國立臺灣大學電機資訊學院電機工程學系

碩士論文

Department of Electrical Engineering
College of Electrical Engineering and Computer Science
National Taiwan University

Master Thesis

透過深度視覺感知
偵測思覺失調症患者的精神異常

Mental Disorder Detection for Schizophrenia Patients via Deep Visual Perception

林炳彰

Bing-Jhang Lin

指導教授：傅立成 博士

Advisor: Li-Chen Fu, Ph.D.

中華民國 110 年 1 月

January, 2021

誌謝

碩士班這兩年的時光裡，我很感謝我所遇到的每個人，大家都是我的貴人，讓我不僅在學術研究方面有著顯著的成長。且對於分析與解決問題的能力也有十足的進步。首先，要感謝的是指導老師 傅立成教授。老師在學術研究上的熱忱和追求卓越的堅持都是我最好的效仿對象。此外，老師在待人處事上也提點了我許多，這些經驗都會是我在未來道路上的智慧與基石，謝謝老師。未來我也會遵循老師的指導，努力成為在社會上有貢獻的人。

再來，感謝中研院 劉庭祿老師。讓我參加實習的計畫，在每週的 meeting 上都能接觸最新的電腦視覺研究，使我可以增廣見聞、獲益良多。這些經歷都讓我對電腦視覺的領域有更加深刻的認識，謝謝老師。希望這次的投稿可以順利。

另外感謝 ACL 的所有夥伴！謝謝安陞學長在這段時間的幫助，透過與學長的討論加深我對於研究的基礎。其次，感謝這次實驗室的畢業好夥伴：靈風、睿庭、彥廷、耘志、瀟越，在口試前夕，彼此提供意見、互相幫忙，一起完成碩士班的最後挑戰！也感謝影像組的學弟：承軒、志淵、體淮，有了你們的支援，讓我可以更專注的做研究。也感謝快樂好朋友：靈風、冠瑋、彥廷，這段時光來共同經歷了許多趣事。也謝謝實驗室助理替我們安排張羅各項事務。最後感謝家人與女朋友，在碩士班的過程中謝謝你們的支持！

在未來的路上，如果大家有需要我的地方且我又有能力的話，我一定大力相挺。

炳彰 January 31, 2021

中文摘要

現代社會中，思覺失調症為一種嚴重的精神疾病，會逐漸改變一個人的精神狀態並導致嚴重的社會成本。一般而言，患者通常無法正常表達自己的真實想法，且在行為上與常人有所不同。根據《精神疾病手冊》，我們察覺到思覺失調症為精神與情緒異常的組合，其中許多症狀皆與情緒和抑鬱高度相關。在本論文中，我們以視覺感知為基礎，對思覺失調症患者在表情與憂鬱上的表徵進行編碼，以推估患者的心理狀態。接著，根據思覺失調症的性質與認知神經學對於人類情緒上的研究，為覺失調症患者建立精神障礙檢測系統，以提供評估讓醫師了解病患的嚴重程度。

由於先前的研究主要透過面部分析以識別人類的情緒和抑鬱狀態，所以它們經常無法準確識別人類的情緒與憂鬱狀態並無法提供令人滿意的性能。主要原因為面部表情是一種不穩定的情緒表示，其容易因面部肌肉的運動而產生不穩定的情緒訊號進而導致不穩定的預測結果。另一方面，由於思覺失調症的性質，患者亦不能表達具有區別性的面部情緒信號。 因此，常規的面部表情識別系統不適合識別思覺失調症患者的情緒狀態和抑鬱水平。

為了克服這些限制，我們提出了跨模態圖卷積網絡，以有效地整合來面部與景況的視覺特徵。通過稀疏圖和圖卷積，我們可以鏈接在語意上相互關聯的視覺訊號，並忽略其他不相關的訊號，如背景像素。通過這種稀疏化的機制，我們可以有效地整合來面部與景況的情緒特徵，從而生成更可靠的深度表示。另一方面，由於我們所提出的模型涵蓋多任務範疇，所以對於各項任務，我們

提出了任務感知目標函數，以為每個任務實現更好的模型收斂。對於情感識別，我們提出了用於度量學習的密度損失，並採用與嵌入密度和正交相關的綜合標準。通過考慮這兩個關鍵屬性，我們可以為嵌入學習形成更強大的正則化。此外，對於抑鬱症的估計，我們採用分類觀點來形成回歸任務的分佈式損失。通過將回歸級別劃分為幾個分類，我們可以以一種細緻的方式強調損失值的影響。由於抑鬱症是一種精神障礙其特徵為持續性的低落情緒，因此我們設計了一種知識轉移方法，即情緒傳遞者，將情感先驗知識轉移到抑鬱症模型中。在每個迷你批處理中，我們採用指數移動平均方案來平滑地傳遞知識，來為看不見的數據實現穩健的模型。通過我們精心設計的多任務學習框架，我們可以準確地偵測病患的精神狀態，並基於所記錄的狀態設計精神障礙檢測算法。

為了驗證我們框架的有效性，我們在包括 CAER 和 AVEC 14 在內的多個基准上進行了一系列實驗。實驗結果表明，我們的方法在 CAER 上的準確度達到 87.23%，在 AVEC 14 上的 MAE / RMSE 上達到 6.82 / 8.50。優於所有先進的 SOTA 方法。此外，我們還將所設計的系統應用於台灣大學醫院所搜集之的思覺失調數據集上，並在所搜集之個案上達到 73.38 mAP。

關鍵字：圖卷積網路、情緒識別、憂鬱評估

ABSTRACT

Nowadays, schizophrenia is a mental illness that will progressively change a person's mental state and cause serious social problems. In principle, patients are unable to express their real thinking ordinarily, and their behaviors are often different from the normal people. According to the medical literature about mental illness, we observe that schizophrenia involves a range of problems with thinking, behavior, and emotions and its symptoms are highly correlated to the emotion and depression. Thus, in this thesis, we employ the visual perception as basis to encode the representations of the emotion and depression in order to infer the mental state of the schizophrenia patient. Then, we follow the nature of schizophrenia and the emotion of human behavior in cognitive science to realize a mental disorder detection system for schizophrenia in order to provide an assessment.

Since the previous studies mainly focus on facial analysis to recognize human emotion and depression, they frequently fail to provide satisfactory performance. Facial expressions are extremely unstable emotional signals. They often result in unstable prediction results due to facial muscle movements. On the other hand, due to the nature of schizophrenia patients, they hardly can express a discriminative facial expression. Thus, conventional facial expression recognition systems are not suitable to identify the emotional status and depressive level of schizophrenia patients.

To overcome these limitations, we propose Cross-Modality Graph Convolutional Network (CMGCN) to effectively integrate visual features from different modalities, including the face and context. With the sparse graph and graph

convolution, we can relate the relevant visual cues and drop other irrelevant ones. By doing so, we can effectively integrate the emotional features from the face and context modalities, and thus yield a more robust representation for both emotion recognition and depression estimation. On the other hand, we propose task-aware objective functions to achieve better model convergence for each task. For emotion recognition, we propose Density Loss for metric learning with comprehensive criteria relevant to the density of each class in the deep embedding space to form a powerful regularization to embed learning. In addition, for depression estimation, we take the classification viewpoint to form Distributed Loss for the regression task. By dividing the regressive level into several bins, we can emphasize the influence of the loss value in a meticulous way. As depression is a disorder characterized by a low emotional status, we design a knowledge transfer approach, namely, Emotion Passer, to transfer the prior knowledge on emotion to the depression model. In each training iteration, we exploit the exponential moving average scheme to smoothly pass the knowledge and realize a robust model for the unseen data. By the well designed multi-task learning framework, we can recognize the emotion and estimate the depression of patients and then achieve a more accurate mental disorder detection.

To verify the effectiveness of our learning framework, we perform a series of experiments on several benchmark datasets, including CAER and AVEC 14. The experimental results show that our method achieves 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 methods. In addition, we apply our system on the data about schizophrenia patients in National Taiwan University Hospital and achieve 73.38 in mAP.

Keywords: Graph Convolutional Networks, Emotion Recognition, Depression Estimation

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

In this chapter, we first describe the research motivation and an introduction of this thesis in Section 1.1, and then elaborate on the literature review in Section 1.2. After that, we highlight the contributions of this thesis in Section 1.3. Finally, we conclude with the organization of this thesis in Section 1.4.

1.1 Motivation

Schizophrenia [1] is a mental illness that will progressively change a person's mental state and cause serious social problems. In principle, patients are unable to express their real thinking ordinarily, and their behaviors are often different from the normal people. Schizophrenia involves a range of problems with thinking, behavior, and emotions. According to the medical literature about mental illness [2, 3], schizophrenia is defined as a spectrum with several dimensions to describe the degree of mental disorder. Formally, the psychotic disorder can be described as five core characteristics, including delusions, hallucinations, disorganized speech, extremely disorganized or abnormal motor behavior, and negative symptoms. Specifically, the first three characteristics are the major properties to diagnosis schizophrenia, and the rest of the characteristics are auxiliary properties for diagnosis. During the psychotic period, even though the patient's mood is not discriminative enough, doctors can still feel the mood from the patient during the psychological counseling and infer the severity of mental disorders further. Therefore, in this thesis, our main goal is to provide an assessment for schizophrenia patients rather than diagnosis of the schizophrenia. 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. Besides, the mood aspect is one of the important references for

psychologists to understand the severity of schizophrenia patients. Specifically, 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.

The previous studies [4-6] assume that the human face can provide the most discriminative emotional signals, and hence have already done extensive discussions based on facial analysis. Most approaches recognize human's emotion based on facial expression analysis [4-8]. Some others employ the so-called facial action encoding system to analyze face movements in order to recognize human's facial expression [9, 10]. Due to the significant visual appearance changes, encoding the discriminative feature from the given facial image is the crux for recognition. Over the last decade, conventional methods extract the hand-crafted features, such as SIFT [11] or HOG [12]. However, these methods require some domain knowledge and aren't easy to be generalized to real-world scenarios. Instead of designing certain strategies to extract specific features, recent deep convolutional neural networks (CNNs) based approaches have made significant progress by learning the data distribution from the given dataset.

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 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 [13, 14] 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 solve these limitations of facial analysis, it is necessary to consider the context information to realize an accurate emotion recognition model. Furthermore, certain symptoms of schizophrenia are highly related to depression [1-3]. In contrast to schizophrenia patients without depression, patients with depression yield less favorable therapy courses and poorer outcomes.

We are thus motivated to design a multi-task learning framework to realize a mental disorder detection system for schizophrenia patients via emotion recognition and depression estimation. To tackle the shortcomings of facial analysis, we design Cross-Modality Graph Convolutional Networks (CMGCN) to effectively integrate the visual cues from different modalities, including the face and context. Specifically, in our CMGCN, we exploit an affinity graph to describe the pixel-wise correlations between different modalities and then introduce a sampling scheme to construct a sparse graph for graph convolution. As our sampling scheme is designed to connect those pixel pairs with similar semantic meaning, that is emotional class or depressive level, and to drop other semantic irrelevant pairs, such as background pixel pairs. Thus, our CMGCN can greatly reduce the irrelevant information, that is background pixels and result in a more robust representation for recognition or 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, our Distributed Loss can emphasize the influence of the loss in a meticulous way. Furthermore, as depression is a disorder characterized by a low emotional status, we present a knowledge transfer strategy, namely, Emotion Passer, to effectively pass the prior knowledge on emotion to the depression model. In each training iteration, our Emotion Passer 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. It has been demonstrated through extensive experiments that the performance of our multi-task learning framework can surpass those of the state-of-the-art (SOTA) works in each task. Finally, for the mental disorder detection, we follow the observation about human emotion in cognitive science and the nature of schizophrenia to design an algorithm to detect the mental disorders, *i.e.*, *Mania* and *Depression*.

1.2 Literature Review

We first give a brief review of the emotion recognition in Section 1.2.1, and then discuss the algorithm of depression estimation in Section 1.2.2.

1.2.1 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, the previous studies [4-10] 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*. [15] adopted the body pose to identify six emotion categories. Chen *et al*. [16] proposed a context fusion network to recognize human emotion by integrating events, objects, and scenes. Kosti *et al*. [17] 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*. [18] presented a two-stream architecture followed by a fusion network for CAER, as shown in Figure 1-1. 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*. [19] 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.

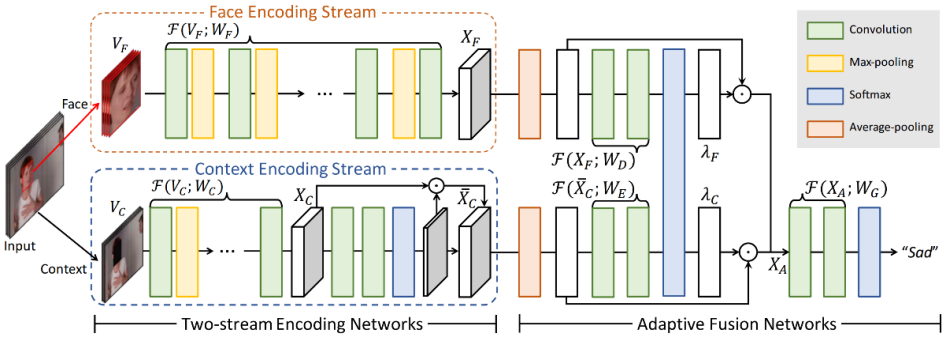


Figure 2. Network configuration of CAER-Net, consisting of two-stream encoding networks and adaptive fusion networks.

Figure 1-1: An overview of CAER-Net [18].

1.2.2 Depression Estimation

According to the machine learning perspective, depression estimation is essentially a regression problem. In the public benchmark datasets, such as AVEC 2013 [20] and AVEC 2014 [21], the depression level is categorized based on Beck Depression Inventory-II (BDI-II) [22], as shown in Table 1-1.

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 studies employed methods to extract hand-crafted visual features. The baseline approach in AVEC 2013 [20] adopted Local Phase Quantization (LPQ) followed by a regressor for learning and prediction. Unlike the single hand-crafted descriptor, Cummins *et al*. [23]

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

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

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*. [24] 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*. [25] 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*. [26] 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.

1.3 Contributions

In this thesis, we design a mental disorder detection system in order to provide an assessment to schizophrenia patients. Our system consists of two phases, including learning and detection. For the learning phase, we draw on the public benchmark datasets about emotion recognition and depression estimation to design a multi-task learning framework, which allows us to learn a multi-task model to precisely infer the mental state of schizophrenia patients. On the other hand, for detection phase, we follow the

characteristics about schizophrenia patients and the observation of the emotion in cognitive science to detect mental disorders about the mood aspact of schizophrenia patients. Our main contributions can be characterized as follows.

I. To better integrate the visual cues from different modalities, such as the face and context, we design a novel fusion network, namely Cross-Modality Graph Convolutional Networks (CMGCN). To leverage the property of a graph, we build an affinity graph to describe the pixel-wise correlations between different modalities, that are the face and context, and then introduce a sampling scheme to intensify the sparsity of the given graph. By doing so, it can allow us to successfully connect those pixel pairs with similar semantic meaning and greatly suppress the interactions between irrelevant pixel pairs. Thus, it can result in a robust representation for emotion recognition or depression estimation.

II. To realize better model convergence for our multi-task learning, we propose the task-aware objective functions to learn the robust embedding for each task. Specifically, for emotion recognition, we propose Density Loss for metric learning with more comprehensive criteria concerning embedding density (*the density of each class in the embedding space*) and the orthogonal relation (*the regularization of inter-class pairs*), which facilitates us to form a powerful regularization for embedding learning. On the other hand, for depression estimation, we exploit a classification viewpoint to form Distributed Loss to emphasize the loss value meticulously. By synergizing with MSE Loss, it can learn a robust regressor, further yielding better performance.

III. Following the natural relation between depression and emotion, we propose Emotion Passer to effectively transfer the prior knowledge on emotion to the

depression model. We draw on the Exponential Moving Average (EMA) mechanism to smoothly adjust the influence of the emotion model. In this way, the training procedure can be greatly reduced and achieve a robust model for depression estimation.

IV. According to the medical literature about of mental illness and the nature of schizophrenia, we propose an algorithm in detecting the mental disorders about the mood aspect to provide an assessment for schizophrenia patients. We first employ our multi-task model to infer the real-time mental status, including emotional status and depressive level. By recording the entire mental state of the patient, we adopt the sliding window scheme to detect the abnormal mental patterns.

1.4 Thesis Organization

In Chapter 1, we first present the motivation of this thesis, and then go through the history of some related works. After that, we briefly elaborate on the contributions, including the multi-task learning framework and the detection algorithm. Finally, we conclude with the organization of this thesis.

In Chapter 2 , we build up some prerequisite knowledge related to our research. Particularly, we present some background knowledge adopted in our research. First of all, we introduce Convolutional Neural Networks (CNNs), including the concept, operation, and classic algorithms. Then, we describe the operation of Graph Convolutional Network (GCN). After that, we elaborate on the metric learning techniques employed for model convergence. Finally, we illustrate the concept of Transfer Learning and the common strategies for transferring prior knowledge.

In Chapter 3, we first introduce the limitation of facial analysis, and then elaborate

on the proposed Cross-Modality Graph Convolutional Network (CMGCN) to integrate the visual cues from different modalities, including the face and context. After integrating the representation, we design the task-aware objective function to realize better model convergence. With the comprehensive criteria, including embedding density (*the density of each class in the embedding space*) and the orthogonal relation (*the regularization of inter-class pairs*), the proposed Density Loss can form a useful regularization for embedding learning and accomplish a robust model for emotion recognition. On the other hand, for depression estimation, the proposed Distributed Loss exploits the classification viewpoint to emphasize the intensity of the loss value. By doing so, we can train the regressor in a meticulous way. In addition, we further synergize Distributed Loss and MSE Loss with a joint learning manner to result in a robust regressor. Moreover, we propose a simple but efficient approach, namely, Emotion Passer, to transfer the prior knowledge. With Exponential Moving Average (EMA) scheme, we can take the smooth rather than protruding way to transfer the prior knowledge on emotions to the depression model. Finally, we introduce how we implement the mental disorder detection algorithm to provide an assessment to schizophrenia patients.

In Chapter 4, the experimental results demonstrate the effectiveness of the proposed multi-task learning framework. To ensure each module is useful, we then adopt a series of ablation studies for verification. Comparing with other state-of-the-art approaches, the proposed multi-task learning can overcome the limitation of facial analysis and achieve impressive results. 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.

In Chapter 5, we conclude contributions of this thesis and suggest some future works.

Chapter 2 Preliminaries

In this chapter, some prerequisite knowledge is introduced. First, we briefly give background information on deep neural networks in Section 2.1, including *convolutional neural network (CNN)* and *graph convolutional network (GCN)*. Second, we discuss the classic metric learning techniques for model convergence in Section 2.2, which are typically built on *classwise* and *pairwise* scenarios. Third, concepts of transfer learning are presented in Section 2.3.

2.1 Deep Neural Networks

In this section, we first elaborate on the concept of convolution and several classic CNN architectures in Section 2.1.1. Then, Section 2.1.2, we introduce the vanish gradient problem and show the solution scheme. Finally, we present some knowledge about GCN in Section 2.1.3, which is an important related work in this thesis.

2.1.1 Convolutional Neural Networks (CNNs)

The development of CNN can be traced back to AlexNext [27], which first applied the convolution operation to accomplish the task of image classification and realized the impressive performance in ImageNet [28]. Then, CNN has achieved a considerable amount of success in many computer vision topics, such as classification and retrieval [27, 29], object detection and segmentation [30, 31], and action recognition [32, 33] tasks.

For the convenience of presentation, we here adopt the image classification as an example to simplify the explanation. Convolution operation plays a fundamental role in the computer vision area. The essential concept of the convolution is to extract the low-level features from the given image, such as edges, corners, even textures. To effectively extract the specific features for inference, the previous studies based on the hand-crafted

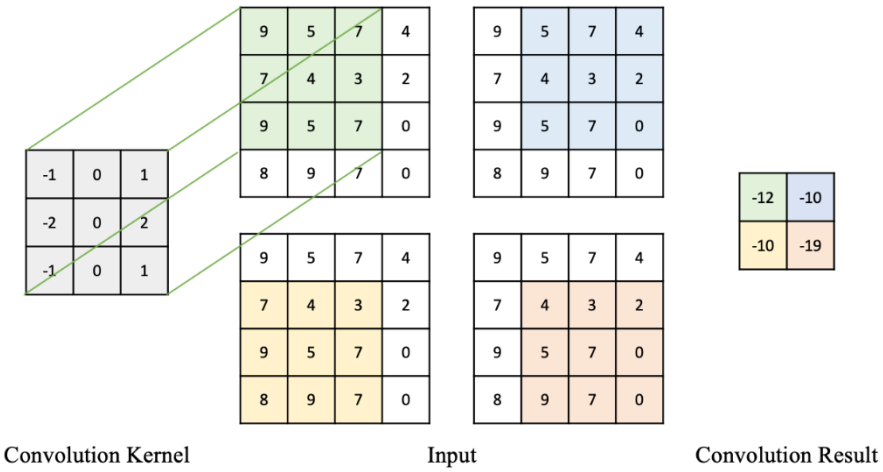


Figure 2-1. A standard 3 × 3 convolutional operation with stride 1.

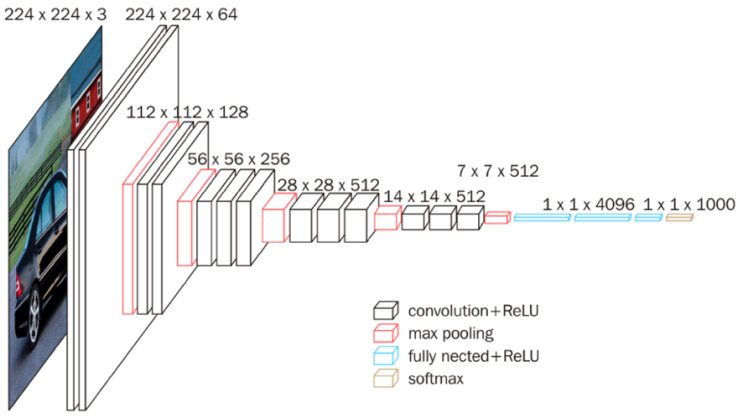
feature, such as SIFT [11] or HOG [12], require some domain knowledge to design a proper kernel of the convolution. However, it often requires many efforts of trial and error to search for the most suitable hyper-parameters for real-world scenarios.

Instead of using the hand-crafted features for design, CNN adopts the learnable parameters for the convolution kernel. With the backpropagation in the end-to-end learning manner, the learnable kernel can effectively fit the data distribution of the target domain, further extracting robust features. In Figure 2-1, we illustrate the basic concept of the convolution operation.

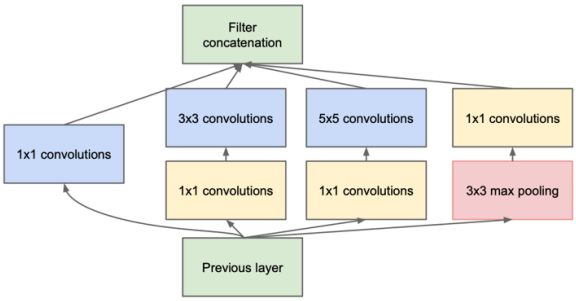
The typical Deep Neural Networks (DNN), also called Multi-Layer Perceptron (MLP), takes a node-to-node learning manner via several flatten layers to fit the data distribution of the target domain. Due to the huge parameters in the fully-connected layers, it often requires many computational costs for operation. By contrast, the convolutional layer of CNN exploits the kernel to extract features from the given image, which takes spatial information into consideration. By sliding the kernel across the image, it can take the shared-weights manner to capture the desired information; thus, CNN can greatly reduce the computation costs and realize better performance in many tasks relevant to the

computer vision. As a matter of fact, each convolutional layer typically consists of multiple kernels to extract various features from the given images. Zeiler *et al*. [29] introduce a way of deconvolution to visualize the response of feature maps. Notably, the results show that shallow layers extract low-level features, such as edges, corners, and colors, whereas the deeper layers are responsible for generating high-level features.

In 2014, Simonyan *et al*. proposed VGG [34], which is a deeper CNN compared with AlexNet [27], and achieve the runner-up of ImageNet in that year. Google research team also proposed a famous Inception Net [35] in the same year. Unlike VGG, Inception Net aims to widen the structure and became the championship in ImageNet 2014. Concretely, it worked by adopting several convolutional kernels with different sizes in a signal layer to extract multi-scale features. Figure 2-2 shows the configurations of VGG16 and Inception Net.



(a) VGG16 [34]



(b) Inception blocks [35]

Figure 2-2: The architecture of VGG16 and Inception Block.

2.1.2 Residual Network

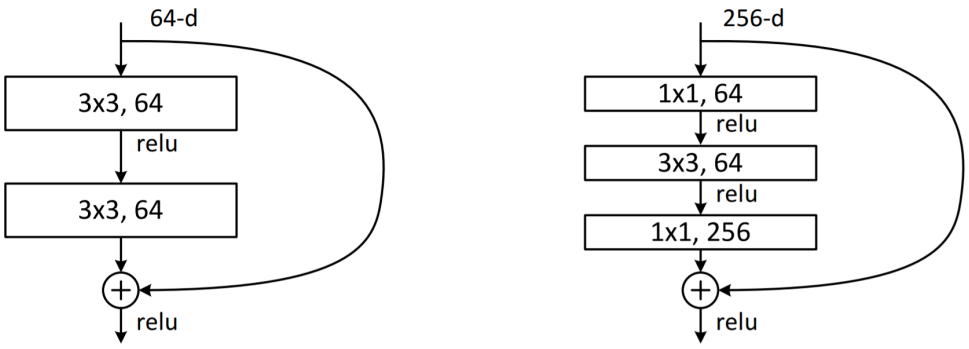


Figure 2-3: The difference between the flat convolutional layers and the residual connections [36].

According to Universal Approximation Theorem (UAT) [37], the feed-forward neural network has the ability to fit any non-linear mapping closely. Thus, researchers started to design a deeper network for strengthening the generalization and robustness. However, the vanishing gradient problem [38] occurs when we train a deep model with gradient-based learning algorithms. During the backpropagation stage, as the chain rules compute the gradients, the gradients of shallow layers may be shrinked too much, consequently resulting in a futile learning procedure.

To address this serious problem, He *et al*. [36]. proposed Residual Network (ResNet) with shortcut (skip) connection to carry more important information in the previous layer to the next layers. Because additional gradients will be provided by a residual path, it can significantly ease the burdens of many differentiations of deep networks. By doing this novel design, the deeper networks are again trainable. Figure 2-3 shows the shortcut connection and the difference between the flat convolutional layers and the residual connections. Formally, the residual connection can be expressed as:

$$output = F(x) + x \tag{2-1}$$

where $F(x)$ denotes a non-linear mapping for input signal x.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 2-4: The family of ResNet [36], each convolutional block involves several residual blocks.

The details of ResNet family are shown in Figure 2-4. Generally, ResNet can be divided into five stages, including four convolutional blocks and 1 classification layer. With the powerful generalization capability, it is often acted as the backbone of deep learning models to extract high-level representations.

2.1.3 Graph Convolutional Networks

Graph convolutional networks (GCN) have extensively discussed in recent years and shows its powerful capability in handling the non-Euclidean data structure (graphical data). As a matter of fact, the graph convolution is a general case of the convolution. The convolution operation takes the rigid space (Euclidean data structure) to seek the higher response pixels from the given image. Departing from the rigid space, we can cast the pixel into several nodes and adopt the graphical structure the represent the given instance, see Figure 2-5. By doing so, we can apply the graph convolution to seek the higher response node and thus yield a high-level graph feature.

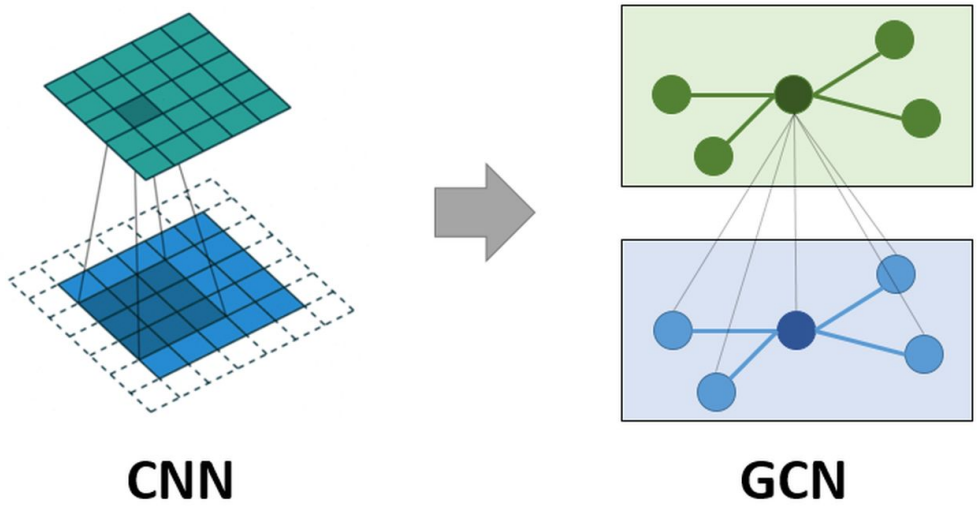


Figure 2-5: The difference between the convolution and the graph convolution.

Since the number of neighbors of graphical structure is various; thus, we have to define an edge set to describe the neighboring relations between each node. Typically, we adopt a graph (adjacent matrix) to describe the connectivity between each vertex. According to the [39], the conventional GCN can be expressed as:

$$\hat{A} = D^{-1}A, \qquad Z = X(\hat{A}XW) \tag{2-2}$$

where A denotes the adjacent matrix (graph) to describe the connectivity between vertices, and D represents the degree matrix of A, which is a diagonal matrix containing the information about the degree of each vertex, where $D_{ii} = \sum_j A_{ij}$. $\hat{A}$ is the normalized graph. X and W respectively indicate the given feature and the embedding weight matrix. $\sigma(\cdot)$ is the activation function.

Due to the powerful capability in handling the non-Euclidean data structure, many studies apply this technique to solve specific tasks. In [39], the famous GCN is proposed to address the semi-supervised learning problem. However, its learning manner can be

viewed as a transductive way and cannot quickly fit the new graph structure. Thus, GraphSAGE [40] designs an inductive way to learn the node representation. DeepGCN [41] takes advantage of CNN to train a very deep GCN successfully, and other studies [42, 43] learn to process a 3D point cloud via GCN because of its capability.

In this thesis, we construct a sparse cross-modality graph for conventional GCN to effectively integrate features from multi-modalities. Since the constructed graph has a non-Euclidean structure with sparse property, adopting GCN is an effective and natural way in this thesis; further details would be revealed in Section 3.3.

2.2 Metric Learning

Metric learning is a fundamental approach for model convergence which aims to learn an embedding to encode data points of the same class to stay together while those of different classes to be far apart. This is typically realized by designing a loss function to promote intra-class compactness and inter-class separability effectively. According to the given labels, metric learning can be classified into *classwise* and *pairwise*. The former prefers to employ a classification loss to optimize the similarity between samples in the feature space and weight vectors. The latter often assigns training samples in the feature space into pair or triplet relations and carries out a metric function to optimize the similarity between samples.

2.2.1 *Classwise* Scenario

Classwise scenario denotes that the ground-truth label of each sample is accessible; thus, we can approximate a feature vector of each class to globally guide samples. Specifically, it will first calculate a similarity score to describe the relationships between samples of each class and their feature vectors (or centers), then employs the classification loss function to promote the feature discrimination.

Softmax Loss is the most popular classification technique and is also called Categorical Cross-Entropy Loss. It is a combination of Softmax activation function and Cross-Entropy Loss. Concretely, it first imposes Softmax activation function to generate a probability distribution of the learned classes based on the given similarity score, then enforces Cross-Entropy loss to maximize the likelihood of the target class. Formally, Softmax Loss can be expressed as:

$$\mathcal{L}_{softmax} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{w_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{C} e^{w_j^T x_i + b_j}} \tag{2-3}$$

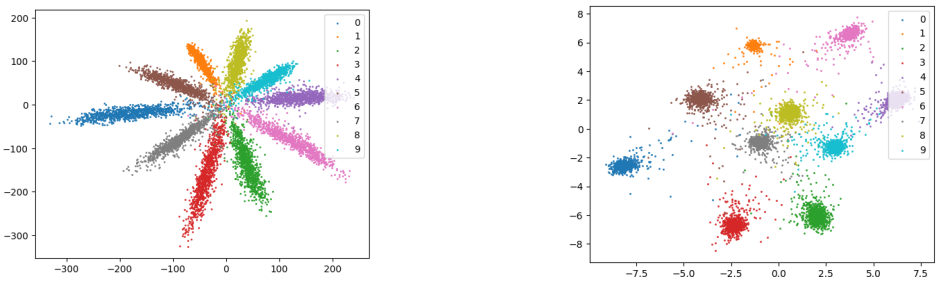
where N and C are the batch size and the total number of classes, respectively. $X \in \mathbb{R}^{N\times d}$ indicates a batch of features and $x_i \in \mathbb{R}^d$ denotes the i^{th} sample belonging to the ${y_i}^{th}$ class. $W \in \mathbb{R}^{d\times C}$ denotes the classification weight matrix, which is the learned center of each class, and $b_j \in \mathbb{R}^C$ is the bias term.

However, Softmax Loss easily leads to sparse feature distribution due to adopting the inner product as the similarity measurement. The nature of the inner product mainly focuses on optimizing the direction of each feature while the magnitude is ignored. As we can see in Figure 2-6 (a), although the feature distribution seems to be separable, the intra-class compactness is significantly low, not robust to the unseen classes.

To solve this problem, Wang *et al*. proposed Center Loss to further promote intra-class compactness [44]. This objective function exploits additional embedding centers to congregate intra-class features. Specifically, it aims to minimize the distance between features and its corresponding center. Formally, Center Loss can be expressed as:

$$\mathcal{L}_{center} = -\frac{1}{2}\sum_{i=1}^{N}\left\|x_i - c_{y_i}\right\|_2^2 \tag{2-4}$$

where $c_{y_i} \in \mathbb{R}^{d\times C}$ indicates an additional center embedding of each class.



(a) (b)

Figure 2-6: Visualized results of features training on MNIST [45]. (a) Softmax Loss may result in the embedding with low intra-class compactness. (b) With the assistance of Center Loss, the trained model can encode samples well, thereby achieving higher intra-class compactness.

Although Center Loss can improve the intra-class compactness, the memory consumptions and computational costs have to be concerned. Because the critical term is calculated from $\mathbb{R}^{d\times C}$ center embeddings, it requires more efforts to compare the difference between samples and these centers for optimization. Besides, we need to adjust the influence of Center Loss carefully. As the Euclidean distance is considered in Center Loss, the range of loss is unbounded, easily resulting in explosion gradient problem because of the huge loss value.

Recent studies, which consider projecting features and classification weights into a bounded compactness sphere space, design various techniques by adopting different kinds of penalties to control the distribution of the embedding features, thereby resulting in a robust model. An angular softmax (A-softmax) [46] is proposed to map the features and the corresponding weights into the angular space. CosFace [47] and ArcFace [48] impose different margin penalty on the target weight for controlling intra-class compactness. As a matter of fact, these angular losses can be unified as a kind of sphere mapping, and it can be expressed as a general form by:

$$f_m(\theta) = \cos(m_1\theta + m_2) - m_3$$

$$\mathcal{L}_{sphere\_mapping} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cdot f_m\left(\theta_{y_i}\right)}}{e^{s\cdot f_m\left(\theta_{y_i}\right)} + \sum_{j=1, j\neq y_i}^{C} e^{s\cdot\cos\theta_j}} \tag{2-5}$$

where the margin penalties of SphereFace [46], ArcFace [48], and CosFace [47] are respectively denoted as m_1, m_2 and m_3. For other notations, N denotes the batch size and C is the number of possible label classes. θ_{y_i} is the angle between the feature vector and its target weight vector and θ_j is the angle between the feature and other feature vectors.

Since the above sphere mapping techniques mainly focus on designing different penalties for intra-class perspective, the viewpoint of inter-class separability is neglected. RegularFace [49] instead adopts an inter-class viewpoint for learning. It works by imposing a regularization term with the orthogonal property to regulate the similarity between inter-class weights. The regularization term of RegularFace can be expressed as follows:

$$\mathcal{L}_{reg} = \frac{1}{C}\sum_{i=1}^{C}\max_{j\neq i}\langle\frac{w_i}{\|w_i\|},\frac{w_j}{\|w_j\|}\rangle \tag{2-6}$$

where w_i and w_j denote the weight of i^{th} and j^{th} classes, respectively. C is the number of possible label classes.

However, this kind of regularization may lead to huge memory usage and ineffective learning procedure for large-scale datasets with large numbers of classes. The critical term is calculated from $C \times C$ cosine-similarity matrix; thus, it may not be suitable for large-scale classes.

From the above *classwise* techniques, the learning manner is limited to the number of classes of the given dataset; thus, it may lead to a rigid training procedure. To realize a flexible training procedure, a *pairwise* scenario is proposed by directly optimizing the similarity between features. Consequently, the limitation of the label classes will be ignored, and the training procedure becomes flexible.

2.2.2 *Pairwise* Scenario

A *pairwise* scenario indicates that only have partial label information is accessible in the mini-batch. Specifically, we only know the pair or triplet relations of each sample; therefore, we cannot employ a classification weights matrix to promote feature discrimination globally. One of the representative approaches is Triplet Loss [50, 51]. Its basic idea is to minimize the distance between an anchor point and a positive point and maximize the distance between an anchor point and a negative point, see Figure 2-7.

Concretely, Triplet Loss first randomly forms a set with many triplet pairs, then adopts a fixed margin m to pull the anchor point closer to the positive point than to the negative point. Generally, Triplet Loss can be expressed as follows:

$$\mathcal{L}_{tri} = \frac{1}{|\Gamma|} \sum_{(i,j,k) \in \Gamma} [d_{ij} - d_{ik} + m]_+ \tag{2-7}$$

where Γ is a set with many triplet pairs, i, j, and k respectively denotes the index of the anchor, positive and negative points. $f(\cdot)$ is the embedding function to encode the original data point to the high-level feature, $d_{ij} = \left\| f(x_i^a) - f(x_j^p) \right\|_2$ and $d_{ik} = \| f(x_i^a) - f(x_k^n) \|_2$ respectively indicate the Euclidean distance between the anchor and positive point and the distance between the anchor and negative point. $[\cdot]_+$ denotes the hinge function to ignore the negative value.

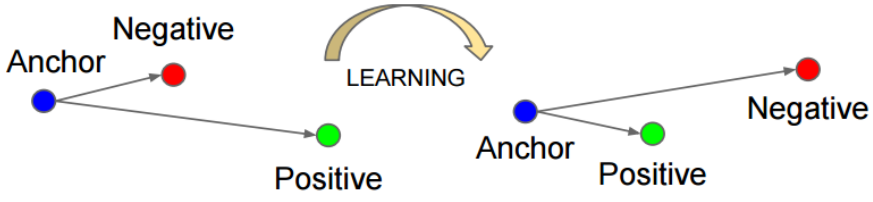


Figure 2-7: The basic idea of Triplet Loss.

However, due to training with random sampling, it inevitably causes the mini-batch involving too many redundant pairs and fails to include a good number of informative samples. It is prone to slow convergence and model degradation, which could seriously limit the targeted performance improvement. Thus, previous studies extensively explore to design mining and weighting schemes. Hermans *et al*. proposed a Batch Hard Triplet Loss [52] to learn with more informative samples. During the training stage of their scenario, each training batch is formed with *P* classes and *K* images for each class, and then the objective function explores the hardest positive and negative samples according to the given anchor for the model training. By rewriting Equation ((2-7), Batch Hard Triplet Loss can be expressed as follows:

$$\mathcal{L}_{hard\_tri} = \frac{1}{N}\sum_{i=0}^{N}\sum_{j=1,j\neq i}^{N}\left[\max_{y_i=y_j}\left(d_{ij}^{p}\right) - \min_{y_i\neq y_j}\left(d_{ij}^{n}\right) + m\right]_{+} \tag{2-8}$$

where $\max_{y_i=y_j}(\cdot)$ and $\min_{y_i\neq y_j}(\cdot)$ denote the mining scheme to seek the hardest positive and negative sample based on the given anchor.

Different from Triplet Loss family, which pulls one positive point while pushes a negative one simultaneously, *N*-Pair Loss [53] and Lifted Structure Loss [54] explore more negative samples for interaction. *N*-Pair Loss aims to *recognize one positive sample from* $N-1$ *negative samples of* $N-1$ *classes* and can be expressed as:

$$\mathcal{L}_{n-pair} = \frac{1}{N}\sum_{i}^{N} log\left\{1 + \sum_{j\neq i} exp\left(\langle f_i^a, f_j^n\rangle - \langle f_i^a, f_j^p\rangle\right)\right\} \tag{2-9}$$

where $f_i = f(x_i)$ and $\{(x_i^a, x_i^*)\}_{i=1}^{N}$ are the *N*-Pairs samples from *N* different classes. Here, x_i^a and x_i^p indicate the anchor and the positive sample respectively. x_j^n denotes the negative sample.

Lifted Structure Loss tends to *identify one positive sample from all corresponding negative samples*. Concretely, this objective function works by pulling a positive pair as close as possible and pushing all negative samples to a position farther than the margin m. Formally, Lifted Structure Loss can be expressed as:

$$\mathcal{L}_{lifted} = \frac{1}{2|\mathrm{P}|} \sum_{(i,j) \in \mathrm{P}} \left[d_{ij} + log \left(\sum_{(i,j) \in \mathrm{N}} exp(\alpha - d_{ik}) + \sum_{(l,k) \in \mathrm{N}} exp(\alpha - d_{ik}) \right) \right]_+ \quad (2\text{-}10)$$

where P and N respectively represent the set of positive pairs and negative pairs.

Instead of using a portion of informative samples to capture the structure of the embedding space, Ranked List Loss [55] exploits all pairs to construct a comprehensive structure for metric learning. Concretely, this objective function first mines non-trivial positive and negative samples, then weights the mined samples based on their loss value to emphasize the importance of each pair. On the other hand, they observe that the distribution of intra-class data may be dropped, and thus they propose a hyper-sphere constraint to preserve the intra-class similarity structure. Formally, Ranked List Loss can be expressed as:

$$\mathcal{L}_{ranked} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq i}^{N} (1 - y_{ij}) w_{ij}^{n} \left[d_{ij} - \alpha \right]_+ + y_{ij} w_{ij}^{p} \left[d_{ij} - (\alpha - m) \right]_+ \quad (2\text{-}11)$$

where $y_{ij} = 1$ if $y_i = y_j$, $y_{ij} = 0$ otherwise. w_{ij}^{*} denotes the weighting for positive and negative pairs.

Multi-Similarity (MS) Loss [56] extensively discusses the type of similarity pairs, including self-similarity and relative similarity, and designs a principled approach in mining and weighting informative pairs. Since most existing methods only explore either self-similarity or relative similarity for optimization, the performance is limited considerably. Thus, they propose an algorithm to fully consider multiple similarities

during weighting in collecting more informative pairs for better learning. Formally, MS Loss can be expressed as:

$$\mathcal{L}_{ms} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{\alpha}log\left[1+\sum_{k\in\mathcal{P}_i}e^{-\alpha(s_{ik}-\lambda)}\right]+\frac{1}{\beta}log\left[1+\sum_{k\in\mathcal{N}_i}e^{\beta(s_{ik}-\lambda)}\right]\right\} \quad (2\text{-}12)$$

where $\mathcal{P}_i$ and $\mathcal{N}_i$ indicate the mined positive pairs and negative pairs according to given anchor x_i. α, β, and λ are hyper-parameters as in Binomial Deviance Loss [57].

Common *pairwise* metric learning aims to maximize intra-class similarity s_p and minimize inter-class similarity s_n; typically, their learning manner can be expressed as seeking to reduce $(s_n - s_p)$. Circle loss [58] observes that the learning manner of previous studies is inflexible and easily converges to ambiguous results. To solve these problems, they instead propose a self-paced weighting, which measures the disparity between the optimal solution and the similarity itself, to dynamically adjust the gradient of each sample. By doing so, it forms a better decision boundary and realizes the flexible optimization, further resulting in better performance. Moreover, they also propose a unified perspective for two elemental learning paradigms, learning with *classwise* labels and *pairwise* labels. Finally, Circle Loss can be expressed as:

$$\mathcal{L}_{circle} = log\left[1+\sum_{j=1}^{L}exp\left(\gamma\alpha_j^n s_j^n\right)\sum_{i=1}^{L}exp\left(-\gamma\alpha_i^p s_i^p\right)\right] \quad (2\text{-}13)$$

where α_j^n and α_i^p are non-negative weighting factors; s_j^n and s_i^p are the similarities of the negative pair and positive pair. γ is a radius of the hypersphere.

In summary, previous techniques often realize the objective of metric learning by various mining and weighting schemes; however, an essential property is often neglected, that is embedding density. Notably, due to the nature of data distribution, the distribution of each class may still be sparse and with varied density. To solve these problems, we

here propose Density Loss for metric learning with comprehensive criteria relevant to the embedding density. By enforcing the orthogonal relation and the embedding density, we can form a useful regularization for metric learning to realize the robust embedding. We will introduce the details of our Density Loss for pairwise metric learning in Section 3.4.1.

2.3 Transfer Learning

In this section, we will give a brief introduction to transfer learning. Conventional deep learning algorithms have been mainly designed to work in isolation and often are trained to solve specific tasks. Once the embedding space changes, the trained model has to be rebuilt from scratch. To overcome the isolated learning manner, transfer learning focuses on utilizing knowledge acquired for one task (source domain) to solve related ones (target domain) [59]. Generally, there are several types of transfer learning. However, we here only introduce the concept of fine-tuning and multi-task learning since these two strategies are more relevant to our learning framework.

A simple and intuitive way for fine-turning is that we first initialize the model by a lot of the source data then adopt a small number of target data to fine-tune the initialized model, see Figure 2-8. However, because the number of target data is too small, it is easy to cause overfitting, that the model only performs well on the target training set and cannot remain the performance on the source data.

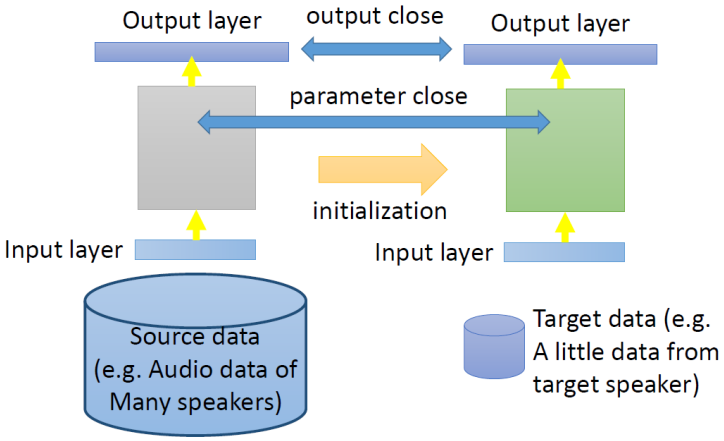


Figure 2-8: The basic idea of Fine-Tuning.

To address the serious overfitting problems, Layer Transfer is proposed, see Figure 2-9. It mainly works by copying specific layers from the source model to the target model and fine-tuning the rest of the layers. As the training parameters are significantly reduced, Layer Transfer can effectively prevent the overfitting problem.

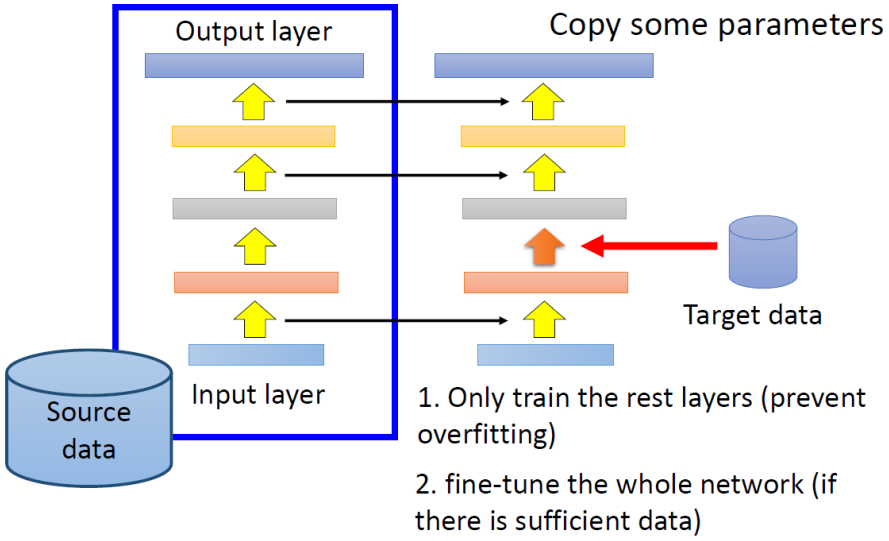


Figure 2-9: The basic idea of Layer Transfer.

Departing from the fine-tuning scheme, multi-task learning exploits a more rigorous criteria to evaluate the model performance. Concretely, it not only pursues the higher performance on the target domain but also asks to maintain the performance on the source domain. A natural way for multi-task learning is simultaneously adopted several domains

for learning, including the source and target, see Figure 2-10. However, the difference between each domain should be concerned, several works [60-62] apply the domain adaptation methods to get higher performance on the source domain and improve the results on the target domain. In this thesis, we propose Emotion Passer to realize the cross-task knowledge passing. By adopting Exponential Moving Average (EMA) mechanism, our Emotion Passer can transfer the knowledge with a fine way and realize a better performance on target domain.

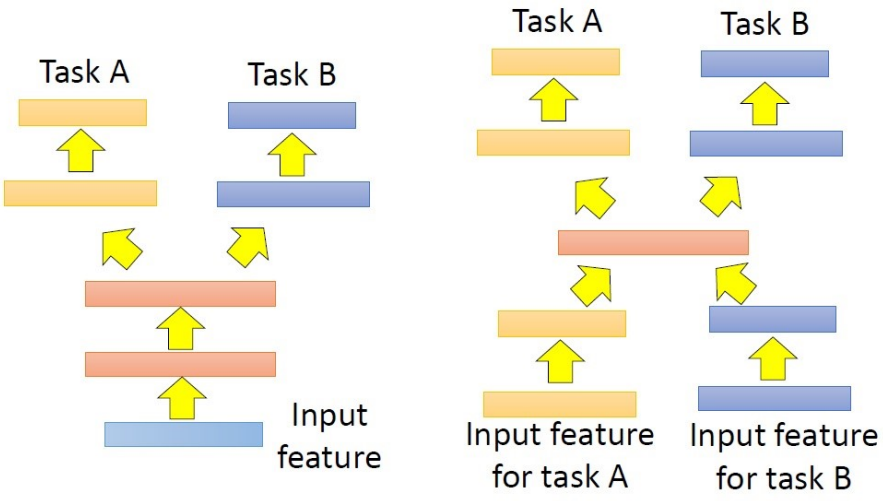


Figure 2-10: The basic idea of Multi-Task Learning.

Chapter 3 Methodology

In this thesis, we propose a multi-task learning framework to realize a mental disorder detection for schizophrenia patients via emotion recognition and depression estimation. The organization in this chapter is as follows: Section 3.1 gives a briefly overview of our mental disorder detection system, including both learning and detection parts. Then, we introduce an overview of our multi-task learning framework in Section 3.2. In Section 3.3, we first elaborate on the shortcomings of facial analysis. Then, to overcome the limitation of facial analysis, we propose Cross-Modality Graph Convolutional Network (CMGCN) to integrate the information from different modalities, including the face and context. After obtaining a robust representation, we design novel task-aware objective functions to realize a better model convergence for each task, including emotion recognition and depression estimation. The details of each objective function will be introduced in Section 3.4. Given that depression is a disorder characterized by a low emotional status, we propose Emotion Passer to effectively transfer the prior knowledge on emotion to the depression model in Section 3.5. Finally, in Section 3.6, we illustrate an algorithm to detect the mental disorders about the mood aspect from the given schizophrenia patient in order to provide an assessment for doctors to understand the severity of the patient. By such novel design, our algorithm accomplishes impressive performance in many public benchmarks.

3.1 System Overview

Our mental disorder detection system is illustrated in Figure 3-1. 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

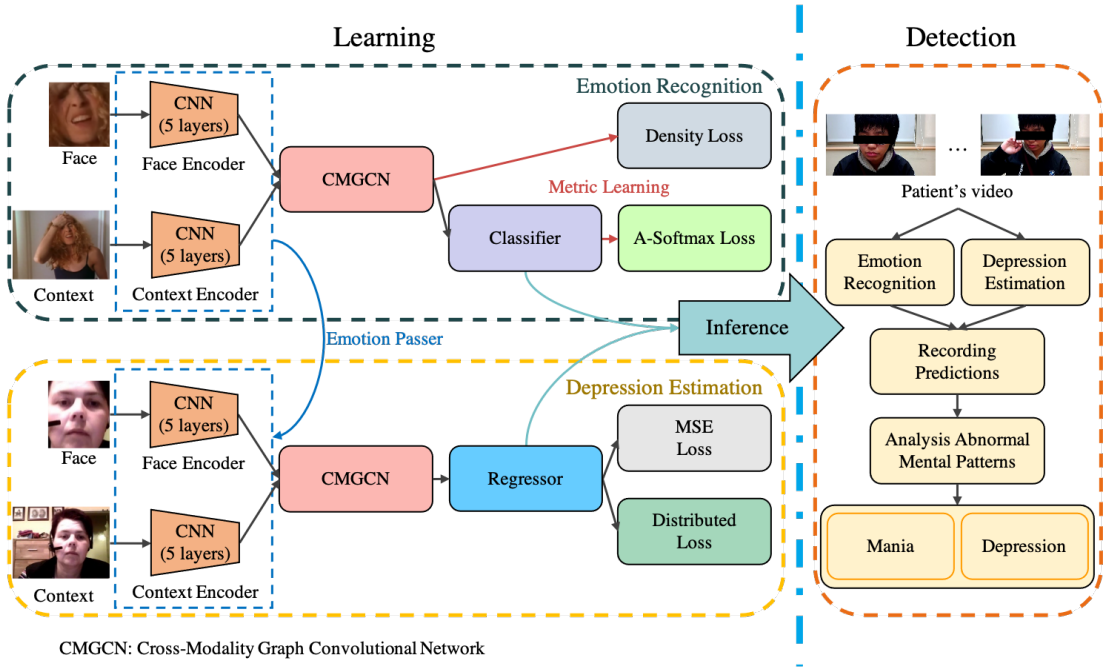


Figure 3-1: Our mental disorder detection system.

FER systems in inferring the emotional status and depressive level of humans. The details of our design will be introduced in Section 3.3, Section 3.4, and Section 3.5. 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*. The details of our detection algorithm will be elaborated in Section 3.6.

3.2 Multi-task Learning Framework

First of all, we introduce our multi-task learning framework, which is shown on the learning part in Figure 3-1. Our design mainly consists of four main components: 1) Cross-Modality Graph Convolutional Network (CMGCN), 2) Density Loss for Emotion Recognition, 3) Distributed Loss for Depression Estimation, and 4) Emotion Passer. 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. 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 Section 3.4. On the other hand, for cross-task knowledge transfer, we present Emotion Passer in Section 3.5 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 approach can successfully outperform the other state-of-the-art algorithms. Further, our system can detect the mental disorders of schizophrenia patients from National Taiwan Hospital (NTUH) via affective observation, and thus provide an assessment for doctors to estimate the severity of schizophrenia patients.

3.3 Cross-Modality Graph Convolutional Network (CMGCN)

To estimate the human mental state, the previous studies in Facial Expression Recognition (FER) suppose that facial expression comprises the most discriminative emotional responses [4-10, 24-26]; thus, algorithms based on facial analysis have been extensively discussed. However, conventional FER systems experience a failure to infer the real-time emotional status and depressive level accurately. As we can see from Figure

Figure 3-2: Comparison of face and context emotional signals. The cropped face of each frame usually expresses with different emotional signals, so FER systems often fail to recognize emotions accurately. If we consider the whole information, including the face and the context, we can get a more certain signal for recognition [18].

3-2, because of the facial muscle movements, it is ambiguous to estimate the emotion only with the cropped faces. On the other hand, in cognitive science, some studies [13, 14] have shown that people recognize the emotions of others not only from their faces but also from their surrounding contexts, such as actions, interactions with others, and the overall human's behaviors. Therefore, it is important to design a fusion mechanism to integrate features from different modalities, that are the face and context. In the following, we first introduce the previous fusion method, also called separate representation learning, which takes two isolated MLP (Multi-Layer Perceptron) layers to fuse the visual features from different modalities. After that, we introduce the proposed joint representation learning approach.

An intuitive idea [18] is to weight the features from different modalities to emphasize the importance of each separately. From Figure 3-3, the previous study adopts Global Average Pooling (GAP) to generate the representation of each modality in isolation, and

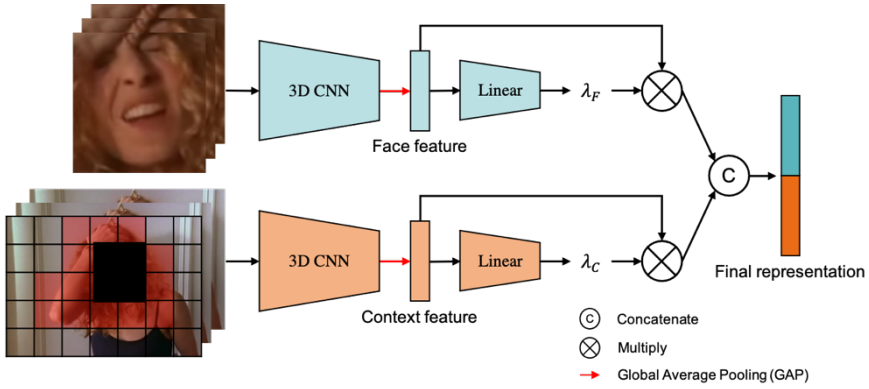


Figure 3-3: The existing separate representation [18].

then separately learns different weights via several isolated MLP layers. Further, they design an attention mechanism to let the model can pay more attention to those area with discriminative signals in the context images. Specifically, it works by mask out the face region from the given context image. However, this fusion mechanism will include many irrelevant emotional regions, such as background pixels. As we can see from the context image, it only involves a few critical regions, marked by red pixels, for recognition and estimation, whereas others are irrelevant emotional pixels, such as the background. Besides, once the face alignment fails to capture the target human face, this separate scheme may not generate a reliable weighting to reduce the effect of the wrong facial information. Another critical issue is the computational costs, as the GAP considers all pixels from the given feature map, and it is necessary to calculate the update factor of all parameters during the backpropagation stage. Typically, it will lead to an inefficient training procedure.

To tackle these problems, 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 3-4. 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

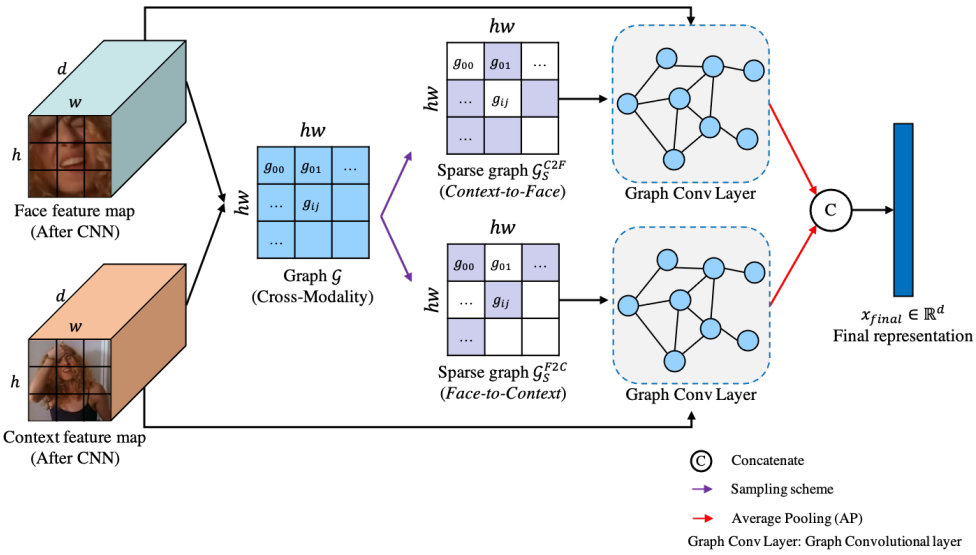


Figure 3-4: An overview of our CMGCN.

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.

3.3.1 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 [63]. 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 graph $\mathcal{G} \in \mathbb{R}^{hw \times hw}$ 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:

$$g_{ij} = \langle \frac{x_i^f}{\|x_i^f\|}, \frac{x_j^c}{\|x_j^c\|} \rangle \tag{3-1}$$

where x_i^f and x_j^c denote the pixel from face and context feature maps, respectively.

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 $\mathcal{G}$, 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, as shown in Figure 3-5. The kernel function can be expressed as follows:

$$g_{ij} = \exp(g_{ij} - 1)^p \tag{3-2}$$

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

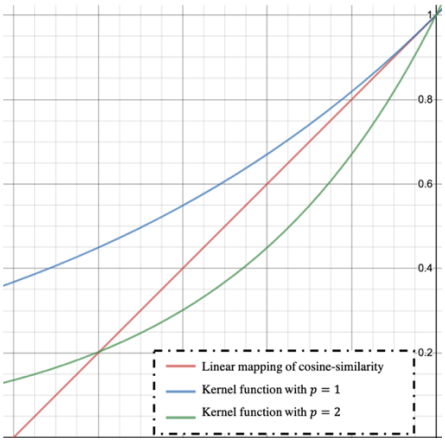


Figure 3-5: The comparison between the linear mapping of cosine-similarity, the kernel function in Equ. (3-2) with $p = 1$, and the kernel function in Equ. (3-2) with $p = 2$.

3.3.2 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 3-6. 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.

Figure 3-6: 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, as shown in Figure 3-7. Here, we briefly introduce three sampling schemes to achieve our intention: 1) epsilon ball (ϵ-ball), 2) k-nearest neighbors (k-NN), and 3) Bernoulli sampling. Note that these sampling strategies will generate a mask $M \in \mathbb{R}^{hw \times hw}$ 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. For the ϵ-ball, it will keep the elements which are higher than the defined ϵ, and it can be expressed as:

$$m_{ij} = \begin{cases} 1, & if\ g_{ij} > \epsilon \\ 0, & otherwise \end{cases} \tag{3-3}$$

where g_{ij} is an element in the given graph $\mathcal{G}$, which is calculated by Equ. (3-1), and ϵ is a defined parameter to keep/drop elements.

For the *k*-NN strategy, it will preserve the top-*k* elements of each row from the given graph, which denotes to include the top-*k* similar pixel pairs of different modalities of each pixel, *i.e.*, from the face to context or the other way around.

$$m_{ij} = \begin{cases} 1, & if\ x_j \in \mathcal{N}(x_i, k) \\ 0, & otherwise \end{cases} \tag{3-4}$$

where x_i and x_j respectively denote the *i*-th pixel in the face/context modality and the *j*-th pixel in the context/face modality. Follow [64], the k-nearest neighbor of x_i can be defined as $\mathcal{N}(x_i, k)$ and k is the hyper-parameter to control the number of selected samples.

For the Bernoulli sampling, it will respond to those elements with higher similarity in the given graph $\mathcal{G}$. Concretely, this sampling strategy will independently and randomly respond m_{ij} with 1/0 based on the g_{ij} itself, and it thus provides many more various combinations of the graph $\mathcal{G}$. Because the m_{ij} becomes a random variable in the Bernoulli processing, it can be expressed as a form of probability by

$$P(m_{ij} = 1) = g_{ij};\ P(m_{ij} = 0) = 1 - g_{ij} \tag{3-5}$$

where $g_{ij} \in [0, 1]$ can be regarded as a probability for the Bernoulli sampling.

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 in Equ. (3-5), 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. As we mentioned earlier, the sampling scheme will return a binary mask $M \in \mathbb{R}^{hw \times hw}$ with the same shape as the given graph $\mathcal{G}$, where an entry m_{ij} with respect to 1/0 means that its counterpart in $\mathcal{G}$ would be kept/dropped. The resulting sparse graph $\mathcal{G}_{sparse}$ can be expressed as follows:

$$g_{ij}^{sparse} = g_{ij} \times m_{ij} \tag{3-6}$$

where m_{ij} denotes the entry from the sampling results M and g_{ij} is the similarity from the given graph calculated by Equ. (3-2).

After building the sparse cross-modality graph $\mathcal{G}_{sparse}$, we then introduce a general GCN [39] 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

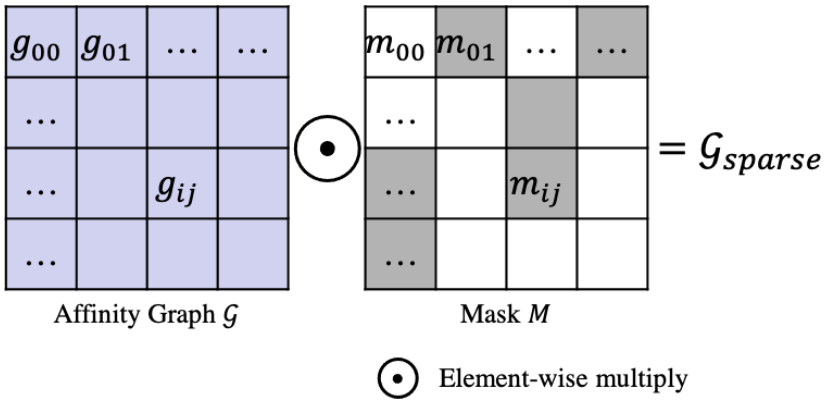


Figure 3-7: The concept of our sampling mechanism.

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:

$$\hat{A} = D^{-1}\hat{A},$$
$$X_{\mathcal{G}} = \sigma\left(\hat{A}XW\right) \tag{3-7}$$

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

3.3.3 Bidirectional Fusion

To acquire a comprehensive representation, we here consider a bidirectional way to explore critical regions among multi-modalities, as shown in Figure 3-8. 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:

$$X_{final} = \left[GAP\left(X_{\mathcal{G}}^{f} + X^{f}\right), GAP\left(X_{\mathcal{G}}^{c} + X^{c}\right)\right] \tag{3-8}$$

where X^{f} and X^{c} are face and context features, $X_{\mathcal{G}}^{f}$ and $X_{\mathcal{G}}^{c}$ indicate the graph features based on Equ. (3-7), $[\cdot,\cdot]$ denotes the concatenation operation.

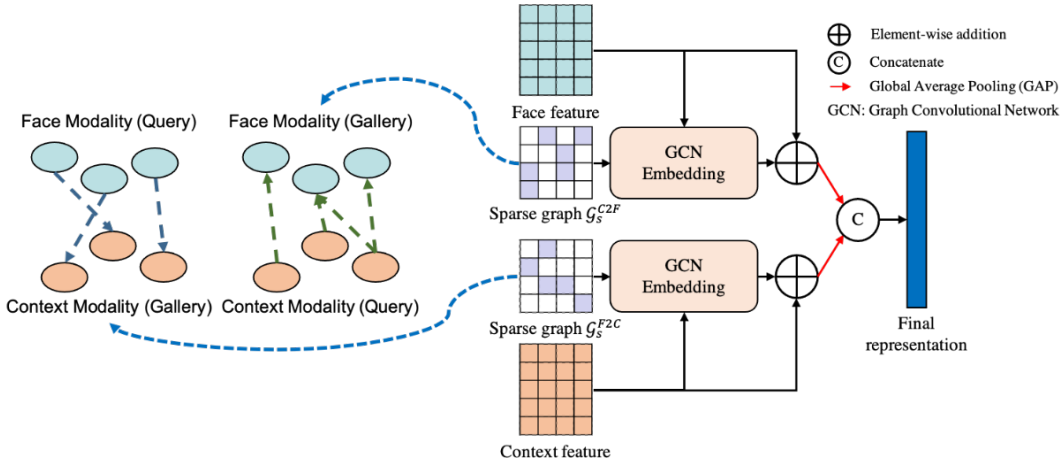


Figure 3-8: Our bidirectional fusion scheme.

3.4 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 subsections, we first elaborate on the proposed *density loss* for emotion recognition and then describe the *distributed loss* for depression estimation.

3.4.1 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). In Figure 3-9, we give a brief introduce about the concept of metric learning. The details of the previous metric learning methods are shown in Section 2.2.

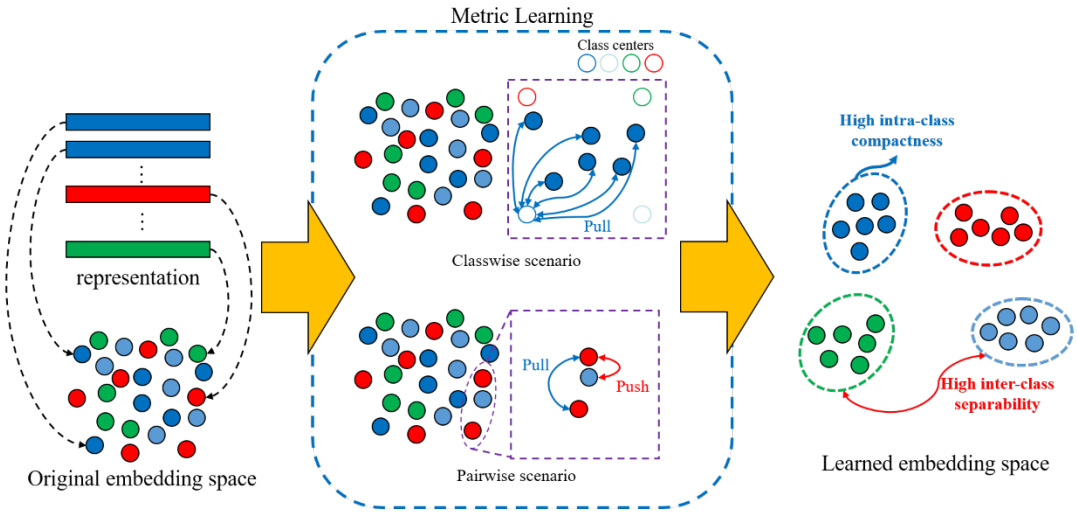


Figure 3-9: The concept of metric learning.

However, 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 and are introduced in Section 2.2. 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:

$$\mu_i^p = \sum_{j=1}^{N} s_{ij}^p \, , where \; y_i = y_j;$$
$$\mu_i^n = \sum_{j=1}^{N} s_{ij}^n \, , where \; y_i \neq y_j \tag{3-9}$$

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.

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 (see Figure 3-10), no matter how close an underlying sample is. The emphasizing terms of the proposed density loss are defined as follows:

$$w_{ij} = \begin{cases} exp\left(\left[\mu_i^p - s_{ij}^p\right]_+\right), & if \; y_i = y_j \\ exp\left(\left[s_{ij}^n - \mu_i^n\right]_+\right), & otherwise \end{cases} \tag{3-10}$$

where μ_i^p and μ_i^n indicate the density of intra-class pairs and inter-class pairs according to the given anchor i, respectively. $[\cdot]_+$ denotes the hinge function in order to drop satisfied samples (greater/less than intra-class/inter-class density). We here consider using the exponential function to keep the weight of the satisfied samples as $e^0 = 1$, and enlarge the penalty of unsatisfied samples.

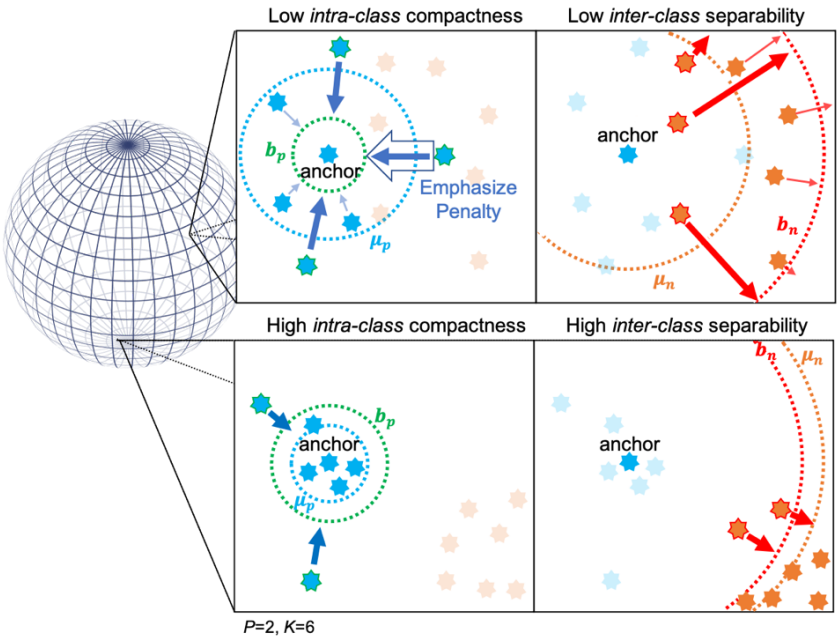


Figure 3-10: The density of the proposed Density Loss. Even when the intra-class compactness is already high, the penalty will still be properly emphasized based on the intra-class density to promote feature discrimination.

Unlike learning to reduce the disparity between positive and negative samples ($s_n - s_p$) [58], 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 which is shown in Figure 3-11 can be regarded as a tolerated area of each class, and b_n denotes the distributed region of other classes to realize better regularization.

$$l_{ij} = \begin{cases} \left[b_p - s_{ij}^p\right]_+, & if\ y_i = y_j \\ \left[s_{ij}^n - b_n\right]_+, & otherwise \end{cases} \tag{3-11}$$

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:

$$\mathcal{L}_{density} = \frac{1}{N}\left(\sum_{i=1}^{N}\left(w_{ij}l_{ij}\right)^p\right)^{\frac{1}{p}} \tag{3-12}$$

where l_{ij} denotes the loss value defined in Equ. (3-11) and $p > 1$ which specifies the underlying norm function.

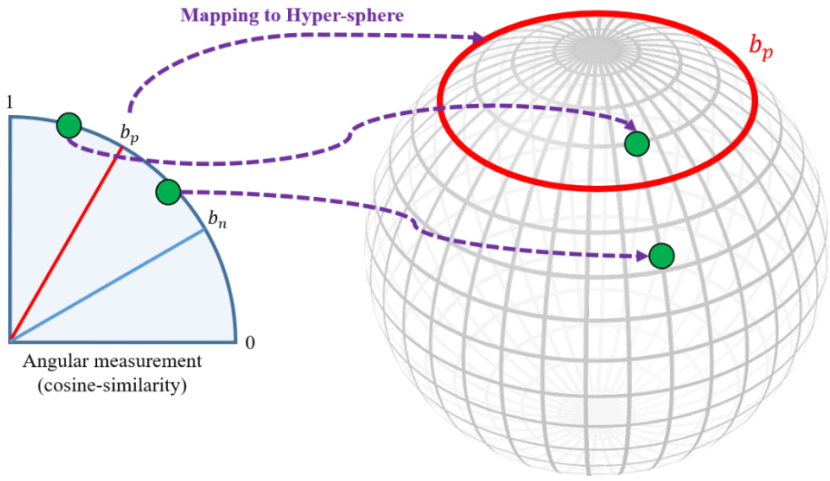


Figure 3-11: The correlation between cosine-similarity and hyper-sphere. After obtaining the pairwise similarity, each pair of samples, marked by a green point, will locate on the hyper-sphere. The intra-class boundary b_p, marked by a red line, will become a tolerated area of each class in the hyper-sphere.

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. (3-11). For the classwise metric learning, we take an Angular softmax (A-softmax), which is expressed in Equ. (2-5), 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. The details of the classwise metric learning are described in Section 2.2.1.

3.4.2 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 [20, 21, 23-26]. 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:

$$\mathcal{L}_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{3-13}$$

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.

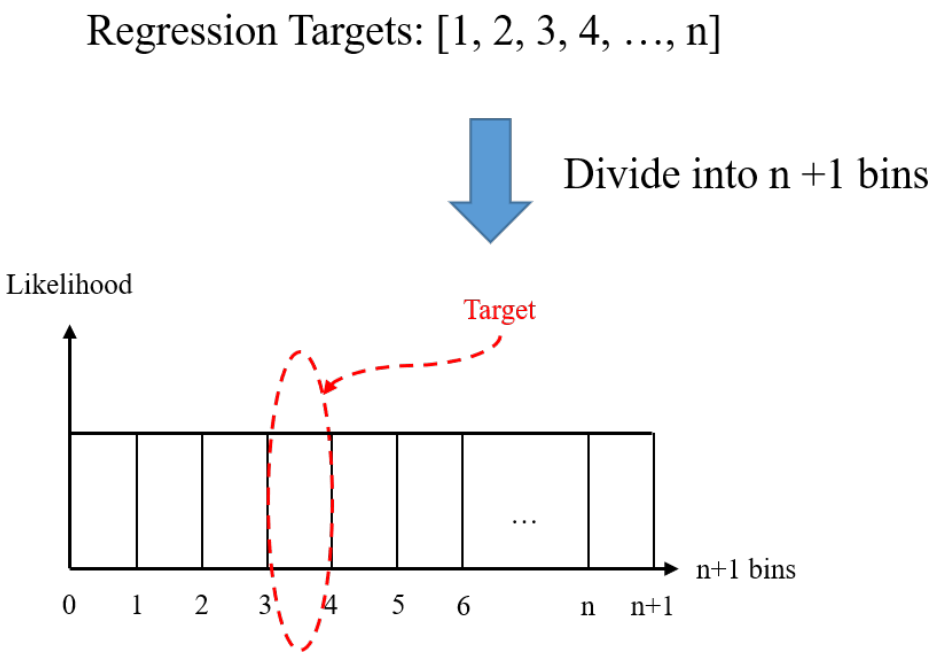


Figure 3-12: The illustration of the concept of our Distributed Loss. We introduce the classification viewpoint into the regression task.

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 [20, 21], 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 3-12.

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

$$\mathcal{L}_{distributed} = \frac{1}{n}\sum_{i=1}^{n} -log(s_{y_i}) - log(s_{y_i+1}) \tag{3-14}$$

where s_{y_i} and s_{y_i+1} indicate the likelihood scores of the target bins.

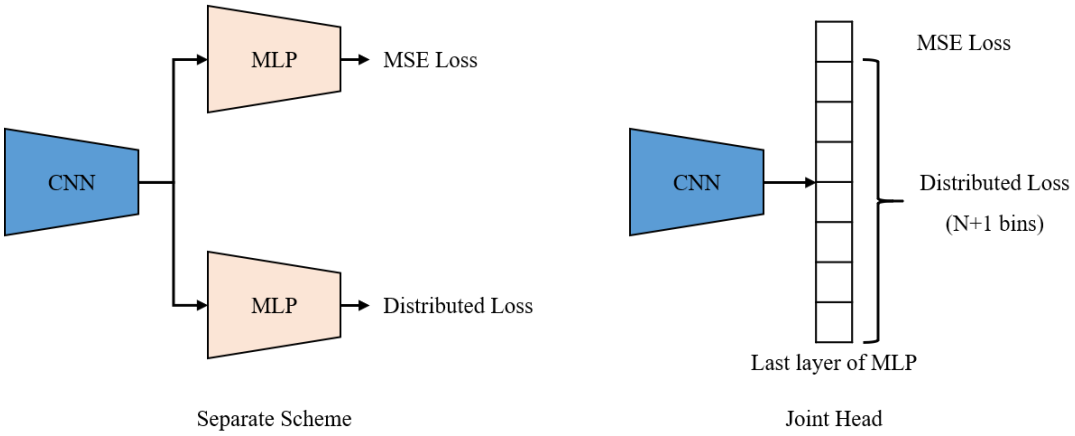


Figure 3-13: The difference between separate scheme and our joint head.

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) [65] architecture, see Figure 3-13. 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.

3.5 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:

$$\theta_t^d \leftarrow \alpha\theta_{t-1}^d + (1-\alpha)\theta_t^e \tag{3-15}$$

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.

3.6 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, 3], 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.

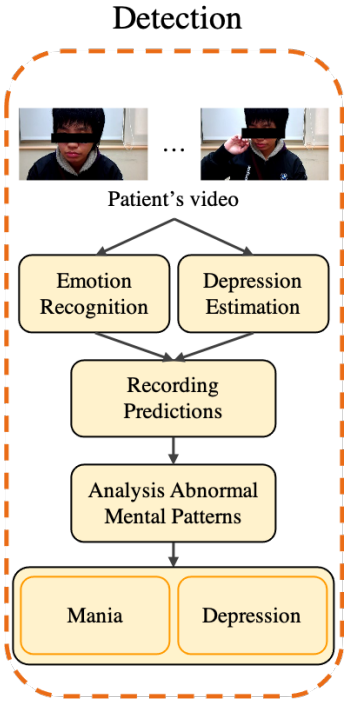


Figure 3-14: The detection flow of our detection algorithm.

In Figure 3-14, 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 [13, 14] 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, 3] 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 [13, 14] to design how we aggregate the current mental state and the past states. In the previous studies [13, 14], emotion essentially is an action readiness characterized by happening quickly, short duration, and non-periodic happening. According to the statistics [66], the duration of an emotion lasts between 0.5 to 4 seconds. Thus, we here consider to use the emotional status and depressive 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 3-15. 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.

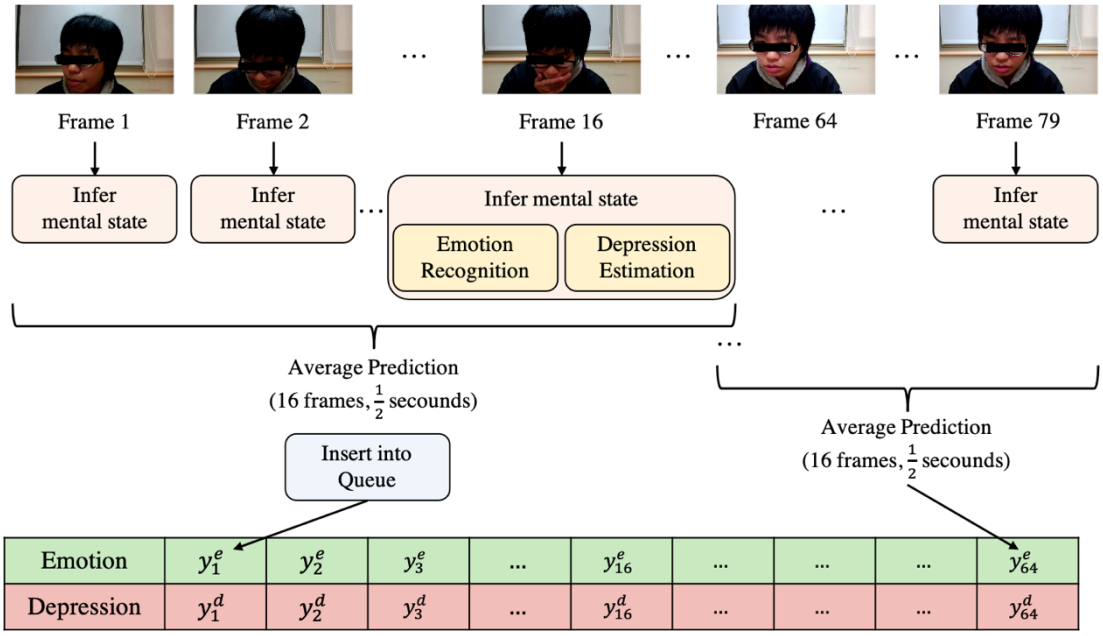


Figure 3-15: Illustration of how we record the mental state of patients.

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:

$$entropy = -\sum_{i=1}^{C} P(\hat{y}_i) \log\big(P(\hat{y}_i)\big) \tag{3-16}$$

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

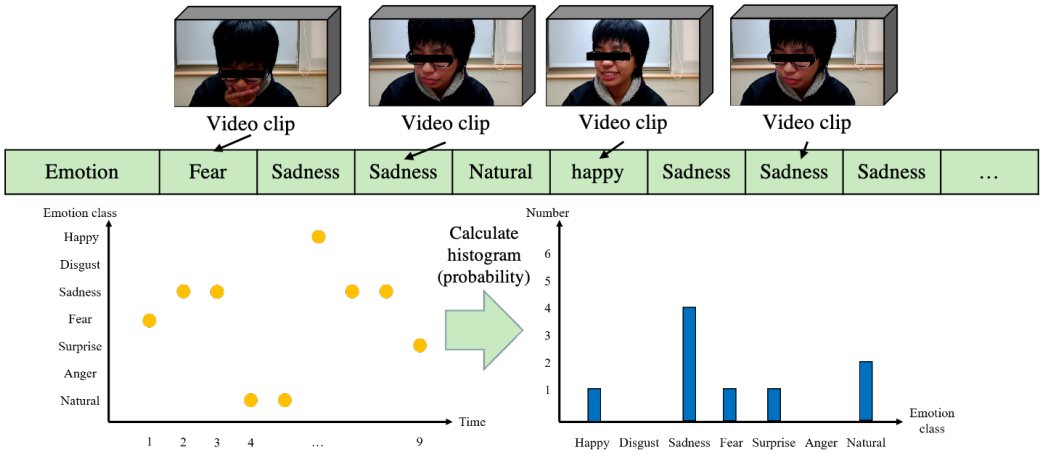


Figure 3-16: The unstable emotional status.

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 3-16. Therefore, the entropy of the given emotional statuss is high. Thus, we will give a Mania alert for notification. In Figure 3-16, 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 3-17.

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

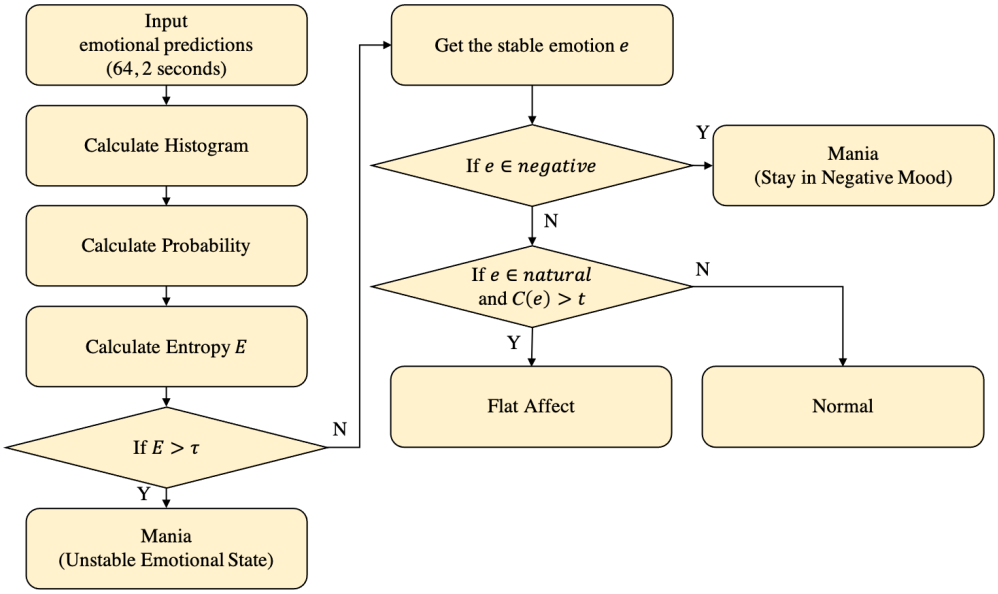


Figure 3-17: The detection flow based on the emotional status.

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 [22], 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:

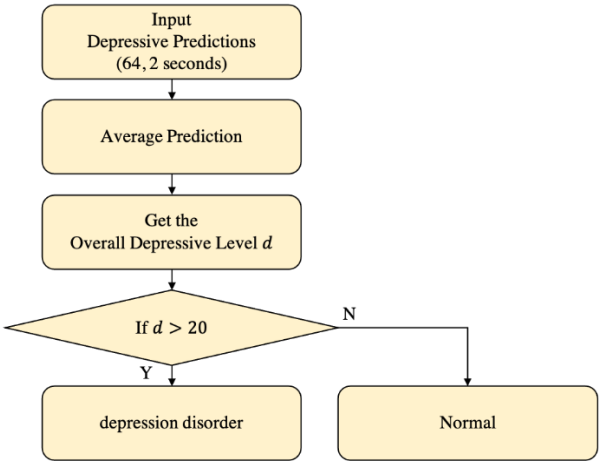


Figure 3-18: The detection flow based on the depressive level.

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.

Chapter 4 Experiments

In this chapter, we will first introduce the setting of our experimental in Section 4.1, and then provide our implementation details in Section 4.2. In Section 4.2, we explain the datasets which will need in our experiments. Section 4.4 provide various ablation studies to identify our design architecture. Finally, Section 4.5 demonstrate our experimental results and show the superiority of our performance.

4.1 Configuration

In Table 4-1, we list the specification of our experiment setup. In this thesis, we exploit PyTorch as the Application Programming Interface (API) to build up our deep learning model. The proposed model is trained on a personal computer equipped with NVIDIA GeForce GTX 2080 GPU with 8 G memory.

Table 4-1: Specification of Environment

Central Processing Unit (CPU)	Intel – 9600 K
Graphic Processing Unit (GPU)	NVIDIA GeForce GTX 12080
Random Access Memory (RAM)	32.0 GB
Operating System (OS)	Windows 10
Deep Learning API	Pytorch 1.7.1

4.2 Datasets

We evaluate our learning framework on two public datasets, including CAER [18] and AVEC 14 [21], both of which are the most well known and challenging. These two datasets will be introduced in Section 4.2.1 and Section 4.2.2, and both of them are highly related to the mental state of humans. The evaluation metrics such as accuracy, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) will be mentioned in the end.

4.2.1 Context-Aware Emotion Recognition (CAER) Dataset

CAER [18] 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 Figure 4-1, we demonstrate some images from CAER dataset.

Figure 4-1: The example frames from CAER [18].

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 4-2.

Table 4-2: Amount of video clips and frames in each category on CAER.

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

4.2.2 Audio-Visual Emotion recognition Challenge 2014 (AVEC 14) Dataset

AVEC 14 depression dataset is proposed for the Audio/Visual Emotion Challenge 2014 [21], 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 [22]), as shown in Table 1-1. Some sample images of this benchmark are shown in Figure 4-2.

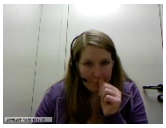
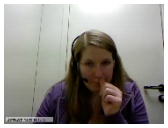
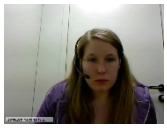
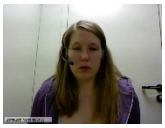
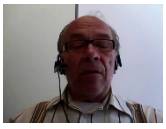
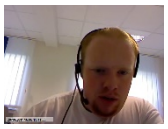
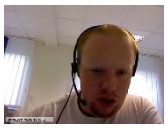
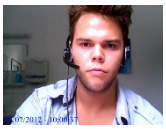
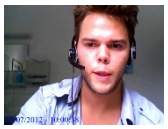
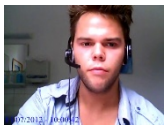
Frames				BDI II Score	Depression Severity
				0	None
				15	Mild
				24	Moderate
				44	Severe

Figure 4-2: Example video frames with depression value score in AVEC 14.

4.2.3 Evaluation Metrics

In this thesis, we evaluate our method by using the accuracy for emotion recognition, and Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) [67] 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:

$$accuracy = \frac{\#\ of\ correct\ predictions}{\#\ of\ predictions} \tag{4-1}$$

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

$$\begin{aligned} MAE &= \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \\ RMSE &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \end{aligned} \tag{4-2}$$

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.

4.3 Training 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 [18] 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 [52] (*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.

4.4 Ablation Studies

To verify the effectiveness of the proposed method, 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.

4.4.1 The influence of CMGCN

First of all, we investigate our CMGCN with different modalities in detail, and the results are shown in Table 4-3. 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.

Table 4-3: Comparison of our CMGCN in different modalities.

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

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.

Table 4-4: Comparison of our CMGCN with different sampling schemes.

Sampling Scheme	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

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. The details of each sampling are introduced in Section 3.3.2. As we can see from Table 4-4, $\mathcal{G}_{full}$ 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 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.

4.4.2 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 4-5, for the Baseline group, we consider the A-Softmax Loss with $s = 20$ to train the entire model, where more details are shown in Section 2.2.1. $\mathcal{L}_{hard\_tri}$ and $\mathcal{L}_{density}$ respectively indicate the Hard Triplet Loss and our Density Loss, where the formulation of Hard Triplet Loss $\mathcal{L}_{hard\_tri}$ are defined by Equ. (2-8) in Section 2.2.2. 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 $\mathcal{L}_{hard\_tri}$, the intra-class compactness can be enhanced the with 3.51% improvement in accuracy on CAER. However, $\mathcal{L}_{hard\_tri}$ may easily cause ambiguous optimization results, are shown in Section 2.2.2. 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 ($s_p - s_n > m$), where more details are shown in Section 2.2.2. 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

Table 4-5: Comparison of our Density Loss.

Method	CAER
	Accuracy (%)
Baseline	81.75
Baseline + $\mathcal{L}_{hard\_tri}$	85.26
Baseline + $\mathcal{L}_{density}$	87.23

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.

4.4.3 The influence of Distributed Loss

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 4-6, $\mathcal{L}_{MSE}$ and $\mathcal{L}_{Distri}$ 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 $\mathcal{L}_{MSE}$ and $\mathcal{L}_{Distri}$. The Joint group indicates that we adopt a single MLP (Multi-Layer Perceptron) layer to jointly learn with $\mathcal{L}_{MSE}$ and $\mathcal{L}_{Distri}$.

Table 4-6: Comparison of our Distributed Loss.

Method	Head Type	AVEC 14	
		MAE	RMSE
$\mathcal{L}_{MSE}$	--	8.0123	9.7324
$\mathcal{L}_{Distri}$	--	7.9850	9.6507
$\mathcal{L}_{MSE}$ + $\mathcal{L}_{Distri}$	Separate	7.4287	9.0254
$\mathcal{L}_{MSE}$ + $\mathcal{L}_{Distri}$	Joint	6.8206	8.5078

From Table 4-6, our $\mathcal{L}_{Distri}$ can perform better than $\mathcal{L}_{MSE}$. 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 $\mathcal{L}_{MSE}$. 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.*, $\mathcal{L}_{MSE}$ stream predicts the large value while $\mathcal{L}_{Distri}$ predicts the bin located at the low level. To mitigate this ambiguity, we follow YOLO [65] to design a joint head for prediction; thus, the performance can be greatly improved due to the joint learning manner.

4.4.4 The influence of Emotion Passer and Joint Head

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 4-7, 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.

Table 4-7: Comparison of our Emotion Passer.

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

4.5 In Comparison with State-Of-The-Arts (SOTA) Works

In this section, we compare the proposed method with several SOTA approaches.

4.5.1 The result on CAER

For a fair comparison, we follow the same scenarios in [18] to demonstrate the effectiveness of the proposed approach. From Table 4-8, we can see that the deeper backbones, such as ResNet [36], can achieve the better performance than AlexNet [27] because it can extract more discriminative features. CAER-Net-S [18] 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 [68] 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 [18] can better fuse the temporal information and achieve advanced performance with 77.04% on accuracy.

Table 4-8: Emotion Recognition: Comparison of the SOTA on CAER.

Method	Data type	Modality	CAER
			Accuracy (%)
ImageNet-AlexNet [27]	Image	Face + Context	47.36
ImageNet-ResNet [36]	Image	Face + Context	57.33
Fine-tuned AlexNet [27]	Image	Face + Context	61.73
Fine-tuned ResNet [36]	Image	Face + Context	68.46
CAER-Net-S [18]	Image	Face	70.09
		Context	65.65
		Face + Context	73.51
Sports-1M-C3D [68]	Video	Face + Context	66.38
Fine-tuned C3D [68]	Video	Face + Context	71.02
CAER-Net [18]	Video	Face	74.13
		Context	75.57
		Face + Context	77.04
Ours	Video	Face + Context	**87.23**

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 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

4.5.2 The result on AVEC 14

We follow the same scenario in [21, 25] to evaluate our depression model on AVEC 14. The quantitative results are shown in Table 4-9, our depression model outperforms all other SOTA methods with 6.82/8.5 on MAE/RMSE. As we can see, other methods [21, 25, 69-73] 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 [73], 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 [25, 72, 73].

Table 4-9: Depression Estimation: Comparison of the SOTA on AVEC 14.

Method	Data type	Modality	AVEC 14	
			MAE	RMSE
Baseline [21]	Video	Face	8.86	10.86
UUIMSidorov [69]	Video	Face	11.20	13.87
InaoeBuap [70]	Video	Face	9.35	11.91
Brunel [71]	Video	Face	8.44	10.50
BU-CMPE [72]	Video	Face	7.96	9.97
DCNN [25]	Video	Face	7.47	9.55
RNN-C3D [73]	Video	Face	7.22	9.20
Ours	Video	Face + Context	**6.82**	**8.50**

4.5.3 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 4-3.

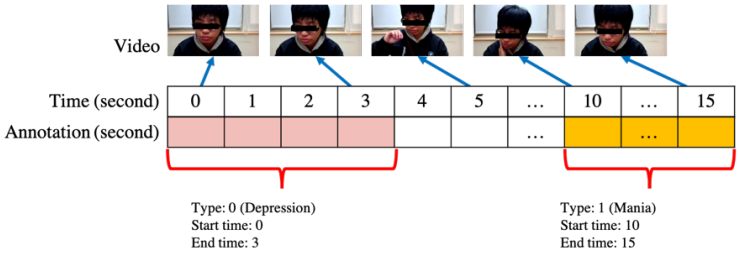


Figure 4-3: An example about the temporal annotation.

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:

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

$$AP_j = \frac{1}{n_j} \sum_{i=1}^{n_j} F(j, y_i, p_i) \tag{4-3}$$

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

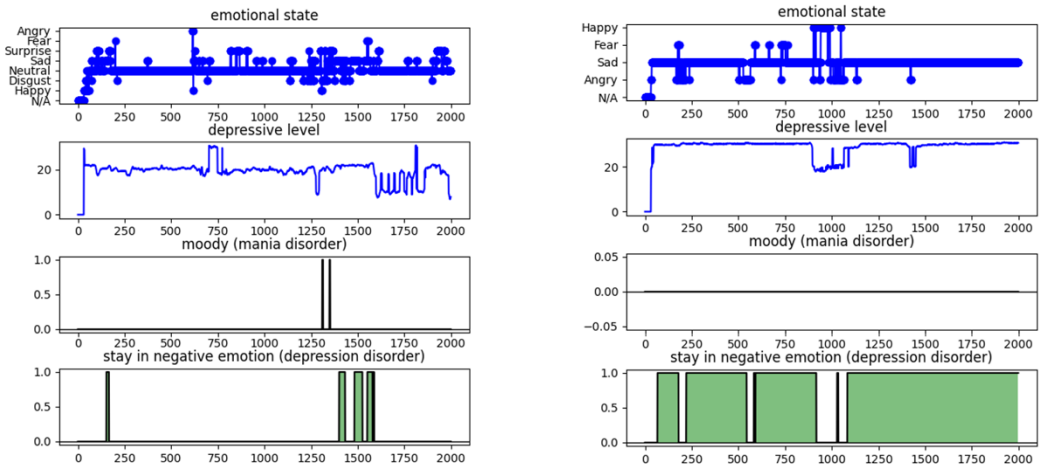
where C and j denote the total number of classes in the given video and the label of j-th class, respectively; n_j is the number of frames corresponding to the given class j, and y_i, and p_i respectively denote the ground-truth and the predicted disorder class of i-th frame.

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 Chapter 3. As we can see from Table 4-10, the results show that our algorithm can successfully detect Mania disorder and Depression

Table 4-10: The performance of our Mental Disorder Detection System.

Case	Class	AP	mAP
Case 1	Mania	72.56	--
	Depression	75.32	--
	Overall	--	73.94
Case 2	Mania	69.21	--
	Depression	76.43	--
	Overall	--	72.82
Overall	--	--	73.38

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.



(a) (b)

Figure 4-4: The prediction results in a short duration. (a) for case 1 and (b) for case 2.

Chapter 5 Conclusion and Future Works

In this thesis, 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.

REFERENCE

[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] T. Gonzalez and C. Chiodo, "ICD 10," *Foot & Ankle International,* vol. 36, pp. 1110-1116, 2015.

[4] C. F. Benitez-Quiroz, R. Srinivasan, and A. Martínez, "EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 5562-5570, 2016.

[5] 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.

[6] S. Li, W. Deng, and J. Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2584-2593, 2017.

[7] F. Xiaoyi, "Facial expression recognition based on local binary patterns and coarse-to-fine classification," in *The Fourth International Conference on Computer and Information Technology (CIT),* pp. 178-183, 2004.

[8] 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.

[9] 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.

[10] 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)*, pp. 1037-1041, 2020.

[11] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision (IJCV)*, vol. 60, pp. 91-110, 2004.

[12] A. Kläser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in *British Machine Vision Conference (BMVC)*, 2008.

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

[14] E. K. Gray and D. Watson, "Assessing positive and negative affect via self-report," *Handbook of Emotion Elicitation and Assessment*, 2007.

[15] 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, pp. 1238-1246, 2008.

[16] 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.

[17] 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),* pp. 1960-1968, 2017.

[18] J. Lee, S. Kim, S. Kim, J.-I. Park, and K. Sohn, "Context-Aware Emotion Recognition Networks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10142-10151, 2019.

[19] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "EmotiCon: Context-Aware Multimodal Emotion Recognition Using Freg… s Principle," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 14222-14231, 2020.

[20] M. Valstar, B. Schuller, K. Smith, F. Evben, B. Jiang, S. Bilakhia, S. Schieder, 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*, Barcelona, Spain, 2013.

[21] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Evben, B. Jiang, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, Orlando, Florida, USA, 2014.

[22] 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. 588-597, 1996.

[23] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals: a multimodal approach," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge,* 2013.

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

[25] 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.

[26] M. Ding, Y. Huo, J. Hu, and Z. Lu, "DeepInsight: Multi-Task Multi-Scale Deep Learning for Mental Disorder Diagnosis," in *British Machine Vision Conference (BMVC)*, 2018.

[27] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS),* 2012.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F.-F. Li "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision,* vol. 115, no. 3, pp. 211-252, 2015.

[29] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *European Conference on Computer Vision (ECCV)*, pp. 818-833, 2014.

[30] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich Feature Hierarchies for

Accurate Object Detection and Semantic Segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 580-587, 2014.

[31] A. Bochkovskiy, C. Wang, and H. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *ArXiv*, vol. abs/2004.10934, 2020.

[32] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[33] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1933-1941, 2016.

[34] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, C. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-9, 2015.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016.

[37] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks,* vol. 2, no. 5, pp. 359-366, 1989.

[38] S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 6, pp. 107-116, 1998.

[39] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[40] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," in *Advances in Neural Information Processing Systems (NeurIPS)*,

2017.

[41] G. Li, M. Muller, A. Thabet, and B. Ghanem, "DeepGCNs: Can GCNs Go As Deep As CNNs?," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV),* pp. 9266-9275, 2019.

[42] Q. Xu, X. Sun, C. Y. Wu, P. Wang, and U. Neumann, "Grid-GCN for Fast and Scalable Point Cloud Learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5660-5669, 2020.

[43] J. Chen, B. Lei, Q. Song, H. Ying, D. Z. Chen, and J. Wu, "A Hierarchical Graph Network for 3D Object Detection on Point Clouds," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 389-398, 2020.

[44] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition," in *European Conference on Computer Vision (ECCV)*, pp. 499-515, 2016.

[45] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

[46] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep Hypersphere Embedding for Face Recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6738-6746, 2017.

[47] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5265-5274, 2018.

[48] J. Deng and S. Zafeiriou, "ArcFace for Disguised Face Recognition," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 485-493, 2019.

[49] K. Zhao, J. Xu, and M. Cheng, "RegularFace: Deep Face Recognition via Exclusive Regularization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1136-1144, 2019.

[50] C. Wu, R. Manmatha, A. J. Smola, and P. Krähenbühl, "Sampling Matters in Deep

Embedding Learning," in *2017 IEEE International Conference on Computer Vision (ICCV)* , pp. 2859-2867, 2017.

[51] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815-823, 2015.

[52] A. Hermans, L. Beyer, and B. J. A. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," *ArXiv*, vol. abs/1703.07737, 2017.

[53] K. Sohn, "Improved Deep Metric Learning with Multi-class N-pair Loss Objective," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1857-1865, 2016.

[54] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep Metric Learning via Lifted Structured Feature Embedding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4004-4012, 2016.

[55] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. Robertson, "Ranked List Loss for Deep Metric Learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 5202-5211, 2019.

[56] X. Wang, X. Han, W. Huang, D. Dong, and M. Scott, "Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 5017-5025, 2019.

[57] D. Yi, Z. Lei, and S. Li, "Deep Metric Learning for Practical Person Re-Identification," *ArXiv,* vol. abs/1407.4979, 2014.

[58] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zhang, Z. Wang. and Y. Wei, "Circle Loss: A Unified Perspective of Pair Similarity Optimization,"in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 6397-6406, 2020.

[59] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering,* vol. 22, no. 10, pp. 1345-1359, 2010.

[60] Y.-J. Li, F.-E. Yang, Y.-C. Liu, Y.-Y. Yeh, X. Du, and Y.-C. Wang, "Adaptation and re-identification network: An unsupervised deep transfer learning approach

to person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 172-178, 2018.

[61] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," *ArXiv, vol. abs/1611.05244,* 2016.

[62] P. Peng, T. Xiang, Y. Wang, and M. Pontil, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1306-1315, 2016.

[63] L. Yang, X. Zhan, D. Chen, J. Yan, C. C. Loy, and D. Lin, "Learning to Cluster Faces on an Affinity Graph," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2293-2301, 2019.

[64] D. Qin, S. Gammeter, L. Bossard, T. Quack, and L. V. Gool, "Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 777-784, 2011.

[65] J. Redmon, S. Divvala, R. B. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 779-788, 2016.

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

[67] 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.

[68] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4489-4497, 2015.

[69] M. Sidorov and W. Minker, "Emotion Recognition and Depression Diagnosis by Acoustic and Visual Features: A Multimodal Approach," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014.

[70] H. Espinosa, H. Escalante, L. Pineda, M. Montes-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 *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014.

[71] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic Depression Scale Prediction using Facial Expression Dynamics and Regression," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014.

[72] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble CCA for Continuous Emotion Prediction," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014.

[73] 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.