLP) for each objective function. But this strategy easily leads to inconsistent results. One MLP may dominate the model while another makes slightly effects; therefore, the predicted results of these two networks are different. Unlike this separate learning scheme, we here propose a joint head for prediction based on the You Only Look Once (YOLO) [28] architecture, see Figure 5. Since the weights of two objective functions are shared, it can greatly prevent the inconsistent results. By synergizing Distributed loss with MSE loss, we can emphasize the loss value meticulously and realize better performance.

3.1.3 Emotion Passer

In this section, we elaborate on the proposed training strategy, namely, Emotion Passer. As we mentioned earlier, as depression is a disorder characterized by a low emotional status, it is highly correlated to emotion. Therefore, a common training strategy for depression estimation is using several FER datasets (source domain) to pre-train the model, and then adopts the collected depression dataset (target domain) to fine-tune the model. However, this strategy often requires careful selection of the training parameters, such as the learning rate, to handle the shifting of the embedding space. On the other hand, the training procedure will become inefficient due to sequential training with several source datasets.

To promote training efficiency, our Emotion Passer takes the epoch-wise viewpoint to transfer the knowledge. Specifically, in each iteration, we first train the emotion model, and then, we impose Exponential Moving Average (EMA) mechanism to transfer the weights from the emotion model to the depression model. Since the EMA can smoothly adjust the updated factor based on the current iteration, the knowledge transfer procedure is expected to be smooth rather than sluggish. Thus, it can realize a better performance. Formally, the knowledge transfer of Emotion Passer can be expressed as:

$$\theta_t^d \leftarrow \alpha\theta_{t-1}^d + (1-\alpha)\theta_t^e \tag{15}$$

where α and t represent a smoothing coefficient parameter and the current iteration, respectively, d and e denote the depression and emotion models, respectively, and θ denotes the parameters of the model.

3.2 Mental Disorder Detection

In this section, we begin with the introduction of the mental disorder of schizophrenia patients. Then, we elaborate on the relations between the mental disorder and our learning framework and how we implement the system to accomplish the mental disorder detection system in order to provide an assessment to schizophrenia patients.

Schizophrenia is a psychiatric disorder characterized by continuous or relapsing episodes of psychosis [1]. Based on the medical literature about mental illness [2], this mental disorder involves a range of problems with thinking, behavior, and emotions. Typically, the major characteristics of schizophrenia are highly relevant to psychotic disorders, such as delusions, hallucinations, and disorganized speech. As we mentioned earlier, in this thesis, our goal aims to provide an assessment for doctors to evaluate the severity of schizophrenia patients. Thus, we focus on detecting the mental disorder about the mood aspect of the patient during the counseling. As the mental disorder about the mood aspect is highly correlated with emotion and depression, we can draw some clues from emotion recognition and depression estimation to estimate the real-time mental state of patients in order to infer the mental disorder about mood aspect further. Particularly, since the mental disorders about the mood aspect of schizophrenia patients can be described by Mania and Depression. The former is a period of extremely high energy or mood and may cause schizophrenia patients with more severe psychotic symptoms, especially for hallucinations or disorganized speech. While the latter is a low emotional status, and the patients often stay in pervasive sadness and depression. Since the above disorders are highly correlated to emotion and depression, we can draw some clues from emotion recognition and depression estimation to estimate the real-time mental state of patients in order to infer mental disorders further.

In Figure 6, we illustrate the detection flow of our algorithm. Given an untrimmed video of a patient, we first adopt emotion recognition and depression estimation to infer the emotional status and depressive level of the patient. Next, we follow the observation of human behavior in cognitive science [3], [4] to design a sliding window strategy to aggregate the prediction with a specific duration. Finally, we will analyze the abnormal mental pattern based on the signs [2] of Mania and Depression to detect mental disorders.

To infer the real-time mental state, which denotes the emotional status and depressive level of patients, we follow the observation of human emotion in cognitive science [3], [4] to design how we aggregate the current mental state and the past states. Further, previous studies [3], [4] also clam that emotion essentially is an action readiness characterized by happening quickly, short duration, and non-periodic happening. According to the statistics [29], the duration of an emotion lasts between 0.5 to 4 seconds. Thus, we here consider to use the emotional status and depressive

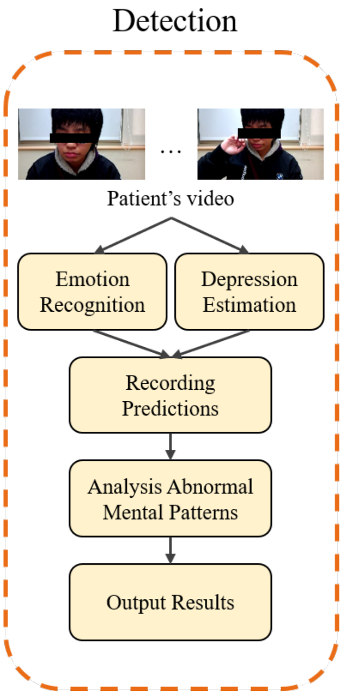

Fig. 6. The detection flow of our detection algorithm.

level with 2 second duration, that is the average period of the emotions, for analyzing the abnormal mental patterns for mental disorders. Specifically, in our implementation, we use the camera with 32 frames per second (fps) to record the video and adopt the sliding window strategy to sequentially capture frames for mental disorder detection, as shown in Figure 7. For a real-time prediction, we build a video clip with 16 frames (0.5 seconds) for the process of emotion recognition and depression estimation. After that, we implement a queue to gather 64 predictions for analysis to detect abnormal mental patterns.

Having obtained the sequential predictions (64 predictions; 2 second duration), we then dissect the recording to detect abnormal mental patterns. Since the nature of Mania and Depression is quite different, we design a detection algorithm, which is consulted to the sign of each disorder, to detect the abnormal patterns. For the Mania, schizophrenia patients may be fearful or suspicious of friends. Also, during a hallucination, people may arise with severe episodes of mania, and their emotional status may be unstable. Thus, we introduce the entropy to measure the uncertainty of the given prediction queue. The entropy can be expressed as follows:

$$entropy = -\sum_{i=1}^{C} P(\hat{y}_i)\log(P(\hat{y}_i)) \tag{16}$$

where C and $P(\hat{y}_i)$ denote the total number of emotional classes and the probability of i-th emotional class, respectively.

Since schizophrenia patients with Mania may behave in various emotions in a short time-serious, its emotional status will distribute uniformly, marked by blue in Figure 8.

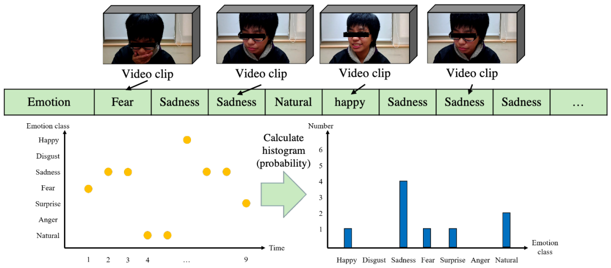

Fig. 8. The unstable emotional status.

Therefore, the entropy of the given emotional status is high. Thus, we will give a Mania alert for notification. In Figure 8, we show an example about unstable emotional status. In contrast, if the patients stay in a stable emotion, it will result in low entropy value. For the stable emotional status, we first get the emotion class with the max probability to represent the stable emotion. Then, we will check that if the patient stays in negative emotions, such as fear or anger, we will give a Mania alert for notification. On the other hand, if the patient stays in neutral emotional class for a long time (10 seconds), we will give a flat affect alert for notification. The flow of the disorder detection based on the patient's emotional status is shown in Figure 9.

In our implementation, we first transform the them into histogram to count occurrences of each emotional class, and then divide the histogram by total number of occurrences to generate the probability. After that, we can calculate the entropy of the given emotion distribution to check the

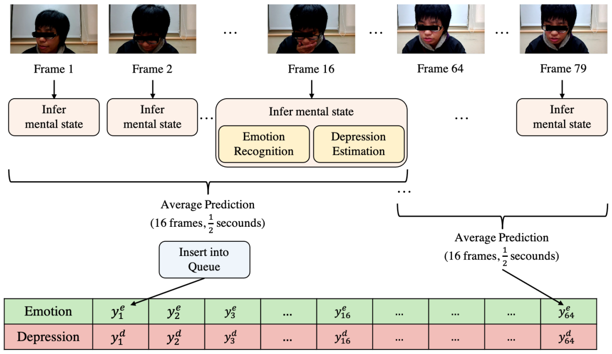

Fig. 7. Illustration of how we record the mental state of patients.

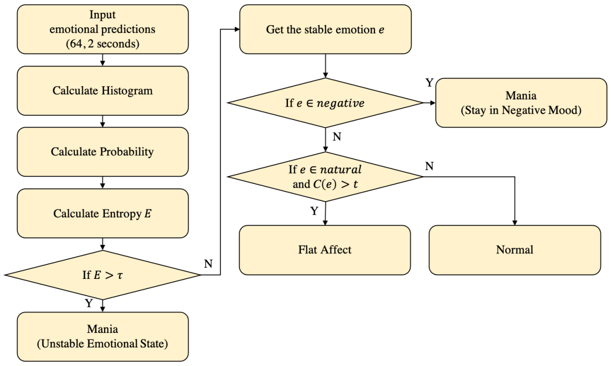
Input emotional predictions (64, 2 seconds)
Calculate Histogram
Calculate Probability
Calculate Entropy E
If $E > \tau$
N
Y
Mania (Unstable Emotional State)
Get the stable emotion e
If $e \in negative$
Y
Mania (Stay in Negative Mood)
N
If $e \in natural$ and $C(e) > t$
N
Y
Flat Affect
Normal

Fig. 9. The detection flow based on the emotional status.

emotional status of patients. Finally, we follow the rule mentioned above to detect the Mania disorder.

On the other hand, to detect the depression disorder, we here adopt the predicted depressive level for analysis. As our depression module learns to predict the depressive level by the AVEC-14 dataset, which is annotated based on Beck Depression Inventory-II [19], we here also follow the same criteria to determine the depression severity in the detection stage. The details of BDI-II can be interpreted as follows: 0 13: indicates no or minimal depression, 14 19: indicates mild depression, 20 28: indicates moderate depression, 29 45: indicates severe depression. If the patient's depressive level is higher than 20, we will give a depression disorder alert for notification because it may denote the moderate depression. The flow of the disorder detection based on the patient's depressive level is shown as follows:

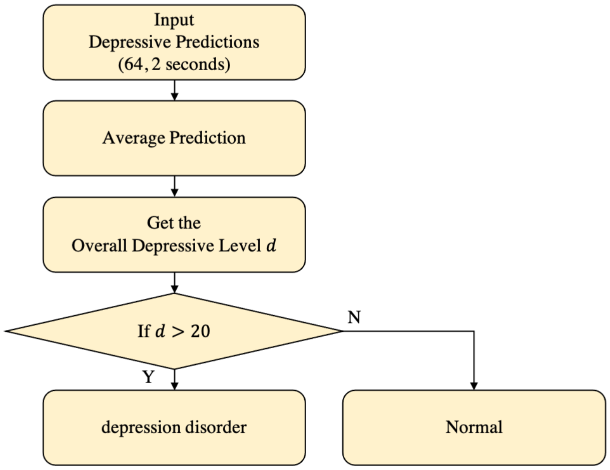

Fig. 10. The detection flow based on the emotional status.

In our implementation, we first average the predicted depressive levels to one scalar in order to represent the overall depressive level in the mentioned duration (2 seconds). Then, we check the value following the BDI-II criteria to detect the depression disorder.

4 Experiments

In this section, we will first introduce the datasets and the evaluation metrics which will need in our experiments, and then elaborate on the training details of our learning framework. After that, we provide various ablation studies to identify our design architecture. Finally, we demonstrate our experimental results and show the superiority of our performance.

4.1 Datasets & evaluation metrics

We evaluate our learning framework on two public datasets, including CAER [15] and AVEC 14 [18], both of which are the most well known and challenging. The evaluation metrics such as accuracy, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) will be mentioned in the end.

4.1.1 Datasets

CAER is a collection of video-clips from TV shows with 7 discrete emotion annotations, including Anger, Disgust, Fear, Happy, Sadness, Surprise, and Neutral. The dataset involves 13201 clips and about 1.1M frames for training and testing. In CAER benchmark, the videos range from short (around 30 frames) to longer clips (more than 120 frames). The average length of the image sequence is 90 frames. Besides, they extract about 70K static images from video data to form an image subset, called CAER-S. The details of each category are summarized in Table 2.

AVEC 14 depression dataset is proposed for the Audio/Visual Emotion Challenge 2014 [18], where a subset of the audiovisual depressive language corpus (AViD-Corpus) is used for the depression sub-challenge. In this dataset, the video is recorded in German language and can be classified into Freeform and Northwind scenario. The former is an uncontrolled response of participants to several questions, such as "What is your favorite dish?" or "Discuss a sad childhood memory.". The latter is in a controlled environment, where participants read aloud an excerpt of the fable "The North Wind and the Sun". The level of the depression is labeled with a single value per video using a standardized self-assessed subjective depression questionnaire, namely, the Beck Depression Inventory-II (BDI-II [19]), as shown in Table. 1.

4.1.2 Evaluation metrics

we evaluate our method by using the accuracy for emotion recognition, and Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) [30] for depression estimation. For emotion recognition, a classification task, we use the accuracy to estimate the model performance. The accuracy is defined via the ground-truth labels and the model predictions. Specifically, if one predicted class is the same as its ground-truth label, it is a correct prediction, otherwise, it is an incorrect prediction. By dividing the total number of the correct predictions by the total number of predictions, we can get the accuracy of the current model for the classification task. The accuracy can be expressed as follows:

$$accuracy = \frac{\#ofcorrectpredictions}{\#ofpredictions} \tag{17}$$

In addition, for depression estimation, a regression task, we follow the standard metric, including MAE and RMSE, to measure the overall performance. Different from the CMC metric, MAE and RMSE aim to measure how close the

Category	# of clips	# of frames	%
Anger	1,628	139,681	12.33
Disgust	719	59,630	5.44
Fear	514	46,441	3.89
Happy	2,726	219,377	20.64
Neutral	4,579	377,276	34.69
Sadness	1,473	138,599	11.16
Surprise	1,562	126,873	11.83
Total	13,201	1,107,877	100

TABLE 2
Amount of video clips and frames in each category on CAER.

predicted score is from the ground truth value. Formally, MAE and RMSE can be expressed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{18}$$

$$RMSE = \left(\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2\right)^{\frac{1}{2}} \tag{19}$$

where N is the number of samples, y_i and $\hat{y}_i$ respectively denote the ground truth and the predicted value of the i-th sample.

4.2 Implementation details

For our multi-task learning framework, we adopt CNNs with 5 convolutional layers and the proposed CMGCN as the backbone network. Note that we train the neural network from scratch, which is with learning rate initialized as 10^{-3} and dropped by a factor of 10 every 40 epochs. We follow the same setting in [15] to train our multi-task framework. We first resize the context image into 128×171 and randomly crop it into 112×112. For the facial image, we resize the image into 112×112. Then, we use a PK batch sampler [31] (P classes and K instances/class) to construct a mini-batch. For each mini-batch, there are 7 identities and 10 images per identity. Also, we apply the common data augmentation strategies, including padding, random crops, horizontal flips, to avoid the overfitting problems.

4.3 Ablation studies

To verify the effectiveness of the proposed multi-task learning framework, we conduct a series of ablation studies for it in the following subsection. We first investigate the influence of our CMGCN, and then discuss the influence of the proposed task-aware objective functions, including Density loss and Distributed loss. After that, we demonstrate the effectiveness of our Emotion Passer for knowledge transfer.

4.3.1 The influence of CMGCN

First of all, we investigate our CMGCN with different modalities in detail, and the results are shown in Table 3. For a fair comparison, we set Density loss for emotion recognition and Distributed loss for depression estimation. In addition, we employ Emotion Passer to transfer the prior knowledge on emotion to depression model. For baseline model, we directly concatenate the features from different modalities and achieve 70.09% in accuracy on CAER and 12.40/14.80 in MAE/RMSE on AVEC 14. Face indicates that we only apply CMGCN on the face modality alone; Context indicates that we only apply CMGCN on the context modality alone; Face and Context together denotes that we adopt CMGCN to integrate the visual features from different modalities.

As we can see, the Face can be regarded as an attention mechanism concerning face modality, where it focuses on linking relevant pixels and dropping the background pixels around face. Although the context information is not involving for training, our CMGCN can improve the performance by 9.32% in accuracy on CAER and 2.66/1.61 in MAE/RMSE on AVEC 14. For the Context group, the effect of CMGCN is similar to the Face group, but it affects the context modality only. Since the context involves the entire visual cues, it can avoid the wrong face provided from face modality and result in the best performance on CAER with 89.62% in accuracy. In contrast, for AVEC 14, because the scene is in the laboratory, the environment is not a serious problem, such as illumination and noise, to detect the correct face. Thus, we can easily capture the correct face from the given image so only considering the context modality cannot significantly improve the performance in this benchmark. By integrating the features from different modalities, our CMGCN can achieve the best performance on AVEC 14 with 6.82/8.5 in MAE/RMSE and significantly improve the performance of Baseline by 17.14% in accuracy on CAER.

Modality	CAER	AVEC 14	
	Accuracy (%)	MAE	RMSE
Baseline	70.09	12.4062	14.8008
Face	79.41	9.7376	13.1881
Context	89.62	9.2008	11.8333
Face + Context	87.23	6.8206	8.5078

TABLE 3
Comparison of our CMGCN in different modalities.

Moreover, as our core idea is to build the sparse graph, we here discuss the influence of the sampling scheme for yielding a sparse graph for GCN embedding. As we can see from Table 4, $\mathcal{G}_{full}$ denotes a graph where we do not impose the sampling scheme to connect/drop relevant/irrelevant pixels, $\mathcal{G}_{topk}$ denotes that where we select the top 5 high entries from another modality for each pixel, $\mathcal{G}_{\epsilon}$ denotes that where we adopt the epsilon ball where threshold is 0.5 to link the relevant pixels, and $\mathcal{G}_{bernoulli}$ denotes that where we introduce the Bernoulli sampling scheme to link pixels based on their similarity. Here, we adopt CMGCN to integrate the visual features from the face and context modalities for a fair comparison. $\mathcal{G}_{full}$ can yield 85.33% in accuracy on CAER and 10.80/13.25 in MAE/RMSE on AVEC 14. It means linking the relevant entries is a crux for integrating the visual features from different modalities. However, for AVEC 14, because the cross-modality graph involves too many irrelevant features, such as background information, it cannot improve the performance well. If we impose the sampling scheme to constitute the sparse graph, such as $\mathcal{G}_{topk}$, $\mathcal{G}_{\epsilon}$, and $\mathcal{G}_{bernoulli}$, we can achieve a better performance, especially for AVEC 14. Because $\mathcal{G}_{topk}$ connects all relevant entries for all pixels, it will perform the worst on CAER since it considers other irrelevant information. Comparing $\mathcal{G}_{bernoulli}$ with $\mathcal{G}_{\epsilon}$, Bernoulli sampling can achieve the better performance due to its sampling nature. Such sampling scheme will not completely ignore lower

Modality	CAER	AVEC 14	
	Accuracy (%)	MAE	RMSE
Baseline	70.09	12.4062	14.8008
$\mathcal{G}_{full}$	85.33	10.8065	13.2533
$\mathcal{G}_{topk}$	84.67	9.0662	11.1652
$\mathcal{G}_{\epsilon}$	86.36	7.6340	9.4368
$\mathcal{G}_{bernoulli}$	87.23	6.8206	8.5078

TABLE 4
Comparison of our CMGCN with different sampling schemes.

Modality	CAER Accuracy (%)
Baseline	81.75
Baseline + cL_{hard_tri}	85.26
Baseline + $cL_{density}$	87.23

TABLE 5
Comparison of our Density loss.

Modality	Head Type	AVEC 14	
		MAE	RMSE
cL_{MSE}	–	8.0123	9.7324
cL_{Distri}	–	7.9850	9.6507
$cL_{MSE} + cL_{Distri}$	Separate	7.4287	9.0254
$cL_{MSE} + cL_{Distri}$	Joint	6.8206	8.5078

TABLE 6
Comparison of our Distributed loss.

similarity elements; thus, the model can learn more kinds of graph features (messy or high-quality) and understand how to intensify the connections between relevant features and suppress other irrelevant ones.

4.3.2 The influence of density loss and distributed loss

To validate the effectiveness of our Density loss, we here adopt CMGCN to integrate the visual features from different modalities. As we can see from Table 5, for the Baseline group, we consider the A-Softmax loss [27] with $s = 20$ to train the entire model. cL_{hard_tri} and $cL_{density}$ respectively indicate the Hard Triplet loss [32] and our Density loss. With the pairwise metric learning, it can greatly improve the performance. Since the A-Softmax loss imposes an approximated weight matrix to guide representations in the embedding space, it may not well promote the intra-class compactness during the early epochs. By the margin constraint of cL_{hard_tri}, the intra-class compactness can be enhanced the with 3.51% improvement in accuracy on CAER. However, cL_{hard_tri} may easily cause ambiguous optimization results [26]. To alleviate this drawback, our Density considers the strict boundary for intra-class and inter-class pairs relevant to orthogonal relation ($b_p = 1 - b_n$). Each intra-class/inter-class pairs are encouraged to be greater/less than the b_p/b_n in the embedding space. Thus, it will not lead to the ambiguous optimization results, which are learned via ($s_p - s_n > m$). By doing so, it can greatly avoid ambiguity. Further, as our density loss is designed with the embedding density, it can adaptively emphasize the loss of each class, and consequently forming a useful regularization for embedding learning. By providing a more comprehensive metric, our Density loss can achieve the best performance.

To evaluate the effectiveness of our Distributed Loss, we here adopt CMGCN to integrate the visual features from different modalities. As we can see from Table 6, cL_{MSE} and cL_{Distri} respectively denote the MSE Loss and the proposed Distributed Loss. Further, we consider two types of head to synergize these two loss functions. The Separate group denotes that we adopt two isolated MLP to separately learn with cL_{MSE} and cL_{Distri}. The Joint group indicates that we adopt a single MLP layer to jointly learn with cL_{MSE} and cL_{Distri}.

From Table 6, our cL_{Distri} can perform better than cL_{MSE}. Due to dividing the regression levels into several bins, we can take the classification viewpoint to emphasize the intensity of the loss value, thus resulting in better performance in comparison with cL_{MSE}. By the two isolated MLP layers, we can take a simple approach to combine these two losses; however, the performance is only slightly improved. Because these MLP layers undergoes separate learning, it may result in ambiguous results for prediction, *e.g.*, cL_{MSE} stream predicts the large value while cL_{Distri} predicts the bin located at the low level. To mitigate this ambiguity, we follow YOLO to design a joint head for prediction; thus, the performance can be greatly improved due to the joint learning manner.

4.3.3 The influence of Emotion Passer

To validate the effectiveness of our Emotion Passer for knowledge transfer, we here train the emotion model and depression model with two scenarios, including learning without Emotion Passer and learning with Emotion Passer. As we can see from Table 7, if the depression model is randomly initialized, it will not be easy to converge and result in the worst performance. By our Emotion Passer, since the prior knowledge on emotion is smoothly transferred to the depression model, it can effectively enhance the training procedure and result in better performance.

4.4 In comparison with state-of-the-arts (SOTA) works

4.4.1 The result on CAER

For a fair comparison, we follow the same scenarios in [15] to demonstrate the effectiveness of the proposed approach. From Table 8, we can see that the deeper backbones, such as ResNet [33], can achieve the better performance than AlexNet [34] because it can extract more discriminative features. CAER-Net-S [15] masks out the human face from the given context image to seek more emotional features for embedding learning. Although this attention mechanism can improve the model performance, it mainly relies on the correct face detected by the face alignment model. Besides, their fusion network considers two isolated MLP (Multi-Layer Perceptron) layers to respectively predict the weights for each modality, this fusion mechanism may lead to improper results when the non-target human face is given in the face stream. When the face is incorrect, the fusion weights cannot represent the importance of the face feature. Other methods [35] consider the case where temporal information is involved to construct robust representations; however, it often requires more computational costs and memory consumptions. By 3D CNNs, CAER-Net [15] can better fuse the temporal information and achieve advanced performance with 77.04% on accuracy. Unlike the above methods, our CMGCN exploits the sampling scheme to constitute a sparse graph to describe the correlations of relevant pixels, and then adopts the graph embedding to

Modality	AVEC 14	
	MAE	RMSE
Ours w/o Emotion Passer	8.9545	11.1884
Ours w/ Emotion Passer	6.8206	8.5078

TABLE 7
Comparison of our Emotion Passer.

Method	Data type	Modality	CAER Accuracy (%)
ImageNet-AlexNet [34]	Image	Face + Context	47.36
ImageNet-ResNet [33]	Image	Face + Context	57.33
Fine-tuned-AlexNet [34]	Image	Face + Context	61.73
Fine-tuned-ResNet [33]	Image	Face + Context	68.46
CAER-Net-S [15]	Image	Face	70.09
	Image	Context	65.65
	Image	Face + Context	73.51
Sports-1M-C3D [35]	Video	Face + Context	66.38
Fine-tuned C3D [35]	Video	Face + Context	71.02
CAER-Net [15]	Video	Face	74.13
	Video	Context	75.57
	Video	Face + Context	77.04
Ours	Video	Face + Context	**87.23**

TABLE 8
Emotion Recognition: Comparison of the SOTA on CAER.

yield the final representation. With the sparse graph, the irrelevant information will be significantly dropped, yielding a better representation. Moreover, the proposed Density Loss can lead to better model convergence via comprehensive criteria relevant to orthogonal property and embedding density. Compared with other SOTA methods, our emotion model can significantly surpass other SOTA methods with a clear margin, over 10% in accuracy. Notably, unlike other methods adopting the deeper backbone, our emotion model only adopts the 2D CNNs with 5 layers to extract the visual features from dual modalities .

4.4.2 The result on AVEC 14

We follow the same scenario in [18], [22] evaluate our depression model on AVEC 14. The quantitative results are shown in Table 9, our depression model outperforms all other SOTA methods with 6.82/8.5 on MAE/RMSE. As we can see, other methods [18], [22], [36], [37], [38], [39], [40] focus on facial analysis; thus, their performance is limited due to the ambiguity caused by the face information. Departing from the facial analysis, our approach additionally models the context features for embedding learning; thus, we can yield a representation with more comprehensive signals and achieve the best performance. Compared with the most advanced learning-based approach, RNN-C3D [40], our method can surpass it by 0.4/0.7 in MAE/RMSE. On the other hand, as the hand-crafted-based methods mainly rely on some assumptions to extract visual features, they are not easy to be generalized to unseen data and often perform worse than the learning-based methods [22], [39], [40].

Method	Data type	Modality	AVEC 14	
			MAE	RMSE
Baseline [18]	Video	Face	8.86	10.86
UUIMSidorov [36]	Video	Face	11.20	13.87
InaoeBuap [37]	Video	Face	9.35	11.91
Brunel [38]	Video	Face	8.44	10.50
BU-CMPE [39]	Video	Face	7.96	9.97
DCNN [22]	Video	Face	7.47	9.55
RNN-C3D [40]	Video	Face	7.22	9.20
Ours	Video	Face + Context	6.82	8.50

TABLE 9
Depression Estimation: Comparison of the SOTA on AVEC 14.

4.5 The experiments of mental disorder detection

To validate the effectiveness of our mental disorder detection system, we collect the video data of schizophrenia patients from the National Taiwan University Hospital (NTUH). Each video is recorded during the psychological counseling, where the psychological counselor will give a conversation with a patient, and is roughly longer than 10 minutes. As we mentioned earlier, Mania and Depression are important signs for doctors to understand the severity of schizophrenia patients. Thus, each video is annotated by two psychologists and is related to temporal annotations with two kinds of disorder classes, which are "Mania" and "Depression". Particularly, the annotated results are the consistent agreement of two psychologists. For example, if the patients with depression disorder in the given video, the psychologists will annotate the class, the start time, and the end time of depression disorder, as shown in Figure 11.

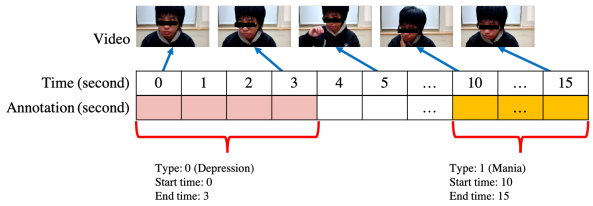

Fig. 11. An example about the temporal annotation.

To estimate the model performance, we here exploit the Mean Average Precision (mAP) to measure the difference between the prediction and the ground-truth. Specifically, if one image frame is located in a duration of a specific mental disorder, the class of this image frame will be annotated as the same as the specific mental disorder, such as mania or depression. According to the ground-truth annotation, we can define the True Positive (TP) and False Positive (FP) as follows: If one frame with the same class as the ground-truth, it is TP, otherwise, it is FP. After that, we can calculate the mAP of each video, and the formulation of mAP is as follows:

$$F(j, y_i, p_i) = \begin{cases} 1, & \text{if } y_i = j \text{ and } y_i = p_i \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

$$AP_j = \frac{1}{n_j} \sum_{i=1}^{n_j} F(j, y_i, p_i) \tag{21}$$

$$mAP = \frac{1}{C} \sum_{j=1}^{C} AP_j \tag{22}$$

where C and j denote the total number of classes in the given video and the label of j-th class, respectively; n_j is the number of frames corresponding to the given class j, and y_i, and p_i respectively denote the ground-truth and the predicted disorder class of i-th frame.

Currently, we get two cases of schizophrenia patients from NTUH with complete temporal annotations. Thus, we here adopt our multi-task model and the proposed detection algorithm for the collected cases and report the AP and mAP of each class of each case in detail. The details of the proposed multi-task learning framework and mental disorder detection algorithm are shown in Section 3.2. As we can see from Table 10, the results show that our algorithm can successfully detect Mania disorder and Depression disorder to a certain degree. Here, our approach can achieve 73.38 in mAP on the collected cases. For Mania disorder, because our detection algorithm mainly relies on the domain knowledge to design the good rules for detection, it may result in worse performance than the depression disorder by 2.76 in AP on case 1 and 7.22 in AP on case 2. In addition, we also show the actual prediction of our system. Since the number of image frames is quite large, we here only show the predicted results in a short duration.

Case	Class	AP	mAP
	Mania	72.56	–
Case 1	Depression	75.32	–
	Overall	–	73.94
	Mania	69.21	–
Case 1	Depression	76.43	–
	Overall	–	72.82
Overall	–	–	73.38

TABLE 10
The performance of our Mental Disorder Detection System.

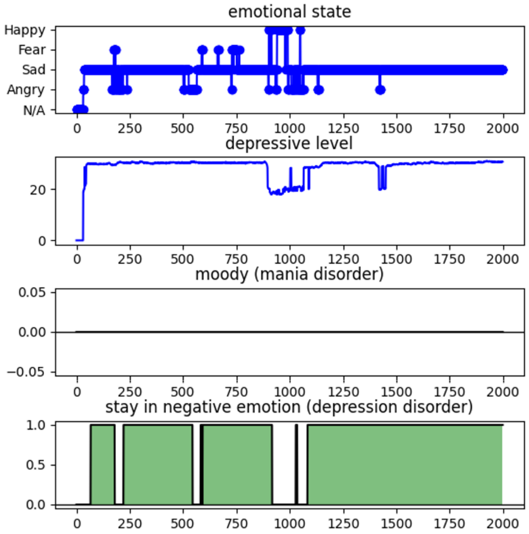

Fig. 12. The detection results in a short duration for case 1.

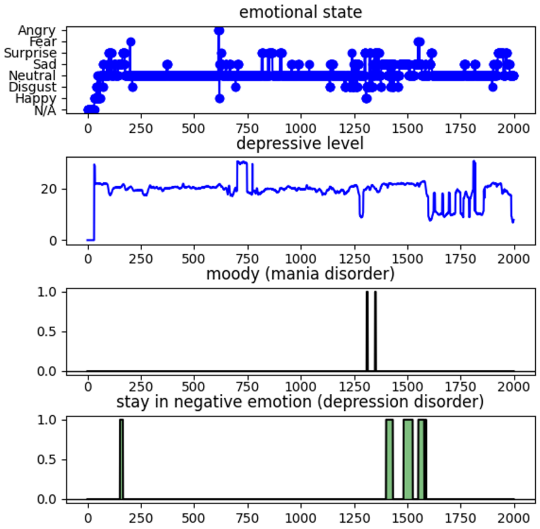

Fig. 13. The detection results in a short duration for case 2.

5 Conclusion

In this paper, we propose a novel multi-task learning framework to detect the mental disorder of schizophrenia patients. By both emotion recognition and depression estimation, our system can infer the mental state of schizophrenia patients, and then it can detect the mental disorders about the mood aspect by analyzing the abnormal patterns of the predicted mental state. To precisely infer the emotional status and depressive level, we design a fusion network, namely Cross-Modality Graph Convolutional Networks (CMGCN), to integrate the visual features from different modalities. Concretely, our CMGCN adopts an affinity graph to describe the correlations between different modalities, then employs the sampling scheme to build the sparse graph. Since the sampling scheme can connect the relevant pixel pairs of different modalities with similar semantic meaning and neglect irrelevant ones, we constitute a representation with more comprehensive emotion signals and consequently result in better performance. In addition, for model convergence of each task, we design task-aware objective functions to form a useful regularization for embedding learning. For emotion recognition, a classification task, we propose Density loss for metric learning with comprehensive criteria relevant to the embedding density and orthogonal property. Based on these two critical attributes, we can form a useful regularization for embedding learning further resulting in better performance. On the other hand, we exploit the classification viewpoint to form Distributed loss for depression estimation, a regression task. By dividing the regression level into several bins, we can emphasize the intensity of the loss in a more meticulous way compared with MSE loss and achieve better performance. Observe depression is a disorder characterized by a low emotional status; thus, we propose a knowledge transfer scheme, namely, Emotion Passer. For each mini-batch, our Emotion Passer exploits EMA to smoothly transfer the emotion prior knowledge to the depression model. Thus, we can greatly promote training efficiency compared with other transfer learning strategies. Finally, with the well design multi-task learning framework, we can accurately record the mental state of patients and thus enforce the proposed disorder detection algorithm to detect the abnormal patterns based on cognitive science. The comprehensive ablation studies consolidate the effectiveness of our method, CMGCN, Density loss, Distributed loss, and Emotion Passer. In the experimental results, our method achieves 87.23% in accuracy on CAER and 6.82/8.50 in MAE/RMSE on AVEC 14, which outperforms all advanced SOTA methods. In addition, for mental disorder detection, our system can achieve 73.38 in mAP on the collected patient' cases from NTUH. This shows that our system can detect the mental disorder of schizophrenia patients to a certain degree. In future work, we plan to combine with the speech perspective to accomplish a more comprehensive representation and learn the model on the dataset collected by NTUH to realize an end-to-end learning manner.

Acknowledgments

The authors would like to thank...

References

[1] A. Vita, S. Barlati, L. D. Peri, G. Deste, and E. Sacchetti, "Schizophrenia," *The Lancet*, vol. 388, 2016.

[2] G. Arbanas, "Diagnostic and statistical manual of mental disorders (dsm-5)," *Alcoholism and psychiatry research*, vol. 51, pp. 61–64, 2015.

[3] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, pp. 169–200, 1992.

[4] E. K. Gray and D. Watson, "Assessing positive and negative affect via self-report." 2007.

[5] X. Feng, "Facial expression recognition based on local binary patterns and coarse-to-fine classification," in *The Fourth International Conference onComputer and Information Technology, 2004. CIT '04.*, 2004, pp. 178–183.

[6] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Transactions on Cybernetics*, vol. 45, pp. 1499–1510, 2015.

[7] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[9] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, pp. 2439–2450, 2019.

[10] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition," *IEEE Transactions on Image Processing*, vol. 24, pp. 189–204, 2015.

[11] M. Singh, B. B. Naib, and A. K. Goel, "Facial emotion detection using action units," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 2020, pp. 1037–1041.

[12] K. Schindler, L. Gool, and B. Gelder, "Recognizing emotions expressed by body pose: A biologically inspired neural model," *Neural networks : the official journal of the International Neural Network Society*, vol. 21 9, pp. 1238–46, 2008.

[13] C. Chen, Z. Wu, and Y.-G. Jiang, "Emotion in context: Deep semantic feature fusion for video emotion recognition," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016.

[14] R. Kosti, J. M. Alvarez, A. Recasens, and À. Lapedriza, "Emotion recognition in context," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1960–1968.

[15] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[16] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using frege's principle," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[17] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013.

[18] M. Valstar, B. Schuller, K. Smith, T. R. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th ACM international workshop on Audio/visual emotion challenge*, 2014.

[19] A. Beck, R. Steer, R. Ball, and W. Ranieri, "Comparison of beck depression inventories -ia and -ii in psychiatric outpatients." *Journal of personality assessment*, vol. 67 3, pp. 588–97, 1996.

[20] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals: a multimodal approach," 2013.

[21] L. Yang, D. Jiang, W. Han, and H. Sahli, "Dcnn and dnn based multi-modal depression recognition," *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 484–489, 2017.

[22] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Transactions on Affective Computing*, vol. 9, pp. 578–584, 2018.

[23] M. Ding, Y. Huo, J. Hu, and Z. Lu, "Deepinsight: Multi-task multi-scale deep learning for mental disorder diagnosis," in *BMVC*, 2018.

[24] L. Yang, X. Zhan, D. Chen, J. Yan, C. C. Loy, and D. Lin, "Learning to cluster faces on an affinity graph," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[25] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[26] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[27] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[29] E. Svetieva and M. G. Frank, "Empathy, emotion dysregulation, and enhanced microexpression recognition ability," *Motivation and Emotion*, vol. 40, pp. 309–320, 2016.

[30] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature," *Geoscientific Model Development*, vol. 7, pp. 1247–1250, 2014.

[31] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016.

[32] A. Hermans*, L. Beyer*, and B. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," *arXiv preprint arXiv:1703.07737*, 2017.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.

[35] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[36] M. Sidorov and W. Minker, "Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach," in *AVEC '14*, 2014.

[37] H. Espinosa, H. Escalante, L. Pineda, M. M. y Gómez, D. Pinto, and V. Reyes-Meza, "Fusing affective dimensions and audio-visual features from segmented video for depression recognition: Inaoe-buap's participation at avec'14 challenge," in *AVEC '14*, 2014.

[38] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic depression scale prediction using facial expression dynamics and regression," in *AVEC '14*, 2014.

[39] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble cca for continuous emotion prediction," in *AVEC '14*, 2014.

[40] M. A. Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.