國立臺灣大學電機資訊學院電機工程學研究所

碩士論文

Graduate Institute of Electrical Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

通過深度視覺感知對思覺失調症患者的心理障礙檢測

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

1. 誌謝

能完成這篇論文，我要特別感謝我的指導教授傅立成老師，也要感謝實驗室的每一位成員對我的支持與鼓勵，謝謝大家。

1. 中文摘要

把台大碩士論文格式規定設定好後，做成一篇範本，方便其他人可以直接修改來用。另外也順便把以前一些Word操作的小技巧寫在裡面，希望對大家有所幫助。

關鍵字：論文格式、Word、範例

1. ABSTRACT

This is an English abstract. This is an English abstract. This is an English abstract. This is an English abstract. This is an English abstract. This is an English abstract. This is an English abstract. This is an English abstract. This is an English abstract. This is an English abstract. This is an English abstract. This is an English abstract. This is an English abstract.

1. CONTENTS

口試委員會審定書 #

[誌謝 i](#_Toc60053436)

[中文摘要 ii](#_Toc60053437)

[ABSTRACT iii](#_Toc60053438)

[CONTENTS iv](#_Toc60053439)

[LIST OF FIGURES vi](#_Toc60053440)

[LIST OF TABLES viii](#_Toc60053441)

[Chapter 1 Introduction 1](#_Toc60053442)

[1.1 Motivation 1](#_Toc60053443)

[1.2 Lecture Review 1](#_Toc60053444)

[1.3 Contributions 1](#_Toc60053445)

[1.4 Thesis Organization 1](#_Toc60053446)

[Chapter 2 Preliminaries 2](#_Toc60053447)

[2.1 Deep Neural Networks 2](#_Toc60053448)

[2.1.1 Convolutional Neural Networks 2](#_Toc60053449)

[2.1.2 Graph Convolutional Networks 2](#_Toc60053450)

[2.2 Metric Learning 2](#_Toc60053451)

[2.2.1 *Classwise* Scenario 2](#_Toc60053452)

[2.2.2 *Pairwise* Scenario 5](#_Toc60053453)

[2.3 Transfer Learning 9](#_Toc60053454)

[Chapter 3 Methodology 10](#_Toc60053455)

[3.1 Framework Overview 10](#_Toc60053456)

[3.2 Cross-Modality Graph Convolutional Networks 11](#_Toc60053457)

[3.2.1 Cross-Modality Graph Construction 15](#_Toc60053458)

[3.2.2 Sampling Scheme for Sparse Graph 16](#_Toc60053459)

[3.2.3 GCN Embedding 17](#_Toc60053460)

[3.2.4 Bidirectional Way for Final Representation 17](#_Toc60053461)

[3.3 Objective Functions 18](#_Toc60053462)

[3.3.1 Density Loss for Emotion Recognition 18](#_Toc60053463)

[3.3.2 Distributed Loss for Depression Estimation 18](#_Toc60053464)

[3.4 Cross-Task Knowledge Passing 18](#_Toc60053465)

[3.5 Mental Disorder Detection 18](#_Toc60053466)

[3.5.1 Mood Disorder 18](#_Toc60053467)

[3.5.2 Bo Disorder 18](#_Toc60053468)

[Chapter 4 Experiments 19](#_Toc60053469)

[4.1 Configuration 19](#_Toc60053470)

[4.2 Training Details 19](#_Toc60053471)

[4.3 Datasets 19](#_Toc60053472)

[4.4 Ablation Study 19](#_Toc60053473)

[4.5 Results 19](#_Toc60053474)

[Chapter 5 Conclusion 20](#_Toc60053475)

[Chapter 6 Future Works 21](#_Toc60053476)

[REFERENCE 22](#_Toc60053477)

1. LIST OF FIGURES

**找不到圖表目錄。**

1. LIST OF TABLES

**找不到圖表目錄。**

# Introduction

In this chapter, we first describe the research motivation in Section 1.1, then the lecture review is discussed in Section 1.2. After that, we concisely introduce the proposed algorithm and its contributions in Section 1.3 and conclude with the organization of this thesis in Section 1.4.

## Motivation

Schizophrenia is a mental disorder that progressively changes the mental state of a person and can be characterized by apparent changes in perception, thoughts, mood, and behavior [1]. Patient may feel depressed, lack motivation, even not able to live a normal life.

## Lecture Review

## Contributions

## Thesis Organization

# Preliminaries

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

## Deep Neural Networks

### Convolutional Neural Networks

### Graph Convolutional Networks

## 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.

### *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, it can be expressed as:

|  |  |
| --- | --- |
|  | ( 2.1 ) |

Where and are the batch size and the total number of classes, respectively. indicates a batch of embedding features and denotes the sample belonging to the class. denotes the classification weight matrix, which is the learned center of each class, and 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 2‑1 (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 [2]. Center loss exploits additional embedding centers and adopts Euclidean distance as a metric function to congregate intra-class features. Formally, it can be expressed as:

|  |  |
| --- | --- |
|  | ( 2.2 ) |

Where indicates an additional center embedding of each class

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |
| Figure 2‑1: Visualized results of features training on MNIST [3]. (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 feature distribution, the memory consumptions and computational costs have to be concerned. Because the critical term is calculated from center embeddings, it requires more effort 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 loss value range is unbounded, easily resulting in futile optimization.

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) [4] is proposed to map the features and the corresponding weights into the angular space. CosFace [5] and ArcFace [6] 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:

|  |  |
| --- | --- |
|  | ( 2.3 ) |

Where the margin penalties of SphereFace [4], ArcFace [6] and CosFace [5] are respectively denoted as , and . For other notations, denotes the batch size and is the number of possible label classes. is the angle between the feature vector and its target weight vector and is the angle between the feature and other weight vectors.

Since the above sphere mapping techniques mainly focus on designing different penalties for intra-class perspective to realize the objective, inter-class separability is neglected. RegularFace [7] 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 can be expressed as:

|  |  |
| --- | --- |
|  | ( 2.4 ) |

Where and denote the and classification weight, respectively. is the number of possible label classes.

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

### *Pairwise* Scenario

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. Thus, we cannot employ a classification weights matrix to promote feature discrimination globally. One of the representative approaches is Triplet Loss [8, 9]. 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 2‑2.

|  |
| --- |
|  |
| Figure 2‑2: The basic idea of Triplet Loss. |

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

|  |  |
| --- | --- |
|  | ( 2.5 ) |

Where Γ is a set with many triplet pairs, , and respectively denotes the index of the anchor, positive and negative points. is a margin constraint. is the embedding function to encode the original data point, and respectively indicate the Euclidean distance between the anchor and positive point and the distance between the anchor and negative point. denotes the hinge function to ignore the negative 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. Formally, the batch hard Triplet loss can be expressed as:

|  |  |
| --- | --- |
|  | ( 2.5 ) |

Different from triplet loss, which pulls one positive point while pushes a negative one simultaneously, these approaches explore more samples with negative classes for interaction. *N*-pair Loss [10] aims to *recognize one positive sample from negative samples of classes*. It can be expressed as:

|  |  |
| --- | --- |
|  | ( 2.6 ) |

Where and are the *N*-pairs samples from *N* different classes. Here, and indicate the anchor and the positive sample respectively. denotes the negative sample.

Lifted Structure Loss [11] tends to *identify one positive sample from all corresponding negative samples*. It works by pulling a positive pair as close as possible and pushing all negative samples to a position farther than the margin .

|  |  |
| --- | --- |
|  | ( 2.7 ) |

Where and respectively represent the sets of positive pairs and negative pairs.

Instead of using a portion of informative samples is incorporated to capture the structure of the embedding space, Ranked List Loss [12] exploits all pairs to construct a comprehensive structure for metric learning. In fact, this objective function first mines both non-trivial positive and negative samples, then weights samples according to 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; thus, they propose a hyper-sphere constraint to preserve the intra-class similarity structure. Formally, it can be expressed as:

|  |  |
| --- | --- |
|  | ( 2.8 ) |

Where if , otherwise. denotes the weighting for positive and negative pairs.

With various mining and weighting schemes, MS Loss [13] 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. Therefore, 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:

|  |  |
| --- | --- |
|  | ( 2.9 ) |

Where and indicate the mined positive pairs and negative pairs according to given anchor . , , and are hyper-parameters as in Binomial Deviance Loss [14].

Departing from mining informative samples, Circle loss [15] observes that the learning manner of previous works is inflexible and easily converges to ambiguous results. To tackle these problems, they propose a self-paced weighting, which measures the disparity between optimal solution and the sample itself, to dynamically adjust the gradient of each sample. Consequently, it leads to flexible optimization and better performance. Besides, they also propose a unified perspective for two elemental learning paradigms, learning with classwise labels and pairwise labels. Circle Loss can be expressed as:

|  |  |
| --- | --- |
|  | ( 2.10 ) |

Where and are non-negative weighting factors; and are the similarity between the anchor and the negative sample and the similarity between the anchor and the positive sample. is a radius of the hypersphere.

However, the embedding density is less discussed and often ignored in previous studies. In the embedding space, the data distribution of each class may have various density so that it is hard to impose a region regularization to restrict the intra-class distribution. In [16], Li *et al.* establish a density regularizer by the k-means center of encoded data in the embedding space, then promote the feature discriminative based on the *P*-way classification formula. This regularization term can be expressed as:

|  |  |
| --- | --- |
|  | ( 2.11 ) |

Where is a density regularization term to measure the distribution of each class, and and denote a set of features with the same class and the corresponding class centroid. is a newly incorporated intermediate variable, which is going to represent as the target density of class .

## Transfer Learning

# Methodology

In this thesis, we propose a multi-task learning framework to realize a mental disorder detection for schizophrenia patients via both emotion recognition and depression estimation. The organization in this chapter is as follows: Section 3.1 briefly introduces the overview our learning framework. In Section 3.2, we first elaborate on the shortcomings of facial analysis for both tasks. We then present a Cross-Modality Graph Convolutional Networks to integrate the information from different modalities effectively: face and context. Importantly, our multi-task learning framework involves many datasets with different domains, and the nature of each task is different, one for classification and another for regression. Thus, In Section 3.3, we respectively design an objective function for each task to realize better learning. On the other hand, we notice that depression estimation is an extension of emotion recognition. In Section 3.4, we introduce an approach to effectively transfer the prior knowledge from the emotion model to the depression one. Finally, in Section 3.5, we illustrate the detection approaches for the mental disorder of schizophrenia patients. By the elaborated design, our algorithm accomplishes impressive performance in many public benchmarks.

## 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) *cross-task knowledge passing (CTKP)*. For the backbone network of each task, we here exploit two-stream architecture, including 2 CNNs with 5 layers, to encode high-level representations from different modalities. After that, we employ our CMGCN to integrate the mental signals from different modalities further resulting in a comprehensive representation for the following processing. As the nature of each task is completely different, we demonstrate density loss and distributed loss to realize a better model convergence. On the other hand, depression estimation is an extension of emotion recognition, we design a knowledge passing scheme, CTKP, to effectively passing the prior knowledge from the emotion model to the depression model. With our well-design multi-task learning framework, our approach can successfully inhibit other state-of-the-art algorithms with a clear margin.

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

## 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 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 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 effectively integrate features from different modalities, including the face and context.

|  |
| --- |
|  |
| 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 [17]. |

An intuitive idea [17] 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 each modality feature, then separately learns different weightings via many linear layers (MLP) to emphasize the importance of each modality. 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 area) for identification, while others are irrelevant emotional pixels, such as the background. On the other hand, once the face alignment fails to capture the target human face, this separate weighting 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.

|  |
| --- |
|  |
| Figure 3‑3: The existing separate representation [17]. Since the GAP considers all irrelevant pixels, it will limit the performance gain and result in the inefficient training procedure. Moreover, this separate weighting scheme mainly relies on the correct input. Once the wrong human face is given, it cannot reduce the effect of the wrong information. |

To tackle these problems, we propose a *cross-modality graph convolutional networks* (CMGCN) to effectively integrate features from different modalities. Particularly, we here exploit the graph viewpoint to model the correlation between different modalities so that we can learn a joint representation meticulously. Moreover, since too many irrelevant regions are included in the context image, we develop a sampling scheme to build a sparse graph to keep critical regions and significantly drop other irrelevant ones. Finally, we introduce GCN to integrate features from different modalities via the 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 correlation for different modalities. Then, we describe the key module, sampling scheme and GCN embedding, to integrate features from different modalities according to the constructed sparse graph. Finally, we conclude the final representation via a bidirectional way.

|  |
| --- |
|  |
| Figure 3‑4: The illustration of our CMGCN. Here, we demonstrate the process from face to context graph. In the implementation, we will carry out our CMGCN parallelly to model the correlation from face to context and from context to face. |

### Cross-Modality Graph Construction

To encode the correlation of different modalities, we consider using an affinity graph to describe the pairwise relations among all pixels, including inner-modality and outer-modality. Given a pair of face and context features maps, the size of their tensor is shown as ; we first reshape the feature maps into for convenient processing. To construct the cross-modality graph 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:

|  |  |
| --- | --- |
|  | ( 3.1 ) |

Where and denote the pixel from face and context feature maps respectively.

### Sampling Scheme for Sparse Graph

Observe that the cross-modality graph 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, there 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 provide discriminative emotional signals while others are background pixels. |

Given the constructed cross-modality graph ,

### GCN Embedding

After building the sparse cross-modality graph , we then introduce a general GCN [18] to construct an embedding to integrate features. Comparing with the typical graph convolution operation, the obtained graph 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:

|  |  |
| --- | --- |
| *,* | ( 3.2 ) |

Where is an adjacency matrix in describing the critical correlations of different modalities, and is a normalization diagonal degree matrix. indicate the activation function for non-linear mapping. and are the input feature map and the embedding of GCN.

### Bidirectional Way for Final Representation

To acquire a comprehensive representation, we here consider a bidirectional way to explore critical regions among multi-modalities. 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:

|  |  |
| --- | --- |
|  | ( 3.3 ) |

Where and are face and context features. and indicate the graph feature of each modality based on Equation ( 3.3 ). denotes the concatenate operation to merge two given features.

## Objective Functions

### Density Loss for Emotion Recognition

### Distributed Loss for Depression Estimation

## Cross-Task Knowledge Passing

## Mental Disorder Detection

### Mood Disorder

### Bipolar Disorder

# Experiments

## Configuration

## Training Details

## Datasets

## Ablation Study

## Results

# Conclusion

This is conclusion …

# Future Works

This is future works

1. REFERENCE

[1] M. J. Owen, A. Sawa, and P. B. Mortensen, "Schizophrenia," *The Lancet,* vol. 388, no. 10039, pp. 86-97, 2016, doi: 10.1016/S0140-6736(15)01121-6.

[2] 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// 2016: Springer International Publishing, pp. 499-515.

[3] 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, doi: 10.1109/5.726791.

[4] 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, doi: 10.1109/CVPR.2017.713.

[5] 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, doi: 10.1109/CVPR.2018.00552.

[6] J. Deng and S. Zafeririou, "ArcFace for Disguised Face Recognition," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 27-28 Oct. 2019 2019, pp. 485-493, doi: 10.1109/ICCVW.2019.00061.

[7] 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, doi: 10.1109/CVPR.2019.00123.

[8] 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, doi: 10.1109/ICCV.2017.309.

[9] 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.

[10] 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.

[11] 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, doi: 10.1109/CVPR.2016.434.

[12] 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.

[13] 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.

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

[15] 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.

[16] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Deep Metric Learning With Density Adaptivity," *IEEE Transactions on Multimedia,* vol. 22, no. 5, pp. 1285-1297, 2020, doi: 10.1109/TMM.2019.2939711.

[17] 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.

[18] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," 2017. [Online]. Available: <https://openreview.net/forum?id=SJU4ayYgl>.