

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

通過深度視覺感知

實現思覺失調症患者的精神障礙檢測

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

炳彰 January xx, 2021

中文摘要

近年來，思覺失調症是一種精神疾病，會逐漸改變一個人的精神狀態並導致嚴重的社會成本。患者通常無法正常表達自己的真實想法，並在行為上與他人有所不同。根據《精神疾病手冊》，我們觀察到思覺失調症的精神障礙與情緒和抑鬱高度相關。因此，我們設計了一多任務學習框架來實現思覺失調症患者的精神障礙檢測。具體而言，我們利用表情識別與憂鬱評估來推算患者的心理狀態，接著根據預測的狀態來檢測異常的模式。

由於先前的研究主要集中在面部分析以識別人的情緒和抑鬱，因此它們經常無法提供令人滿意的性能。面部表情是極為不穩定的情緒表示，常因面部肌肉的運動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，患者不能表達具有區別性的面部表情。因此，常規的面部表情識別系統不適合識別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，以有效地集成來自不同模態的視覺特徵。通過稀疏圖和圖卷積，我們可以鏈接相關的視覺提示，並刪除其他不相關的視覺提示，例如背景像素。通過這樣做，我們可以從內容和面部表情有效地整合情緒特徵，從而產生更可靠的表示方式。另一方面，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屬性有關的綜合標準。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。由於抑鬱症的本質是情感的延伸，我們設計了一種知識轉移方法，即均值傳遞者，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 在內的多個基準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，我們還在幾個細粒度的檢索基準中驗證了密度損失的有效性。與最先進的度量學習技術相比，我們的密度損失也擊敗了所有的方法，並達到最佳的性能。

關鍵字：圖卷積網路、表情識別、憂鬱評估、度量學習

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 someone else. According to the handbook of mental illness, we observe that the mental disorder of schizophrenia is highly correlated with emotion and depression. We are thus motivated to design a multi-task learning framework to realize a mental disorder system for schizophrenia patients. Specifically, our framework adopts emotion recognition and depression estimation to infer the mental state of patients, and then detect the abnormal patterns based on the predicted results.

Since 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, patients cannot express a discriminative facial expression. Thus, conventional facial expression recognition systems are not suitable to identify the emotional state and depressive level of schizophrenia patients.

To overcome these limitations, we propose Cross-Modality Graph Convolutional Networks (CMGCN) to effectively integrate visual features from different modalities, including the face and context. With the sparse graph and graph convolution, we can link the relevant visual cues and drop other irrelevant ones, such as background pixels. By doing so, we can effectively integrate the emotional features from the face and content modalities, and thus yield a more robust

representation for identification. 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 embedding density and orthogonal property. By considering these two critical attributes, we can form powerful regularization for embedding learning. In addition, for depression estimation, we take the classification viewpoint to form Distributed Loss for the regression task. By dividing the regression level into several bins, we can emphasize the influence of the loss value in a meticulous way. As the nature of depression is an extension of emotion, we design a knowledge transfer approach, namely Mean Passer, to transfer the emotion prior knowledge to the depression model. In each mini-batch, we take the exponential moving average scheme to smoothly pass the knowledge and realize a robust model for the unseen data. By the well-design multi-task learning framework, we can precisely identify the mental state of patients and achieve a mental disorder detection system based on the recorded state.

To verify the effectiveness of our framework, we perform series of experiments on several benchmarks, 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 SOTA methods. In addition, we also validate the effectiveness of our Density Loss in several fine-grained benchmarks. Compared with the most advanced metric learning techniques, our Density Loss can beat them with a clear margin by 0.8% on CUB, 1.4% on Cars-196, 0.2% on SOP, and 2.6% on In-Shop.

Keywords: Graph Convolutional Networks, Emotion Recognition, Depression Estimation, Metric Learning

CONTENTS

口試委員會審定書	#
誌謝	i
中文摘要	ii
ABSTRACT	iv
CONTENTS	vi
LIST OF FIGURES	ix
LIST OF TABLES	xi
Chapter 1 Introduction	1
1.1 Motivation	1
1.2 Literature Review	4
1.2.1 Emotion Recognition	4
1.2.2 Depression Estimation	5
1.3 Contributions	7
1.4 Thesis Organization	8
Chapter 2 Preliminaries	10
2.1 Deep Neural Networks	10
2.1.1 Convolutional Neural Networks (CNNs)	10
2.1.2 Residual Network	14
2.1.3 Graph Convolutional Networks	15
2.2 Metric Learning	18
2.2.1 <i>Classwise</i> Scenario	18
2.2.2 <i>Pairwise</i> Scenario	22

2.3	Transfer Learning	26
Chapter 3	Methodology	29
3.1	Framework Overview	29
3.2	Cross-Modality Graph Convolutional Networks.....	31
3.2.1	Cross-Modality Graph Construction	34
3.2.2	Sampling Scheme and GCN Embedding	34
3.2.3	Fusion with Bidirectional Way.....	37
3.3	Objective Functions	38
3.3.1	Density Loss for Emotion Recognition.....	38
3.3.2	Distributed Loss for Depression Estimation	41
3.4	Mean Passer	43
3.5	Mental Disorder Detection	44
Chapter 4	Experiments.....	47
4.1	Configuration.....	47
4.2	Training Details	47
4.3	Datasets.....	48
4.3.1	Context-Aware Emotion Recognition (CAER) Dataset.....	48
4.3.2	Audio-Visual Emotion recognition Challenge 2014 (AVEC 14) Dataset	50
4.3.3	Fine-Grained retrieval benchmarks	51
4.3.4	Evaluation Metrics	51
4.4	Ablation Studies.....	53
4.4.1	The influence of CMGCN.....	53
4.4.2	The influence of Density Loss and Distributed Loss	55
4.4.3	The influence of Distributed Loss.....	56

4.4.4	The influence of Mean Passer and Joint Head	57
4.5	Comparing with State-Of-The-Arts (SOTA)	58
4.5.1	The result on CAER	58
4.5.2	The result on AVEC 14.....	60
4.5.3	The result on Fine-Grained Image retrieval benchmarks.....	61
Chapter 5	Conclusion and Future Works	63
REFERENCE	65

LIST OF FIGURES

Figure 1-1: An overview of CAER-Net [18].	5
Figure 2-1. A standard 3×3 convolutional operation with stride 1.	11
Figure 2-2: The architecture of AlexNet, VGG16, and Inception Block.	13
Figure 2-3: The difference between the flat convolutional layers and the residual connections.	14
Figure 2-4: The family of ResNet [38], each convolutional block involves several residual blocks.	15
Figure 2-5: The difference between the convolution and the graph convolution.	16
Figure 2-6: Visualized results of features training on MNIST [44]. (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.	19
Figure 2-7: The basic idea of Triplet Loss.	22
Figure 2-8: The basic idea of Fine-Tuning.	26
Figure 2-9: The basic idea of Layer Transfer.	27
Figure 2-10: The basic idea of Multi-Task Learning.	28
Figure 3-1: Our multi-task learning framework.	30
Figure 3-2: Comparison of face and context emotional signals. The cropped face of each frame usually expresses with different emotional signals, so the FER system often fails 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].	31
Figure 3-3: The existing separate representation [18].	32

Figure 3-4: An overview of our CMGCN.	33
Figure 3-5: 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 while others are background pixels.	35
Figure 3-6: The concept of our sampling mechanism.	36
Figure 3-7: The bidirectional way to link relevant features.	37
Figure 3-8: The vector fields 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.....	39
Figure 3-9: The illustration of the concept of our Distributed Loss. We introduce the classification viewpoint into the regression task.	41
Figure 3-10: The difference between separate scheme and our joint head.....	42
Figure 3-11: Illustration of our detection framework. We adopt a camera with 32 fps to record the video. Then, we build a video clip with 16 frames (0.5 seconds) for identification, including emotion recognition and depression estimation. Finally, we will gather 64 predictions (2 seconds) for analysis to detect the abnormal pattern.	45
Figure 3-12: Our mental disorder detection flow.	46
Figure 4-1: The example frames from CAER [18].....	49
Figure 4-2: Example video frames with depression value score in AVEC 14.....	50
Figure 4-2: An example of Cumulative Match Characteristic (CMC) curve.	52

LIST OF TABLES

Table 1-1: Beck Depression Inventory-II (BDI-II) score and Depression Severity.	5
Table 4-1: Specification of Environment.....	47
Table 4-2: Amount of video clips in each category on CAER.	49
Table 4-3: Details of each fine-grained retrieval benchmark.	51
Table 4-4: Comparison of our CMGCN in different modalities.....	54
Table 4-5: Comparison of our CMGCN with different sampling schemes.	55
Table 4-6: Comparison of our Density Loss.....	56
Table 4-7: Comparison of our Distributed Loss.	57
Table 4-8: Comparison of our Mean Passer.....	58
Table 4-9: Comparison of the SOTA on CAER.....	59
Table 4-10: Comparison of the SOTA on AVEC 14.	60
Table 4-11: Comparison of the SOTA on CUB and Cars-196.....	61
Table 4-12: Comparison of the SOTA on SOP and In-Shop.	62

Chapter 1 Introduction

In this chapter, we first describe the research motivation in Section 1.1, then elaborate on the lecture 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 someone else. According to the handbook about mental illness [2, 3], the mental disorder of schizophrenia can be described as Bipolar disorder and Depression disorder. Generally, Bipolar disorder is an unstable emotional condition characterized by cycles of abnormal, persistent high mood and low mood. Further, patients may feel abnormally energetic, happy, or irritable. On the other hand, Depression disorder is a low emotional state, and patients often stay in pervasive sadness and depression. Since these disorders are highly correlated with emotion and depression, we can naturally adopt emotion recognition and depression estimation to infer mental states, further realizing a mental disorder detection system.

Previous studies [4-6] assume that the human face can provide the most discriminative emotional signals so that they have already done extensive discussion based on facial analysis. Most approaches identify human emotion based on facial expression analysis [4-8]. Some of the others consider the facial action encoding system to analyze movements of the face to recognize human expression [9, 10]. Due to the

significant visual appearance changes, encoding the discriminative feature from the given facial image is the crux for identification. Over the last decade, conventional methods extract the feature based on the hand-crafted feature such as SIFT [11] or HOG [12]. However, these methods require some domain knowledge and not easy to generalize to real-world scenarios. Instead of designing specific strategies to extract features, recent deep convolutional neural networks (CNNs) based approaches have made significant progress by learning with the distribution of 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, 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 conflicting emotional signals compared to the given 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 express fewer emotional signals [1]. Therefore, facial analysis alone may not be suitable for detecting the emotional state 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 the interaction of time series 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. Compared with schizophrenia patients without depression, patients with depression have a worse treatment course and prognosis.

We are thus motivated to design a multi-task learning framework to realize a mental disorder 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 features 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 convolutional. By doing so, our CMGCN can effectively connect feature with higher similarity and discard other irrelevant ones, such as background pixels. Therefore, it can greatly reduce irrelevant information and thus yield a more comprehensive representation for identification. In addition to this effort, we also design task-aware objective functions to realize better model convergence. For emotion recognition, a classification task, we propose Density Loss for metric learning to form a powerful regularization for embedding learning. In particular, we consider comprehensive criteria relevant to orthogonal property and embedding density to design the metric function; thus, it can 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 an extension of emotion, we present a knowledge transfer strategy, namely Mean Passer, to effectively pass the emotion prior knowledge to the depression model. In each iteration, our Mean 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. Compared with other state-of-the-art (SOTA) studies, our learning framework can achieve the most advanced performance in each task.

1.2 Literature Review

We first give a brief review of the emotion recognition in Section 1.2.1, 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 the given person. In the computer vision area, human emotion is often defined as a set of discrete labels, including happiness, anger, sadness, surprise, disgust, and fear. To identify these labels, previous studies [4-10] mainly rely on facial analysis; thus, they have limited ability to precisely estimate the mental state of the human. To overcome these limitations, some methods adopt other visual clues, 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] adopt the body pose to identify six emotion categories. Chen *et al.* [16] propose a context fusion network to recognize human emotion by integrating events, objects, and scenes. Kosti *et al.* [17] present an end-to-end model for emotion recognition in context by jointly encoding the face and body information. However, these approaches are lacking in 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] present a two-stream architecture followed by a fusion network for CAER, see 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] propose a multi-model approach for the CAER task. Specifically, they exploit the

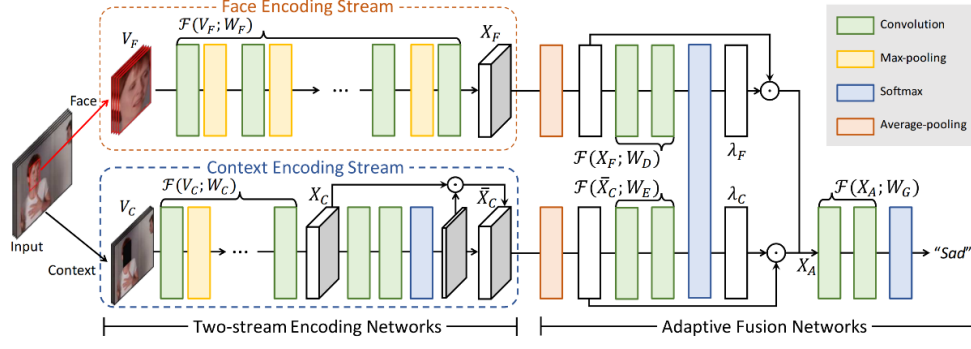


Figure 2. Network configuration of CAER-Net, consisting of two-stream encoding networks and adaptive fusion networks.

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

depth images to model socio-dynamic interactions and employ several sub-networks to infer the perceived emotion in the wild.

1.2.2 Depression Estimation

According to the machine learning perspective, depression estimation is essentially a regression problem. In the public benchmarks, such as AVEC 2013 [20] and AVEC 2014 [21], they annotate the depression level based on Beck Depression Inventory-II (BDI-II) [22], see 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 beginning, many studies employed hand-crafted methods to extract visual features. The baseline approach in AVEC 2013 [20] adopts Local Phase Quantization (LPQ) followed by a regressor for learning and

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

prediction. Unlike the single hand-crafted descriptor, Cummins *et al.* [23] exploit 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 generalize 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] design a multi-model framework to encode more meaningful features for depression estimation. They propose 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] propose a CNNs framework to model the visual information, including facial appearance and dynamics. Specifically, they design a two-stream architecture to encode both spatial and temporal information for learning and prediction. Ding *et al.* [26] propose a deep learning framework, called DeepInsight, to quickly diagnose the Autism Spectrum Disorder (ASD) and Major Depressive Disorder (MDD) based on CNNs. They design a multi-task learning framework for the diagnosis of ASD and MDD. By a share-weight backbone with a multi-scale design, they show an impressive performance on their own datasets.

1.3 Contributions

In this thesis, we design a multi-task learning framework to realize a mental disorder detection system for schizophrenia patients via emotion recognition and depression estimation. Our main contributions can be characterized as follows.

- I.** To better integrate the visual clues from different modalities, such as the face and context, we design a novel fusion network, namely Cross-Modality Graph Convolutional Networks (CMGCN). By taking the graph viewpoint, we build an affinity graph to describe the pixel-wise correlations between different modalities, then introduce a sampling scheme to intensify the sparsity of the given graph. By doing so, it can greatly suppress the interactions between irrelevant pixels and intensify the connections between relevant ones. Thus, it can result in a comprehensive representation for identification.
- 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. For emotion recognition, we propose Density Loss for metric learning with more comprehensive criteria relevant to the embedding density and orthogonal property. As the metric function is designed with the critical attributes, including embedding density and orthogonal property, we can 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 realizing better performance.
- III.** Following the natural relation between depression and emotion, we propose Mean Passer to effectively transfer the emotion prior 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 handbook of mental illness and the nature of schizophrenia, we propose an algorithm in detecting the mental disorder for schizophrenia patients. We first employ our multi-task model to identify the real-time mental state, including emotional state and depressive level. By recording the entire mental state of the patient, we adopt the sliding window scheme to analyze the abnormal patterns.

1.4 Thesis Organization

In this chapter, we first present the motivation of this thesis, 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 Networks (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 passing prior knowledge.

In Chapter 3, we first introduce the limitation of facial analysis, then elaborate on the proposed Cross-Modality Graph Convolutional Network (CMGCN) to integrate the visual clues 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 and orthogonal property, 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 Mean Passer, to transfer the prior knowledge. With Exponential Moving Average (EMA) scheme, we can take the smooth rather than protruding way to transfer the emotion prior knowledge to the depression model. Finally, we introduce how we implement the mental disorder detection algorithm based on cognitive science and the nature of schizophrenia.

In Chapter 4, the experimental results demonstrate the effectiveness of the proposed multi-task learning framework. To ensure each module is useful, we 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.

At last, in Chapter 5, we conclude the contributions of this thesis and 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 story of CNN is started from AlexNet [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 the classification and retrieval [27, 29], object detection and segmentation [30, 31], and action recognition [32, 33] tasks.

For the convenience presentation, we here adopt the image classification as an example to simplify the explanation. Convolution operation plays a fundamental basic 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, previous studies based on the hand-crafted

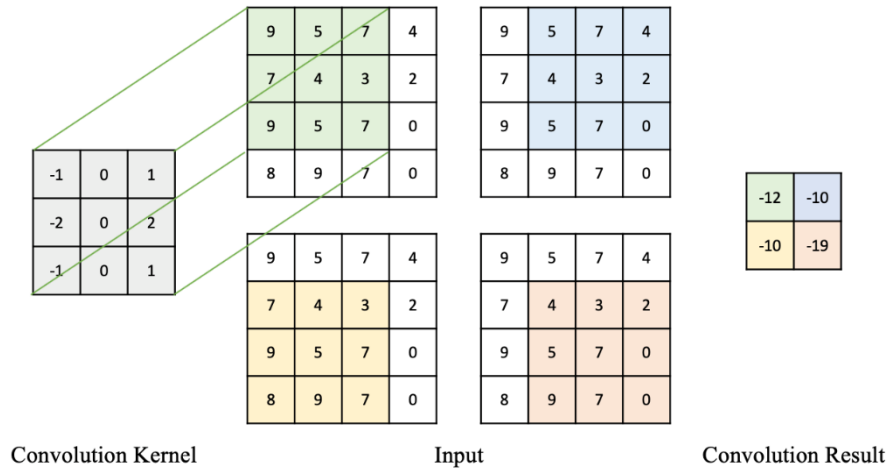


Figure 2-1. A standard 3×3 convolutional operation with stride 1.

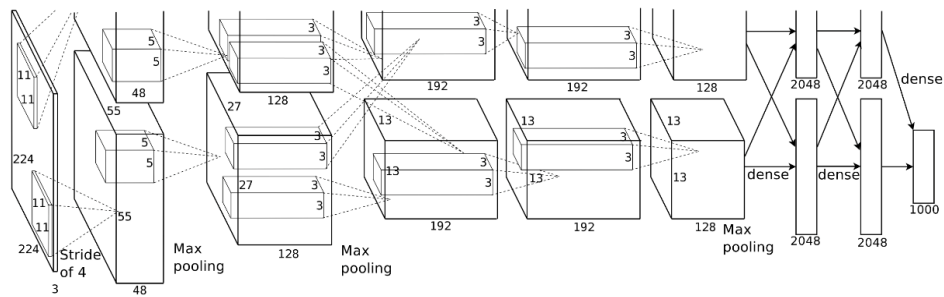
feature, such as SIFT [11] or HOG [12], require some domain knowledge to design a proper kernel of the convolution. However, it often requires many efforts of trial and error to search the most suitable hyper-parameters for real-world scenarios.

Departing from design the hand-crafted features, CNN instead 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 distribution of the target task, further extracting robust features. In Figure 2-1, we illustrate the basic concept of the convolution operation.

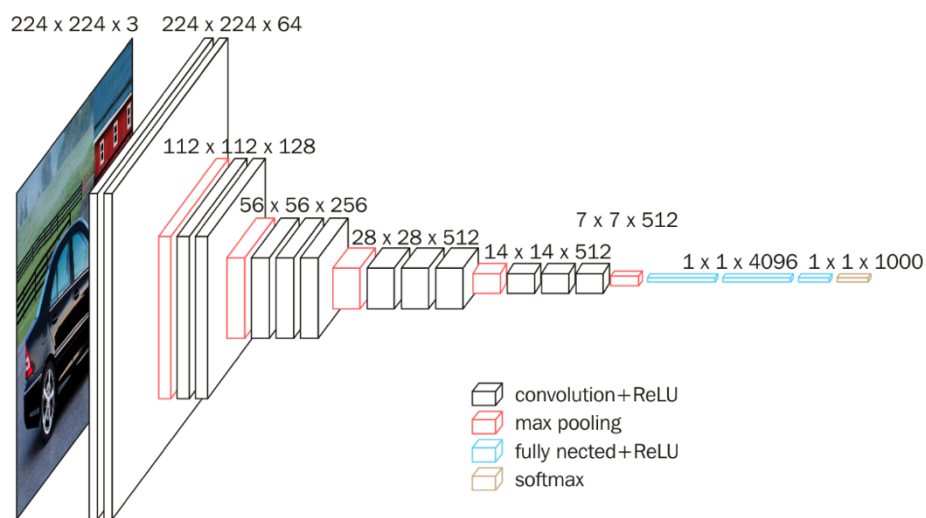
The typical Deep Neural Networks (DNN) take a node-to-node learning manner via several flatten layers to fit the distribution of the given task. Due to the huge parameters in the fully-connected layers, it often requires many computational costs for operation. By contrast, the convolutional layer of 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, while the deeper layers are responsible for yielding high-level features.

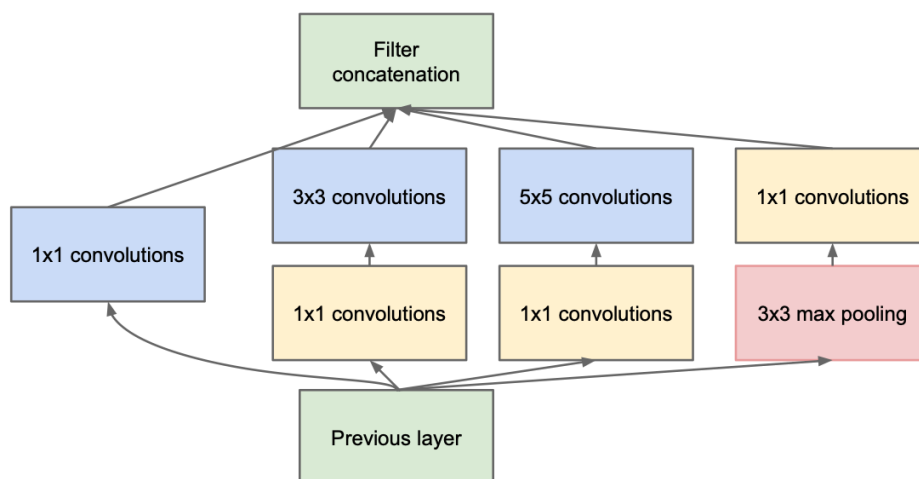
In 2014, Simonyan *et al.* propose VGG [34], which is a deeper CNN compared with AlexNet [27], and achieve the runner-up of ImageNet in that year. Google research team also proposes a famous Inception Net [35] in the same year. Unlike VGG, Inception Net aims to widen the structure and realize the championship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



(a) AlexNet



(b) VGG16



(c) Inception blocks

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

2.1.2 Residual Network

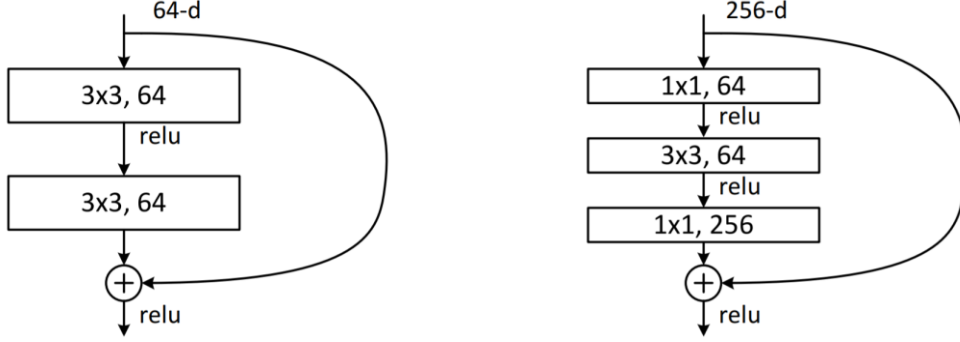


Figure 2-3: The difference between the flat convolutional layers and the residual connections.

According to Universal Approximation Theorem (UAT) [36], the feed-forward neural network has the ability to fit any non-linear mapping closely. Thus, researchers start to design a deeper network for strengthening the generalization and robustness. However, the vanishing gradient problem [37] 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 shrink too much, consequently resulting in a futile learning procedure.

To address this serious problem, He *et al.* [38]. propose 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:

$$output = F(x) + x \quad 2.1$$

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				
		$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv3_x	28×28	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 8$
conv4_x	14×14	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 23$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 36$
conv5_x	7×7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
	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 [38], 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.

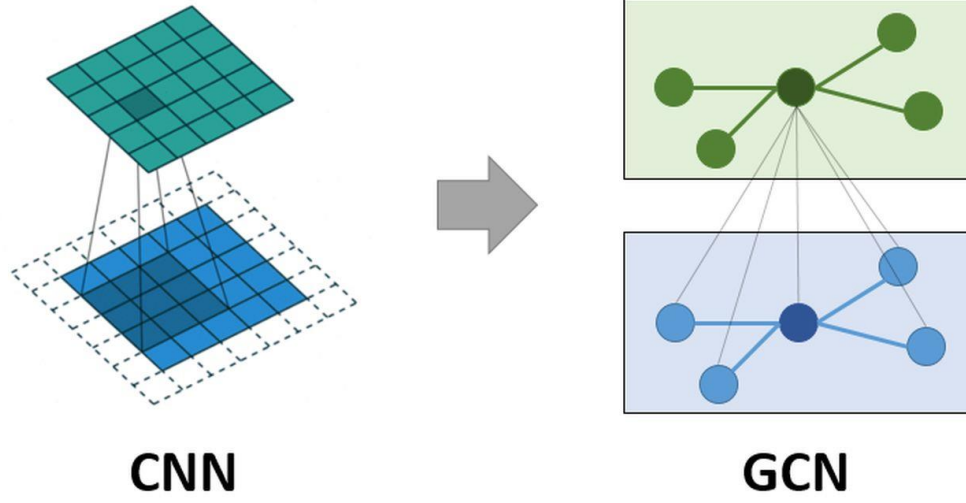


Figure 2-5: The difference between the convolution and the graph convolution.

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 node. According to the [39], the conventional GCN can be expressed as:

$$\begin{aligned}\hat{A} &= D^{-1}\hat{A}, \\ Z &= X(\hat{A}XW)\end{aligned}\tag{2.2}$$

Where \hat{A} denotes the adjacent matrix to describe the connectivity between nodes and D represents the degree matrix. 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 3.2.

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 and weight vectors. The latter often assigns training samples 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:

$$\mathcal{L}_{softmax} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{w_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^C e^{w_j^T x_i + b_j}} \quad 2.3$$

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.* propose Center Loss to further promote intra-class compactness [45]. This objective function exploits additional embedding centers to further congregate intra-class features. Specifically, it aims to minimize the distance between features and its corresponding center. Formally, Center Loss can be expressed as:

$$\mathcal{L}_{center} = -\frac{1}{2} \sum_{i=1}^N \|x_i - c_{y_i}\|_2^2 \quad 2.4$$

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

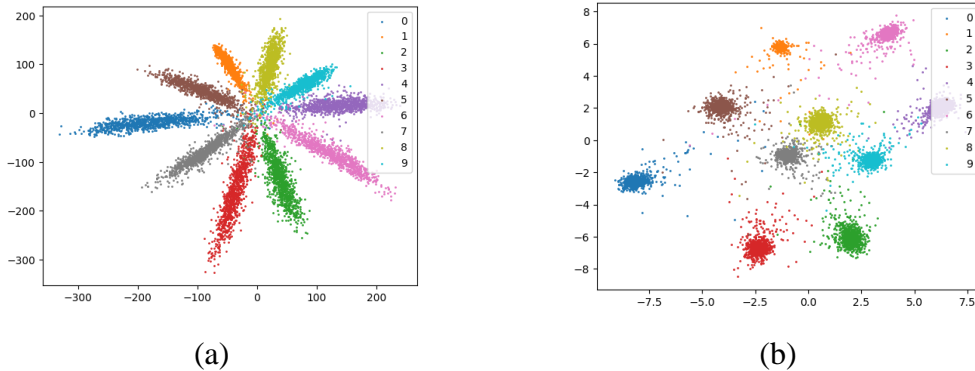


Figure 2-6: Visualized results of features training on MNIST [44]. (a) Softmax Loss may result in the embedding with low intra-class compactness. (b) With the assistance of Center Loss, the trained model can encode samples well, thereby achieving higher intra-class compactness.

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:

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

$$\mathcal{L}_{sphere_mapping} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s \cdot f_m(\theta_{y_i})}}{e^{s \cdot f_m(\theta_{y_i})} + \sum_{j=1, j \neq y_i}^C e^{s \cdot \cos \theta_j}} \quad 2.5$$

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:

$$\mathcal{L}_{reg} = \frac{1}{C} \sum_{i=1}^C \max_{j \neq i} \left\langle \frac{w_i}{\|w_i\|}, \frac{w_j}{\|w_j\|} \right\rangle \quad 2.6$$

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:

$$\mathcal{L}_{triplet} = \frac{1}{|\Gamma|} \sum_{(i,j,k) \in \Gamma} [d_{ij} - d_{ik} + m]_+ \quad 2.7$$

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, $d_{ij} = \|f(x_i^a) - f(x_j^p)\|_2$ 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.

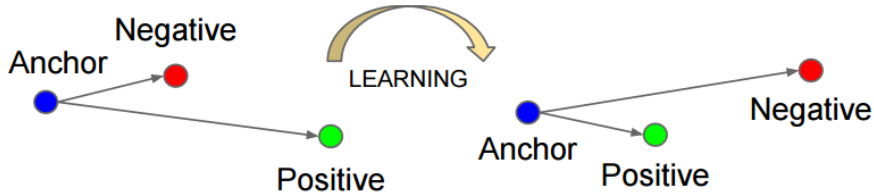


Figure 2-7: The basic idea of Triplet Loss.

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.* propose 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:

$$\mathcal{L}_{hard_triplet} = \frac{1}{N} \sum_{i=0}^N \sum_{j=1, j \neq i}^N \left[\max_{y_i=y_j} (d_{ij}^p) - \min_{y_i \neq y_j} (d_{ij}^n) + m \right]_+ \quad 2.8$$

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:

$$\mathcal{L}_{n-pair} = \frac{1}{N} \sum_i^N \log \left\{ 1 + \sum_{j \neq i} \exp(\langle f_i^a, f_j^n \rangle - \langle f_i^a, f_j^p \rangle) \right\} \quad 2.9$$

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:

$$\mathcal{L}_{lifted} = \frac{1}{2|P|} \sum_{(i,j) \in P} \left[d_{ij} + \log \left(\sum_{(l,k) \in N} \exp(\alpha - d_{lk}) + \sum_{(l,k) \in N} \exp(\alpha - d_{lk}) \right) \right]_+ \quad 2.10$$

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:

$$\mathcal{L}_{ranked} = \frac{1}{N} \sum_{i=1}^N \sum_{j \neq i}^N (1 - y_{ij}) w_{ij}^n [d_{ij} - \alpha]_+ + y_{ij} w_{ij}^p [d_{ij} - (\alpha - m)]_+ \quad 2.11$$

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:

$$\mathcal{L}_{ms} = \frac{1}{N} \sum_{i=1}^N \left\{ \frac{1}{\alpha} \log \left[1 + \sum_{k \in \mathcal{P}_i} e^{-\alpha(s_{ik}-\lambda)} \right] + \frac{1}{\beta} \log \left[1 + \sum_{k \in \mathcal{N}_i} e^{\beta(s_{ik}-\lambda)} \right] \right\} \quad 2.12$$

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:

$$\mathcal{L}_{circle} = \log \left[1 + \sum_{j=1}^L \exp(\gamma \alpha_j^n s_j^n) \sum_{i=1}^L \exp(-\gamma \alpha_i^p s_i^p) \right] \quad 2.13$$

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 boundaries and the embedding density,

we can form a useful regularization for metric learning to realize the robust embedding. We introduce the details of our Density Loss in Section 3.3.1.

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-tuning 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.

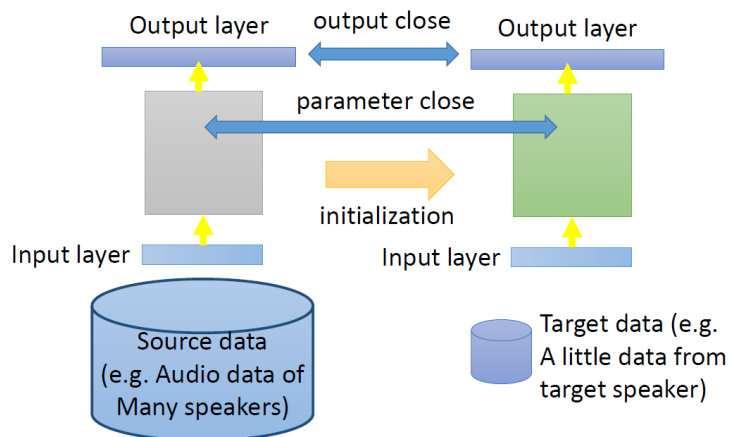


Figure 2-8: The basic idea of Fine-Tuning.

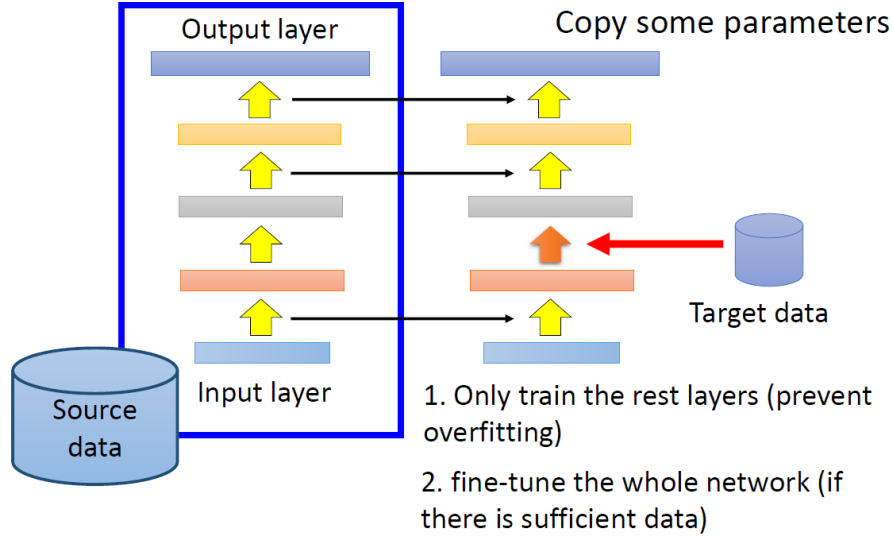


Figure 2-9: The basic idea of Layer Transfer.

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.

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 Mean Passer to realize the cross-task knowledge passing. By adopting EMA mechanism, our Mean Passer can transfer the knowledge with a fine way and realize a better performance on target domain.

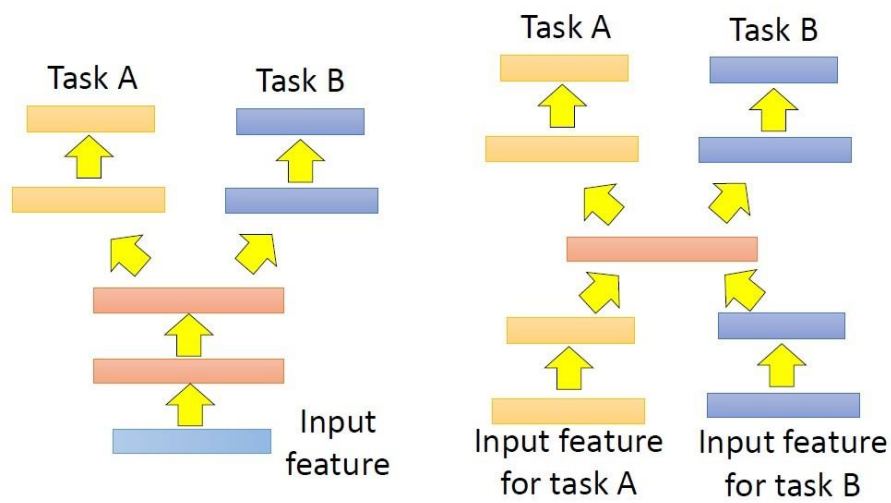


Figure 2-10: The basic idea of Multi-Task Learning.

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 briefly introduces the overview of our learning framework. In Section 3.2, we first elaborate on the shortcomings of facial analysis. Then, to overcome the limitation of facial analysis, we propose Cross-Modality Graph Convolutional Networks (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. The details of each objective function will be introduced in Section 3.3. Observe that depression is an extension of emotion, we propose Mean Passer to effectively transfer the emotion prior knowledge to the depression model in Section 3.4. Finally, in Section 3.5, we illustrate an algorithm to detect the mental disorder of schizophrenia patients, including Mood and Bipolar disorders. By the elaborated design, our algorithm accomplishes impressive performance in many public benchmarks.

3.1 Framework Overview

Our multi-task learning framework is illustrated in Figure 3-1. Our design mainly consists of four main components: 1) Cross-Modality Graph Convolutional Networks (CMGCN), 2) Density Loss for Emotion Recognition, 3) Distributed Loss for Depression Estimation, and 4) Mean 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 $X_f \in$

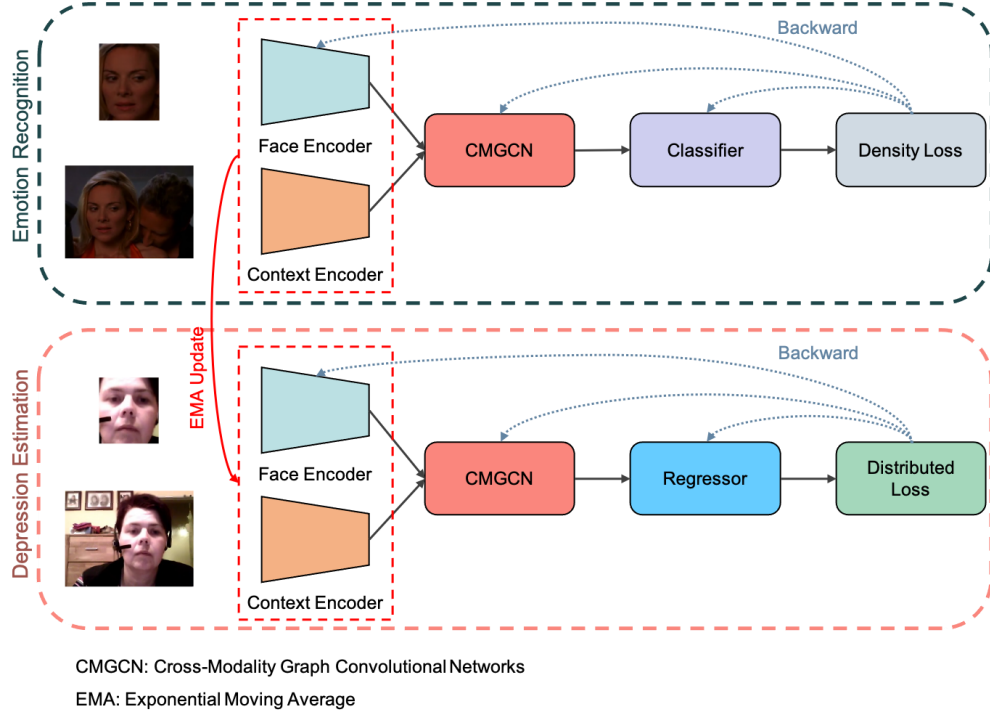


Figure 3-1: Our multi-task learning framework.

$\mathbb{R}^{N \times h \times w \times D}$ 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. 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, see Section 3.3. On the other hand, for cross-task knowledge passing, we present Mean Passer to effectively transfer the prior knowledge from the emotion model to the depression model via Exponential Moving Average (EMA) mechanism. With our well-design multi-task learning framework, our approach can successfully inhibit other state-of-the-art algorithms with a clear margin.

3.2 Cross-Modality Graph Convolutional Networks

To estimate the human mental state, previous studies in Facial Expression Recognition (FER) suppose that facial expression comprises the most discriminative emotional responses; thus, algorithms based on facial analysis have been extensively discussed. However, conventional FER systems often fail to infer the real-time emotional state and depressive level accurately. As we can see from Figure 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 the interactions of time series and the overall behaviors of human appearance. Therefore, it is important to design a fusion mechanism to integrate features from different modalities.



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].

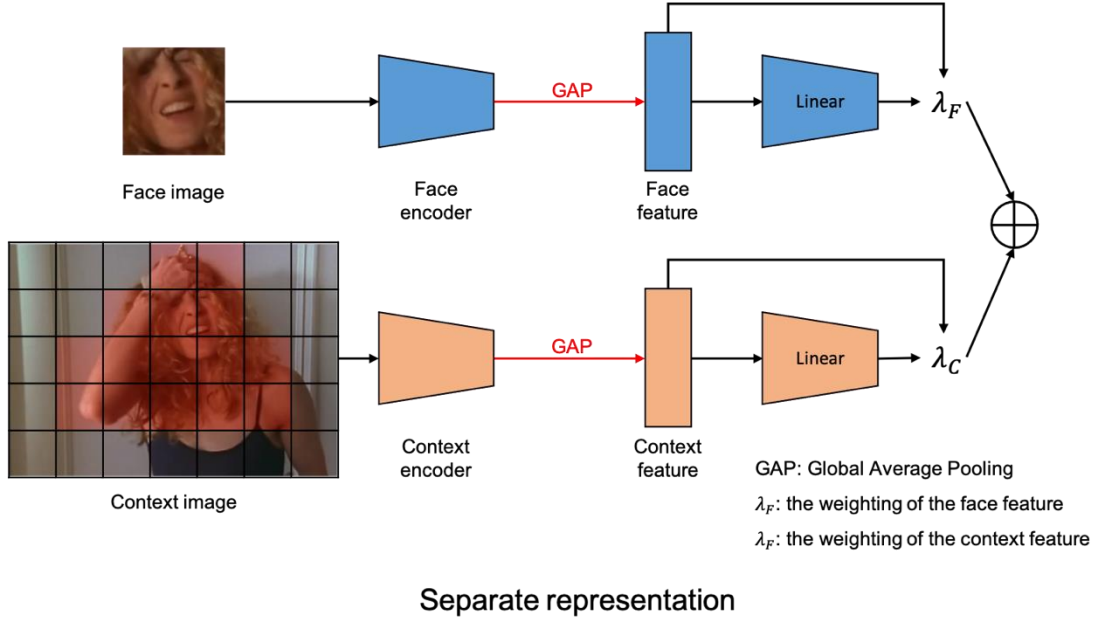


Figure 3-3: The existing separate representation [18].

An intuitive idea [18] is weighting the features from different modalities to emphasize the importance of each other separately. From Figure 3-3, the previous study adopts Global Average Pooling (GAP) to conclude the feature of each modality, then separately learns different weightings via several isolated linear layers (MLP). However, this fusion mechanism will include many irrelevant emotional regions. As we can see from the context image, it only involves a few critical regions (red pixels) for identification, while 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, 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 features from different modalities, see Figure 3-4. Particularly, we here exploit the graph viewpoint to model the correlations between different modalities so that we can learn a joint representation consistently. Since too many irrelevant regions are included in the context image, we develop a sampling scheme to build a sparse graph to keep relevant emotional features and significantly drop other irrelevant 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 way.

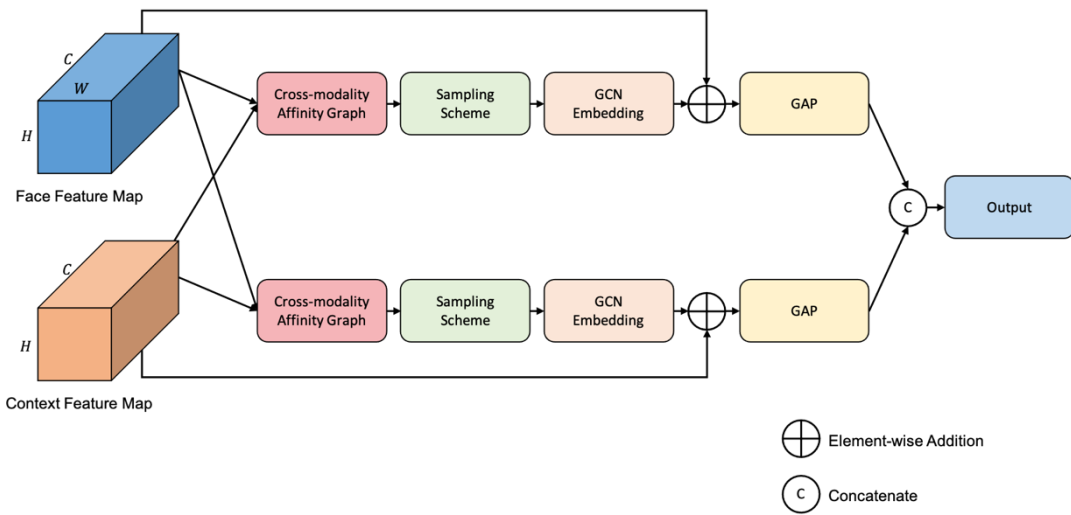


Figure 3-4: An overview of our CMGCN.

3.2.1 Cross-Modality Graph Construction

To encode the correlation between different modalities, we consider using an affinity graph to describe the pairwise relations of pixels crossed different modalities. Given a pair of the face and context features maps, the size of their tensor is shown as $H \times W \times C$; note that we reshape the feature maps into $HW \times C$ 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 is initialized via the cosine similarity. The edge can be formulated as:

$$g_{ij} = \left\langle \frac{x_i^f}{\|x_i^f\|}, \frac{x_j^c}{\|x_j^c\|} \right\rangle \quad 3.1$$

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

3.2.2 Sampling Scheme and GCN Embedding

Observe that the cross-modality graph \mathcal{G} is the fully-connected graph, which links the pairwise correlation of pixels among different modalities, *i.e.*, from face to context or from context to face. 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 irrelevant information, such as background pixels, see Figure 3-5. 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 3-5: 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 while others are background pixels.

Having obtained the graph $\mathcal{G} \in \mathbb{R}^{HW \times HW}$, we can now enhance its sparsity to reduce the influence of other irrelevant (non-target) features, see Figure 3-6. The key idea is to make relevant features have higher similarity (probability) values to be selected in the sampling process so that the GCN embedding could result from the legitimate weighting of combining relevant features. To this effect, we consider using the Bernoulli sampling, which responds to those elements with higher similarity in the given graph. 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 smoothing rather than protruding and achieve better performance. In our implementation, Bernoulli sampling 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 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:

$$\begin{aligned}
 M &= \text{Bernoulli}(\mathcal{G}), \\
 \mathcal{G}_{sparse} &= \mathcal{G} \odot M
 \end{aligned}
 \tag{3.2}$$

Where $\text{Bernoulli}(\cdot)$ denotes the Bernoulli sampling process, note that the sampling

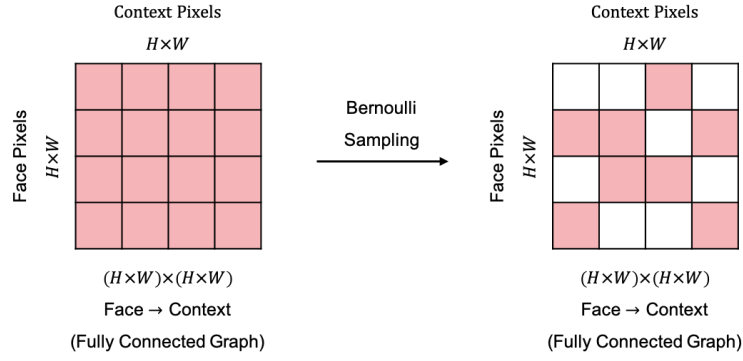


Figure 3-6: The concept of our sampling mechanism.

scheme will respond to each sample individually rather than affected by the entire graph.

\odot symbolizes the element-wise multiplication, M with elements either 0 or 1 represents the sampling result.

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 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:

$$\begin{aligned} \hat{A} &= D^{-1}\hat{A}, \\ X_G &= \sigma(\hat{A}XW) \end{aligned} \tag{3.3}$$

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

3.2.3 Fusion with Bidirectional Way

To acquire a comprehensive representation, we here consider a bidirectional way to explore critical regions among multi-modalities, see Figure 3-7. Specifically, we seek the high correlation 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 concatenate operation to yield the final representation. The final representation of our CMGCN can be expressed as follows:

$$X_{final} = [GAP(X_G^f + X^f), GAP(X_G^c + X^c)] \quad 3.4$$

Where X^f and X^c are face and context features. X_G^f and X_G^c indicate the graph feature of each modality based on Equation (3.3). $[\cdot, \cdot]$ denotes the concatenate operation to merge two given features.

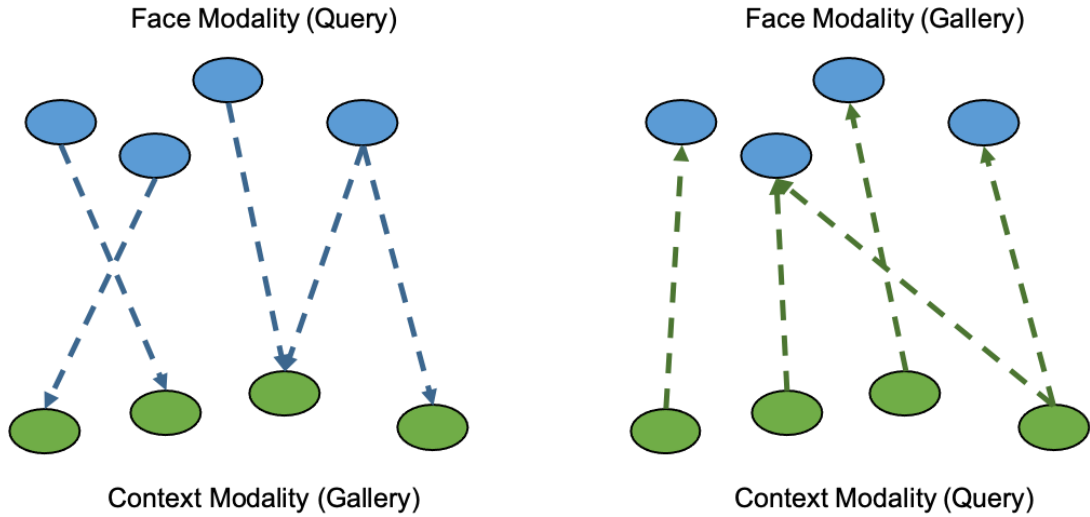


Figure 3-7: The bidirectional way to link relevant features.

3.3 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 then describe the *distributed loss* for depression estimation.

3.3.1 Density Loss for Emotion Recognition

Emotion recognition is essentially a classification task that focuses on learning a discriminative embedding to better identify the emotion from the given human images. To realize this objective, metric learning techniques are exploited to learn an embedding space with high intra-class compactness and inter-class separability. However, we observe that one essential property, embedding density, is often neglected in previous metric learning techniques. Assume that common metric learning can conglomerate each class of samples in an embedding space to a certain extent. The distribution of each class may still be sparse and with varied density, even when we apply the specific mining and weighting strategies to explore informative samples. To solve the issue, we consider enforcing the density prior of each class to form useful regularization for embedding learning.

A natural way to measure the class-wise density is to average all intra-class/inter-class pairs according to the given anchor. By averaging the similarity pairs, we can constitute a vector field to describe the sparsity of the measured class. Taking the intra-class perspective, once the similarity of a pair is less than the constructed vector field ($\mu_i^p > s_{ij}^p$), this sample contributes a degraded factor for the class density. Thus, our Density Loss exploits the disparity between the pair and density field to intensify the

penalty of this similarity pair. It is analogous for inter-class perspective under the opposite optimization situation ($\mu_i^n < s_{ij}^n$). Since always at least one sample is not satisfied (*i.e.*, less or higher than the field), it can penalize the outlier pairs based on the density adaptively and continuously (see Figure 3-8), no matter how close an underlying sample is. The emphasizing terms of the proposed density loss are defined as follows:

$$w_{ij} = \begin{cases} \exp([\mu_i^p - s_{ij}]_+), & \text{if } y_i = y_j \\ \exp([s_{ij} - \mu_i^n]_+), & \text{otherwise} \end{cases} \quad 3.5$$

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

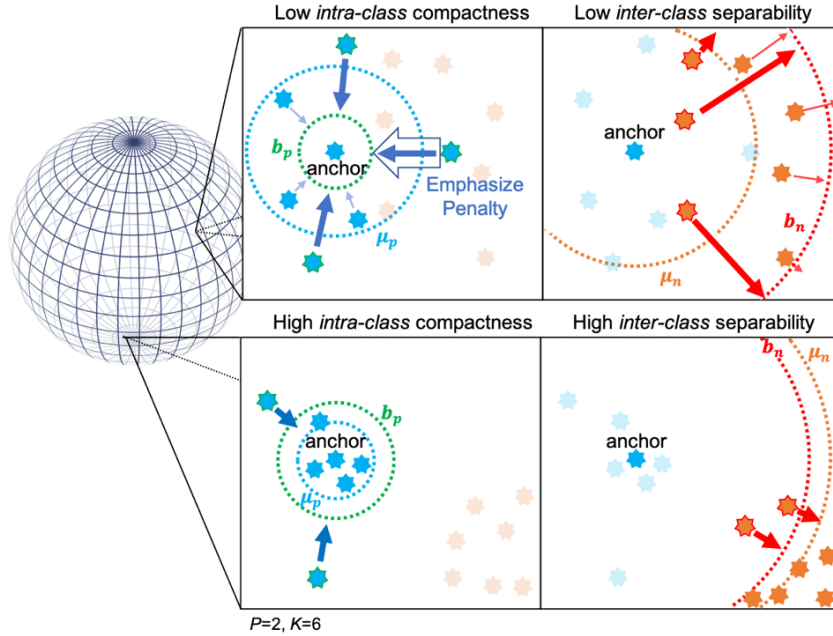


Figure 3-8: The vector fields of the proposed density loss. Even when the intra-class compactness is already high, the penalty will still be properly emphasized based on the intra-class density to promote feature discrimination.

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. Specifically, we consider two boundaries, intra-class b_p and inter-class b_n , to minimize the disparity between each type of the pair on its own. We expect each intra-class pair to be larger than the intra-class boundary and each inter-class pair to be less than the inter-class boundary. Further, to better regularize the distribution of each class, we here design the boundary with a corresponding relation, $b_n = 1 - b_p$. Because the cosine-similarity encodes each data point on a hyper-sphere, the intra-class boundary b_p can be regarded as a tolerated area of each class, and b_n denotes the distributed region of other classes to realize the orthogonal property.

$$L_{ij} = \begin{cases} [b_p - s_{ij}]_+, & \text{if } y_i = y_j \\ [s_{ij} - b_n]_+, & \text{otherwise} \end{cases} \quad 3.6$$

Finally, we multiply the loss l_{ij} with 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 follow:

$$\mathcal{L}_{density} = \frac{1}{N} \left(\sum_{i=1}^N (w_{ij} L_{ij})^p \right)^{\frac{1}{p}} \quad 3.7$$

where L_{ij} denotes the loss value that defines in Equation (3.6) and $p > 1$ specifies the underlying norm function.

3.3.2 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. 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 enlarge the loss value; thus, it cannot take the meticulous way to adjust the strength of the loss. 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., 0 to n . Thus, we can divide the learning target into $n + 1$ bins for learning, see Figure 3-9.

After formulating the loss function via a classification perspective, the learning target (regression value) becomes clear 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}

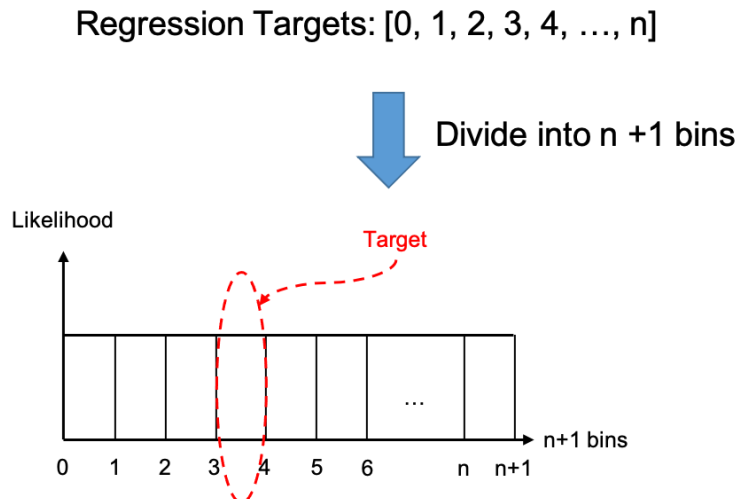


Figure 3-9: The illustration of the concept of our Distributed Loss. We introduce the classification viewpoint into the regression task.

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:

$$\mathcal{L}_{distributed} = -\log(s_{y_i}) - \log(s_{y_i+1}) \quad 3.8$$

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

Finally, we jointly adopt MSE Loss and the proposed Distributed Loss to realize better performance. A common strategy is learning with an isolated linear layer (MLP) for each objective function. But this strategy easily leads to inconsistent results. One MLP may dominate the model while another makes slightly effects; therefore, the predicted results of these two networks are different. Unlike this separate learning scheme, we here propose a joint head for prediction based on the You Only Look Once (YOLO) [63] architecture, see Figure 3-10. 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 take a more meticulous way to fine-tune the intensity of the loss value and realize better performance.

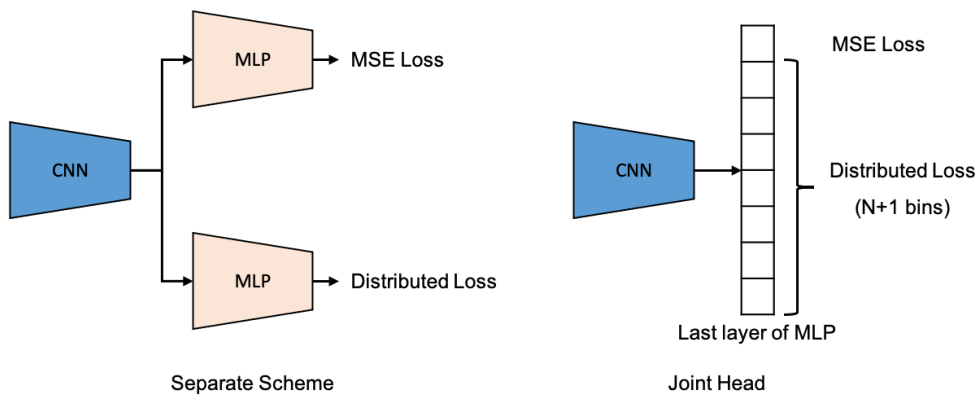


Figure 3-10: The difference between separate scheme and our joint head.

3.4 Mean Passer

In this section, we elaborate on the proposed training strategy, namely *mean passer*. As we mentioned above, depression is an extension of emotion and is a complex state with various emotions. A common training strategy for depression estimation is using several FER datasets (source domain) to pre-train the model then adopts the collected depression dataset (target domain) to fine-tune the model. However, this strategy often requires to carefully select 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 inefficiency due to sequentially training with several source datasets.

To promote training efficiency, our Mean Passer takes the epoch-wise viewpoint to transfer the knowledge. Specifically, in each iteration (epoch), we first train the emotion model. 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 passing procedure is expected to be smoothing rather than protruding. Thus, it can realize a better performance. Formally, the knowledge transfer of Mean Passer can be expressed:

$$\theta_i^d = \alpha\theta_i^d + (1 - \alpha)\theta_i^e \quad 3.9$$

Where α is a smoothing coefficient parameter based on the current iteration i . d and e denote the depression and emotion models, respectively. θ denotes the parameters of the model.

3.5 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 for schizophrenia patients.

Schizophrenia is a psychiatric disorder characterized by continuous or relapsing episodes of psychosis [1]. Based on the handbook about mental illness [2, 3], the mental state of schizophrenia patients can be evaluated by Bipolar disorder and Depression disorder. Concretely, Bipolar disorder is an unstable emotional condition characterized by cycles of abnormal, persistent high mood and low mood. Besides, patients may feel abnormally energetic, happy, or irritable. On the other hand, Depression disorder is a low emotion state, and patients often stay in pervasive sadness and depression.

As we have mentioned before, Bipolar disorder and Depression disorder are important references for doctors to estimate the mental state of schizophrenia patients. As a matter of fact, patients with Bipolar disorder may behave in various moods in a short time-series. On the other hand, Depression disorder may cause patients to stay in a low mood, such as sadness or high depression. Since the above disorders are highly correlated to emotion, we can draw on the emotion recognition and depression estimation to identify the real-time mental state of patients to infer the mental disorder further. Further, in cognitive science, emotion is defined as a state of readiness for the action when humans face stimulus events [13, 14]. Essentially, emotion is an action readiness characterized by happening quickly, short duration, and non-periodic happening. By statistics [64, 65], the duration of emotion is 0.05 to 0.5 seconds. To this effect, we here consider using the prediction of 2 seconds for analysis. Specifically, in our implementation, we use the

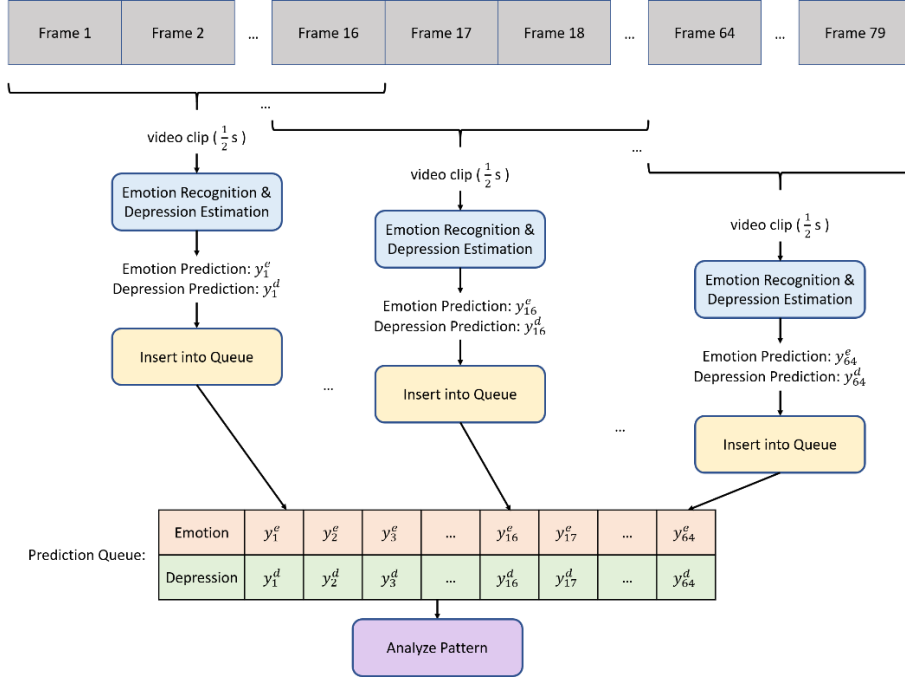


Figure 3-11: Illustration of our detection framework. We adopt a camera with 32 fps to record the video. Then, we build a video clip with 16 frames (0.5 seconds) for identification, including emotion recognition and depression estimation. Finally, we will gather 64 predictions (2 seconds) for analysis to detect the abnormal pattern.

camera with 32 frames per second (fps) to record the video and adopt the sliding window strategy to sequentially capture frames for identification, see Figure 3-11.

Having obtained the sequential predictions, we then dissect the recording to detect the abnormal pattern. For the emotion analysis, we here calculate the entropy of the given emotional predictions to measure disorders. If the entropy is high, the emotion of each class will distribute uniformly; thus, it denotes patients with various emotions in a short time serious. We will give a Bipolar disorder alert to notify doctors. In contrast, if the entropy is low, patients will stay in a stable emotion. Once patients express a stable emotion, we can further check the emotional state. If patients stay in negative emotions, such as sadness, fear, or anger, we will give a Depression disorder alert to notify doctors. For the depression analysis, based on the Beck Depression Inventory-II (BDI-II) [22], our

depression module will output the score with 45 levels, and the BDI-II scores 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. The detection flow is shown on

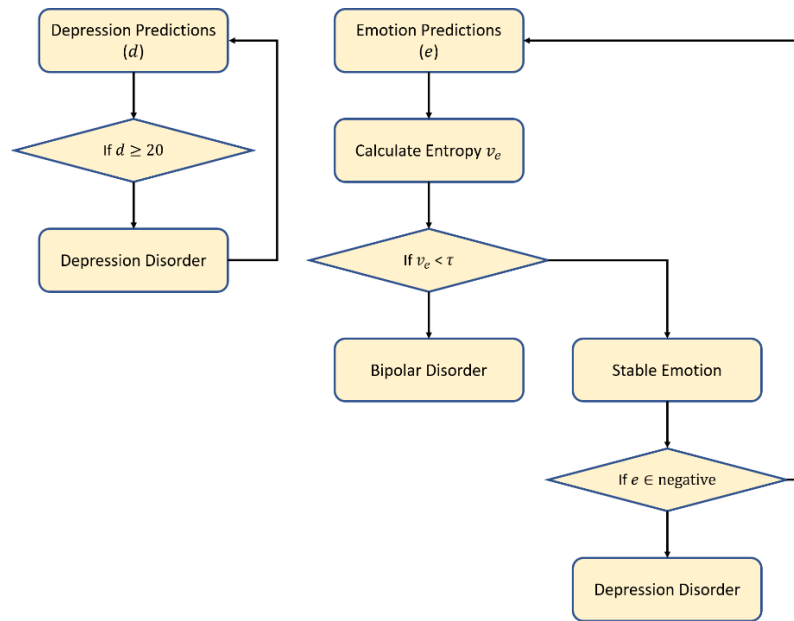


Figure 3-12: Our mental disorder detection flow.

Chapter 4 Experiments

In this chapter, we will first introduce the environment setting in Section 4.1, followed by the implementation details in Section 4.2.

4.1 Configuration

In 錯誤! 找不到參照來源。 , 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 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 with learning rate initialized as 10^{-3} and dropped by a factor of 10 every 40 epochs. We following 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 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.

On the other hand, to validate the effectiveness of the proposed Density Loss for metric learning. We additionally do experiments on the famous fine-grained image retrieval benchmarks, including the CUB [66], Cars-196 [67], Stanford Online Products (SOP) [54], and In-Shop Clothes (In-Shop) [68]. The first two datasets are the small retrieval benchmarks, and the remaining benchmarks are the large-scale datasets. Here, we follow the same settings as [56, 58] for a fair comparison.

4.3 Datasets

We evaluate our learning framework on two public datasets, including CAER [18] and AVEC 14 [21], both of which are the most famous and challenging. These two datasets will be introduced in Section 4.3.1 and Section 4.3.2, and both of them are highly related to the mental state of humans. On the other hand, to evaluate the effectiveness of the proposed Density Loss, we select four fine-grained retrieval benchmarks, including CUB [66], Cars-196 [67], Stanford Online Products (SOP) [54], and In-Shop Clothes (In-Shop) [68]. These fine-grained retrieval datasets will be introduced in Section 0. Evaluation metrics such as Cumulative Match Characteristic (CMC) curve, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) will be mentioned in the end.

4.3.1 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 of sequence length 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 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

4.3.2 Audio-Visual Emotion recognition Challenge 2014 (AVEC 14)

Dataset

AVEC 14 depression database 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, the Beck Depression Inventory-II (BDI-II [22]), see Table 1-1. Some example images of this benchmark are shown in Figure 4-2.



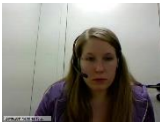
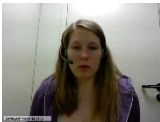






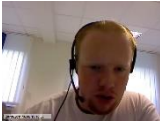





Frames				BDI II Score	Depression Severity
				0	None
				15	Mild
				24	Moderate
				44	Severe

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

4.3.3 Fine-Grained retrieval benchmarks

Table 4-3: Details of each fine-grained retrieval benchmark.

Benchmark	Category	# of images
CUB	200	11,788
Cars-196	196	16,185
SOP	22,634	120,053
In-Shop	7,986	72,712

We select four fine-grained retrieval benchmarks to evaluate the effectiveness of our method, including CUB [66], Cars-196 [67], Stanford Online Products (SOP) [54], and In-Shop Clothes (In-Shop) [68]. The first two datasets are the small retrieval benchmarks, and the remaining benchmarks are the large-scale datasets. Here, we follow the same settings as [56, 58] for a fair comparison. For each benchmark, the first half of classes are roughly used for training and the remaining classes are for testing. The details of each dataset are shown in Table 4-3.

4.3.4 Evaluation Metrics

In this thesis, we evaluate our method by using Cumulative Match Characteristic (CMC) curve, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). For emotion recognition, a classification task, we employ CMC curve to estimate the model performance. From Figure 4-3, the area under the curve (AUC) represents the performance of the classifier. Generally, CMC curve is the one common metric for the classification task. According to the top- k error, the evaluation metric can be expressed as:

$$\begin{aligned}\text{Rank-1} &= 1 - \text{top-1 error} \\ \text{Rank-5} &= 1 - \text{top-5 error}\end{aligned}\tag{4.1}$$

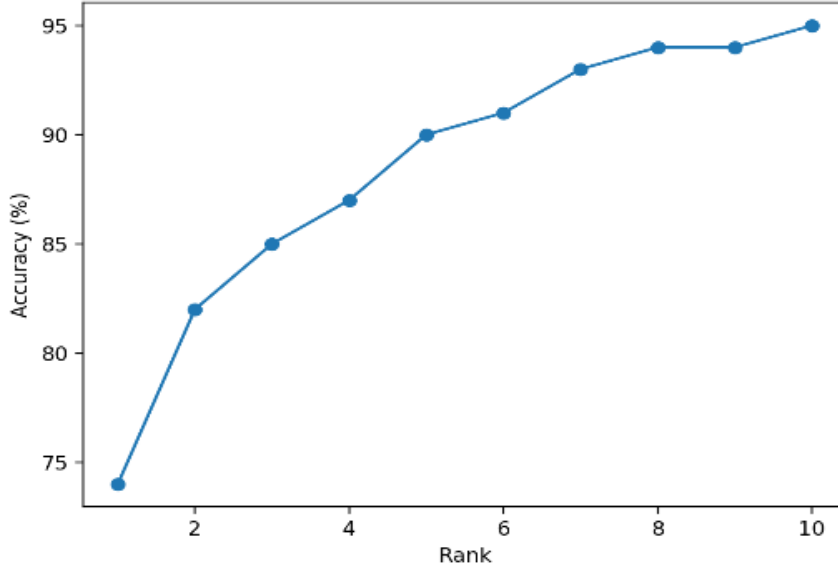


Figure 4-3: An example of Cumulative Match Characteristic (CMC) curve.

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

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

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad 4.2$$

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.

4.4 Ablation Studies

To verify the effectiveness of the proposed method, we conduct a series of ablation studies of them in the following paragraph. We first investigate the influence of our CMGCN, 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 Mean Passer for knowledge transfer.

4.4.1 The influence of CMGCN

First of all, we detailed investigate our CMGCN with different modalities, and the results are shown in Table 4-4. For a fair comparison, we set Density Loss for emotion recognition and Distributed Loss for depression estimation. In addition, we employ mean passer to transfer the emotion prior knowledge to depression model. For baseline model, we directly concatenate the features from different modalities and achieve 70.09% in accuracy on CAER and 12.40/14.80 in MAE/RMSE on AVEC 14. Face indicates that we only apply CMGCN on the face modality alone; Context indicates that we only apply CMGCN on the context modality alone; Face + Context 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 for face modality, it focuses on linking relevant pixels and dropping the background pixels in the face modality. 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 clues, 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

Table 4-4: 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

the scene is in the laboratory, the environment is not a serious problem to detect the correct face. Thus, it 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.

We then discuss the influence of the sampling scheme. As we can see from Table 4-5, \mathcal{G}_{full} denotes that we do not impose the sampling scheme to connect/drop relevant/irrelevant pixels. \mathcal{G}_{topk} denotes that we select the top 5 high entries from another modality for each pixel. \mathcal{G}_ϵ denotes that we adopt the epsilon ball with 0.5 threshold to link the relevant pixels. $\mathcal{G}_{bernoulli}$ denotes that 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

Table 4-5: 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}_{ϵ}	86.36	7.6340	9.4368
$\mathcal{G}_{bernoulli}$	87.23	6.8206	8.5078

as \mathcal{G}_{topk} , \mathcal{G}_{ϵ} , and $\mathcal{G}_{bernoulli}$, we can realize a better performance, especially for AVEC 14. Because \mathcal{G}_{topk} connects all relevant entries for all pixels, it will perform worse on CAER due to considering other irrelevant information. Comparing $\mathcal{G}_{bernoulli}$ with \mathcal{G}_{ϵ} , Bernoulli sampling can achieve the better performance due to its sampling nature. Such the sampling scheme will not completely ignore lower similarity elements; 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.

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 4-6, for the Baseline group, we consider the A-Softmax Loss with $s = 20$ to train the entire model, more details are shown in Section 2.2.1. \mathcal{L}_{tri} and $\mathcal{L}_{density}$ respectively indicate the Hard Triplet Loss and our Density Loss. With the pairwise metric learning, it can greatly improve the performance. Since the A-Softmax Loss impose an approximated weight matrix to guide representations in the embedding space, it may not well promote

Table 4-6: Comparison of our Density Loss.

Method	CAER
	Accuracy
Baseline	81.75
Baseline + \mathcal{L}_{tri}	85.26
Baseline + $\mathcal{L}_{density}$	87.23

the intra-class compactness during the early epochs. By the margin constraint of \mathcal{L}_{tri} , the intra-class compactness can be promoted and improves the with 3.51% in accuracy on CAER. However, \mathcal{L}_{tri} may easily cause ambiguous optimization results, see Section 2.2.2. To alleviate this drawback, our Density considers the strict boundary for intra-class and inter-class pairs relevant to orthogonal property. By doing so, it can greatly avoid ambiguity. Further, as our density loss is designed with the embedding density, it can adaptively emphasize the loss of each class, and consequently forming a useful regularization for embedding learning. With 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 4-7, \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 to jointly learn with \mathcal{L}_{MSE} and \mathcal{L}_{Distri} .

Table 4-7: 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 4-7, 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 compared with \mathcal{L}_{MSE} . By the two isolated MLP, we can take a simple approach to combine these two losses; however, the performance is only slightly improved. Because the MLP is separately 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 alleviate this ambiguity, we follow YOLO to design a joint head for prediction; thus, the performance can be greatly improved due to the joint learning manner.

4.4.4 The influence of Mean Passer and Joint Head

To validate the effectiveness of our Mean Passer for knowledge transfer, we here train the emotion model and depression model with two scenarios, including learning without Mean Passer and learning with Mean Passer. As we can see from Table 4-8, if the depression model is randomly initialized, it will not easy to converge. By our Mean Passer, since the emotion prior knowledge is smoothly transferring to the depression model, it can effectively promote the training procedure and result in better performance.

Table 4-8: Comparison of our Mean Passer.

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

4.5 Comparing with State-Of-The-Arts (SOTA)

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 4-9, we can see that the deeper backbones, such as ResNet [38], can achieve the better performance compared with 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 layers to respectively predict the weights for each modality, this fusion mechanism may lead to improper results. When the face is incorrect, the fusion weights cannot represent the importance of the face feature. Other methods [69] consider involving temporal information 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.

Unlike the above methods, our CMGCN exploits the sampling scheme to constitute a sparse graph to describe the correlations of relevant pixels 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 beat 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 different modalities

Table 4-9: Comparison of the SOTA on CAER.

Method	Data type	Modality	CAER
			Accuracy
ImageNet-AlexNet [27]	Image	Face + Context	47.36
ImageNet-ResNet [38]	Image	Face + Context	57.33
Fine-tuned AlexNet [27]	Image	Face + Context	61.73
Fine-tuned ResNet [38]	Image	Face + Context	68.46
CAER-Net-S [18]	Image	Face	70.09
		Context	65.65
		Face + Context	73.51
Sports-1M-C3D [69]	Video	Face + Context	66.38
Fine-tuned C3D [69]	Video	Face + Context	71.02
CAER-Net [18]	Video	Face	74.13
		Context	75.57
		Face + Context	77.04
Ours	Image	Face + Context	87.23

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 4-10, our depression model outperforms all other SOTA methods with 6.82/8.5 on MAE/RMSE. As we can see, other methods [21, 25, 70-74] 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 [74], 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 generalize to unseen data and often perform worse compared than the learning-based methods [25, 73, 74].

Table 4-10: Comparison of the SOTA on AVEC 14.

Method	Data type	Modality	AVEC 14	
			MAE	RMSE
Baseline [21]	Video	Face	8.86	10.86
UUISidorov [70]	Video	Face	11.20	13.87
InaoeBuap [71]	Video	Face	9.35	11.91
Brunel [72]	Video	Face	8.44	10.50
BU-CMPE [73]	Video	Face	7.96	9.97
DCNN [25]	Video	Face	7.47	9.55
RNN-C3D [74]	Video	Face	7.22	9.20
Ours	Image	Face + Context	6.82	8.50

4.5.3 The result on Fine-Grained Image retrieval benchmarks

In addition, we here compared the proposed Density Loss with other pairwise metric learning techniques. For a fair comparison, we follow the same evaluation rules on Fine-Grained Image retrieval benchmarks, including CUB, Cars-196, SOP, and In-Shop. From Table 4-11 and Table 4-12, we discuss pairwise metric learning techniques with our density loss. The popular learning manner [56, 75] can be viewed as a way to reduce $(s_n - s_p)$, see Section 2.2.2; thus, it may result in various sparsity of each class since the distribution of each group is not well regulated. Unlike popular learning manners, our Density Loss learns with the boundary to regularize the embedding distribution, $[b_p - s_p]_+$ and $[s_n - b_n]_+$, and design the b_p and b_n with a corresponding relation, $b_p = 1 - b_n$. Thus, it can better restrict the distribution of each class and increases R@1 by 1.8% and 1.4% on the CUB and Cas-196 compared with MS loss [56], respectively. Circle loss [58] designs a self-paced weighting to form a better decision boundary for embedding learning; however, it may not well congregates the distribution of each class into a hyper-ball. Since we train the model with a random sampling strategy, the

Table 4-11: Comparison of the SOTA on CUB and Cars-196.

Method	CUB				Cars-196			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
RLL-H [55]	54.7	69.7	79.2	86.9	74.0	83.6	90.1	94.1
HDC [75]	60.7	72.4	81.9	89.2	83.8	89.8	93.6	96.2
SoftTriple [76]	65.4	76.4	84.5	90.4	84.5	90.7	94.5	96.9
MS Loss [56]	65.7	77.0	86.6	91.2	84.1	90.4	94.0	96.5
Circle Loss [58]	66.7	77.4	86.2	91.2	83.4	89.8	94.1	96.5
Density Loss	67.5	77.9	86.0	91.7	85.5	90.8	94.7	97.2

embedding distribution may be sparsity, especially for the edge of each class. Our Density Loss emphasizes the loss of each pair based on the density of each intra-class and inter-class pairs; thus, it can further congregate intra-class features into a more compact hyperball and greatly smooth the burdens of random sampling. Comparing with Circle Loss [58], our Density Loss increases R@1 by 0.8% and 2.1% on the CUB and Cas-196.

For the large-scale benchmark, such as SOP and In-Shop, we report the performance in Table 4-12. Density Loss only slightly improves the Circle loss and MS loss by 0.2% and 0.3% at R@1 performance on SOP, respectively. It denotes that metric learning loss seems not the essential property for this dataset, and it requires more specific embedding strategies to improve the performance. For In-shop benchmark, our Density Loss already surpasses the pairwise metric learning with a clear margin, 89.9% \rightarrow 92.5% at R@1 performance. The results show that the density and orthogonal property are the crux for embedding learning in this dataset.

Table 4-12: Comparison of the SOTA on SOP and In-Shop.

Method	SOP			In-Shop		
	R@1	R@10	R@10 ²	R@1	R@10	R@20
HDC [75]	75.9	88.4	94.9	62.1	84.9	89.0
RLL-H [55]	76.1	89.1	95.4	--	--	--
MS Loss [56]	78.2	90.5	96.0	89.7	97.9	98.5
Circle Loss [58]	78.3	90.5	96.1	--	--	--
Density Loss	78.5	90.0	95.6	92.5	98.2	98.8

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 record the entire mental state of patients, further detecting the abnormal patterns from the given recording.

To precisely estimate the emotional state 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 enforces the sampling scheme to build the sparse graph. Since the sampling scheme can connect the relevant pixels of different modalities 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 the depression is an extension of the emotion; we also propose a knowledge transfer scheme, Mean Passer. For each mini-batch, our Mean 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 Mean 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 outperform all advanced SOTA methods. In addition, we also validate the effectiveness of our Density Loss in several fine-grained retrieval benchmarks. Compared with the most advanced metric learning techniques, which indicate Circle Loss and MS Loss, our Density Loss can beat them with a clear margin 0.8% on CUB, 1.4% on Cars-196, 0.2% on SOP, and 2.6% on In-Shop. In future work, we plan to adopt our method on the dataset collected by NTU hospital and impose more domain knowledge to design the model for real-world scenarios.

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," *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," *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, 2004. CIT '04.*, 16-16 Sept. 2004 2004, pp. 178-183.
- [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, no. 8, 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, no. 1, 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)*, 10-12 June 2020 2020, pp. 1037-1041.
- [11] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," vol. 60, no. 2 %J Int. J. Comput. Vision, pp. 91–110, 2004.
- [12] A. Kläser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in *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," 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 9, pp. 1238-46, 2008.
- [16] C. Chen, Z. Wu, and Y.-G. Jiang, "Emotion in Context: Deep Semantic Feature

- Fusion for Video Emotion Recognition," *Proceedings of the 24th ACM international conference on Multimedia*, 2016.
- [17] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion Recognition in Context," *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)*, 2019, pp. 10142-10151.
 - [19] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frequency Principle," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14222-14231, 2020.
 - [20] M. Valstar *et al.*, "AVEC 2013: the continuous audio/visual emotion and depression recognition challenge," presented at the Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, Barcelona, Spain, 2013.
 - [21] M. Valstar *et al.*, "AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge," presented at the 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. 588-97, 1996.
 - [23] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals: a multimodal approach," *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," *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 *BMVC*, 2018.
 - [27] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems*, vol. 25, 01/01 2012.
 - [28] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211-252, 2015/12/01 2015.
 - [29] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Computer Vision – ECCV 2014*, Cham, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., 2014: Springer International Publishing, pp. 818-833.
 - [30] R. B. Girshick, J. Donahue, T. Darrell, J. J. I. C. o. C. V. Malik, and P. Recognition, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," pp. 580-587, 2014.
 - [31] A. Bochkovskiy, C.-Y. Wang, and H. J. A. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," vol. abs/2004.10934, 2020.
 - [32] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for

- Action Recognition in Videos," in *NIPS*, 2014.
- [33] C. Feichtenhofer, A. Pinz, A. J. I. C. o. C. V. Zisserman, and P. Recognition, "Convolutional Two-Stream Network Fusion for Video Action Recognition," pp. 1933-1941, 2016.
 - [34] K. Simonyan and A. J. C. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," vol. abs/1409.1556, 2015.
 - [35] C. Szegedy *et al.*, "Going deeper with convolutions," pp. 1-9, 2015.
 - [36] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359-366, 1989/01/01/ 1989.
 - [37] 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, 04/01 1998.
 - [38] K. He, X. Zhang, S. Ren, J. J. I. C. o. C. V. Sun, and P. Recognition, "Deep Residual Learning for Image Recognition," pp. 770-778, 2016.
 - [39] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," 2017.
 - [40] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," in *NIPS*, 2017.
 - [41] G. Li, M. M. ller, A. K. Thabet, and B. Ghanem, "DeepGCNs: Can GCNs Go As Deep As CNNs?," *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)*, 13-19 June 2020 2020, pp. 5660-5669.
 - [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)*, 13-19 June 2020 2020, pp. 389-398.
 - [44] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, vol. 86, no. 11, pp. 2278-2324.
 - [45] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition," in *Computer Vision – ECCV 2016*, Cham, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., 2016: Springer International Publishing, pp. 499-515.
 - [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)*, 21-26 July 2017 2017, pp. 6738-6746.
 - [47] H. Wang *et al.*, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18-23 June 2018 2018, pp. 5265-5274.
 - [48] J. Deng and S. Zaferirou, "ArcFace for Disguised Face Recognition," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 27-28 Oct. 2019 2019, pp. 485-493.
 - [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)*, 15-20 June 2019 2019, pp. 1136-1144.

- [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)*, 22-29 Oct. 2017 2017, pp. 2859-2867.
- [51] F. Schroff, D. Kalenichenko, J. J. I. C. o. C. V. Philbin, and P. Recognition, "FaceNet: A unified embedding for face recognition and clustering," pp. 815-823, 2015.
- [52] A. Hermans, L. Beyer, and B. J. A. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," 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*, D. L. a. M. S. a. U. L. a. I. G. a. R. Garnett, Ed., 2016: Curran Associates, Inc., pp. 1857--1865.
- [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)*, 27-30 June 2016 2016, pp. 4004-4012.
- [55] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. Robertson, "Ranked List Loss for Deep Metric Learning," *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," *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 *et al.*, "Circle Loss: A Unified Perspective of Pair Similarity Optimization," *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. Frank 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*, 2018, pp. 172-178.
- [61] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," *arXiv preprint arXiv:1611.05244*, 2016.
- [62] P. Peng *et al.*, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1306-1315.
- [63] J. Redmon, S. Divvala, R. B. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779-788, 2016.
- [64] E. Svetieva and M. G. Frank, "Empathy, emotion dysregulation, and enhanced microexpression recognition ability," *Motivation and Emotion*, vol. 40, pp. 309-320, 2016.
- [65] C. Hurley, A. E. Anker, M. G. Frank, D. Matsumoto, and H. C. Hwang, "Background factors predicting accuracy and improvement in micro expression recognition," *Motivation and Emotion*, vol. 38, pp. 700-714, 2014.
- [66] C. Wah, S. Branson, P. Welinder, P. Perona, and S. J. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," 2011.
- [67] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D Object Representations for

- Fine-Grained Categorization," in *2013 IEEE International Conference on Computer Vision Workshops*, 2-8 Dec. 2013 2013, pp. 554-561.
- [68] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016 2016, pp. 1096-1104.
 - [69] 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)*, 7-13 Dec. 2015 2015, pp. 4489-4497.
 - [70] M. Sidorov and W. Minker, "Emotion Recognition and Depression Diagnosis by Acoustic and Visual Features: A Multimodal Approach," in *AVEC '14*, 2014.
 - [71] 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 *AVEC '14*, 2014.
 - [72] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic Depression Scale Prediction using Facial Expression Dynamics and Regression," in *AVEC '14*, 2014.
 - [73] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble CCA for Continuous Emotion Prediction," in *AVEC '14*, 2014.
 - [74] 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, doi: 10.1109/TAFFC.2018.2870884.
 - [75] H. O. Song, S. Jegelka, V. Rathod, and K. Murphy, "Deep Metric Learning via Facility Location," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 21-26 July 2017 2017, pp. 2206-2214.
 - [76] Q. Qian *et al.*, "SoftTriple Loss: Deep Metric Learning Without Triplet Sampling," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 27 Oct.-2 Nov. 2019 2019, pp. 6449-6457, doi: 10.1109/ICCV.2019.00655.