SEMATIC UNDERSTANDING OF LARGE – SCALE OURDOOR WEB IMAGES: FROM EMOTION RECOGNITION TO SCENE CLASSIFICATION

By

YANYAO LI

Bachelor of Science in Petroleum Engineering

The University of Tulsa

Tulsa, Oklahoma

2016

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
December, 2020

SEMATIC UNDERSTANDING OF LARGE – SCALE OURDOOR WEB IMAGES: FROM EMOTION RECOGNITION TO SCENE CLASSIFICATION

Thesis Approved:

Dr. Guoliang Fan

Thesis Adviser

Dr. Bo Zhang

Dr. Gary G. Yen

ACKNOWLEDGEMENTS

I would like to extend my grateful and sincere appreciation to my committee members: Drs. Guoliang Fan, Bo Zhang, Gary G. Yen who have helped in this endeavor and their willing to mentor. Without their patience, active guidance, cooperation and encouragement, I would not be able to achieve this far.

I am ineffably indebted to my advisor Dr. Guoliang Fan for his conscientious guidance, valuable support, and useful and helpful assistance. His thoughtful instructions and discussions provide me sufficient motivation and support to complete the goal.

I extend my gratitude to my colleagues, Xiaowei Chen, Le Zhou, for giving me help in the research and writing of the thesis.

At last but not least, I am extremely thankful and pay my gratitude to my parents for their understandings and support me morally as well as economically.

Any omission in this brief acknowledgement does not mean lack of gratitude.

Name: YANYAO LI

Date of Degree: DECEMBER, 2020

Title of Study: SEMATIC UNDERSTANDING OF LARGE – SCALE OURDOOR WEB IMAGES: FROM EMOTION RECOGNITION TO SCENE CLASSIFICATION

Major Field: ELECTRICAL ENGINEERING

Abstract: Facial expression recognition and scene-based image clustering are very popular topics in the fields of human-computer interaction and computer vision. Their relationship has been rarely investigated but is a very attractive topic that has many potential applications, such as landscape design, instructions for vacation choices, or plant layout design in the public space. In this research, we use the existing deep learning algorithms to study two issues, i.e., facial expression recognition and scene-based image clustering for large scale outdoor web images. This research paves a path for a future attempt that explores their relationship in real-world images. First, we concentrate on emotion recognition and investigate the performance of the well-known algorithms including Visual Geometry Group Network (VGG network) and Residual Net (ResNet) on the emotions in images captured from a public park. Then we introduce some approaches to address the challenges of the occluded or children's faces. Our proposed pre-processing schemes not only allow the algorithm to detect more faces but also to increase the rate of recognition accuracy under the complex environment. We also investigate the visual analysis of landscape by introducing a set of scene labels for a large set of natural scene images collected from an online source. Then the weakly supervised method – Curriculum Net is applied for scene labeling of our dataset. In Curriculum Net, the training dataset is split into two parts, clean (easy) and noisy (hard) datasets by using a Density Peak Clustering algorithm, from which Curriculum Net is trained from easy to hard data. Particularly, we adopt a more effective density clustering method, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), to improve the clean-noisy separation of training images that leads to the improved scene labeling performance. By summarizing the work in emotion recognition and scene-based image clustering, we prepare e future research to reveal the relationship between the two aspects in real-world scenarios.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER I

INTRODUCTION

The relationship between facial expression and background environment in an image is a very attractive topic that has many potential application scenarios. In the field of landscape design, landscape environment is arranged based on the requirements of human emotion activity, usage of space and scene characteristic to study how to create pleasant environment for people. Public behavior of the expression is the modern content of landscape planning and design with the population growth, multi-cultural communication in the modern information society, and the development of the social science. The landscape design focus on how to create a beautiful environment that makes people happy, imaginative and positive, mainly from the psychological and spiritual needs of the human beings, according to the human behavior in the environment and the law of spiritual feelings. In another way, a sorted list ranked according to people's level of happiness in scenic spots will provide support for people's decision-making on vacation [1]. This study uses millions of social media photos from scenes around the world. The results find that that the great wall of China, Stonehenge, and some parks such as Disneyland got a high score of happiness. Tourists in these places are fulfilled with smiling faces and are happier. Some religious places like temple or monastery are at the bottom of the list indicating more appearance of neutral feelings influenced by complicated factors such as culture and atmosphere. In addition, the environment of the space has an impact on people's emotions. The broad and wide spaces, such as

natural landscape, water area, plants, suburbs and villages, are positively correlated with people's emotions, while the people more likely express sad or angry in the indoor closed space, located in the city or apartment. Also the habit and thinking of emotion expressions are associated with long term changing of people's social abilities, interpersonal adjustment and working environment movement. Emotion recognition accuracy could be more accurate by revealing the association between habitual tendencies to its culture differences. All of them cannot be achieved without the assistance of the understanding of the primary features of the background. Therefore, interest in emotion recognition only no longer meets the development of the science [1].

Due to the high success of various deep learning neural networks, many fields are using it to explore the ability, such as emotion recognition and image clustering. The human-computer interaction of a computer system is a process of information transfer and transformation between human and computer, which is mediated by multimedia information such as voice, graphics and images. The task of human-computer interaction technology is to choose the appropriate mapping method to solve the complexity of its technology implementation. The complexity of technology implementation is mainly reflected in how to enhance computer perception and identify user interaction behavior and state, and then understand their interaction intention. The ease of use of user interaction is manifested in how to make full use of and coordinate multiple human feeling and effect channels, and then reduce the "cognitive effort" required to complete the transaction interaction. The main application fields of facial expression recognition technology include human-computer interaction, intelligent control, security, medical treatment, communication, and so on [2]. Clustering can be used to identify, divide image data sets, and organize navigation. At the same time, we will also use similar images for visualization [3].

However, during the research, we find that the current technology does not have substantial results regarding the detection and recognition in the real-world case scenarios. The training, validation, and test set in most research activities were performed under laboratory conditions,

which may not meet the requirements when the images have complicated elements such as occlusion or sunlight. The images in the standard dataset are mostly captured in bright and single environment, where the face is clear to be marked and analyzed. So, it is difficult to predict which face part or the face key points are occluded. The algorithm should rely more on the non-occluded part of the face than on the occluded part. Occluded regions are usually locally consistent (for example, every other point cannot be occluded), but it is difficult to use this feature as a constraint for occlusion prediction and milestone detection [4]. Secondly, the real world image, especially for the ones in the park can contain many features that are irrelevant like sunglasses. These features make image recognition more difficult. In summary, in this report, with the support of visual analysis of landscape, we want to improve he performance of some current well-known deep learning networks under a real-world scenarios including parks and wild grassland. As a result, in Chapter 2, we present some related work and introduce our work into emotion recognition and weakly supervised image clustering, as well as our new design of landscape labels during the study. Chapter 3 describes our method that is applied to facial emotion recognition in the parks with multiple people and many interference factors. We will also discuss how we address the challenges of occlusion and non-uniform sunlight. Our objective is to improve the rate of recognition. In Chapter 4, the idea of the design of the landscape is introduced. Chapter 5 discusses how we apply our new labels into the Curriculum Net and use a weakly supervised image clustering algorithm to help us classify images according the landscape type in the wild. Then the summary and conclusion will be provided in Chapter 6.

CHAPTER II

REVIEW OF LITERATURE

Our research is based on some existing work that will be discussed in the following parts. In particular, first, we present facial expression recognition. Some popular convolutional neural networks are introduced. Then we focus on image clustering. There will be a comparison between weakly supervised and unsupervised clustering methods which leads us to choose a weakly supervised approach to start our investigation on scene-based image clustering.

2.1 Emotion Recognition

Emotion recognition or Facial expression recognition is the most effective approach to detect emotions. It has many human-computer interaction applications, such as fatigue driving detection and real-time facial expression recognition on a mobile phone. As early as the 20th century, Ekman and other experts proposed seven basic expressions through cross-cultural research, namely anger, fear, disgust, happy, sad, surprise, and neutral [5, 6]. However, continuous research has found that these seven basic expressions cannot fully cover the emotions expressed in people's daily life. To solve this problem, an article in 2014 proposed the concept of consistent expression and pointed out that multiple discrete basic expressions can be combined to form a compound expression. For example, when people encounter unexpected surprises, they should be both happy and surprised [7].

With the continuous improvement of face processing technology (including face detection and face recognition), it is possible to analyze facial expression by computer. Generally speaking, facial expression analysis is very difficult to research, mainly reflected in the accuracy and effectiveness of the expression feature extraction. In particular, the latter is even harder, because the various expressions reflected in the movement of each feature point is similar, for example: opening the mouth does not mean laughing, it may also be crying and surprise [8].

The whole research of facial expression recognition is developing with the development of face recognition. Better methods in the field of face recognition will also be applied to face recognition. This report investigates the advances in facial expression recognition from algorithms. On the database side, emoticon recognition has gradually moved from a small sample size database under the unified control of the traditional laboratory to a large and diverse database in real life. In terms of algorithm, traditional manual design features, even shallow learning features, are no longer well suited to the real world of expression-independent interference

factors, such as light transformation, different head postures, and facial obstruction. As a result, more and more studies begin to apply in-depth learning technology to facial expression recognition to solve these problems [7, 9].

At present, there are three main recognition features used: grayscale feature, motion feature, and frequency feature. Grayscale features are processed from the grayscale values of the expression images. Different expressions have different gray values to get the basis for recognition. In this case, it is required that the image be fully preprocessed for illumination, angle, and other factors so that the gray values obtained are normalized. Motion features utilize the motion information of the main facial expression points under different facial situations for recognition. Frequency domain characteristics mainly make use of the differences of emoticon images under different frequency decomposition. Fast speed is its prominent feature [8, 10, 11]. There are three main directions in the specific expression recognition methods: global and local recognition, deformation and motion extraction, geometric and facial features. In the global recognition method, facial expressions are analyzed as a whole to find out the image differences under different expressions, whether from the distortions of the face or the movement of the face.

Machine learning is the general term of a class of algorithms. These algorithms attempt to mine the hidden rules from a large number of historical data and use them for prediction or classification. More specifically, machine learning can be seen as a function. The input is sample data, and the output is the expected result. However, the function is too complex to express formally. It should be noted that the goal of machine learning is to make the learned functions fit the "new samples", not just the training samples. The ability of functions learned to apply to the new samples is called generalization [12].

According to the learning theory, the machine learning model can be divided into supervised learning [13], semi-supervised learning [14], unsupervised learning, transfer learning [15], and

reinforcement learning [16]. When the training samples are labeled, it is supervised learning; when some of the training samples are labeled, it is semi-supervised learning; when all the training samples are unlabeled, it is unsupervised learning. Transfer learning is to transfer the trained model parameters to the new model to help the new model training. Reinforcement learning is a learning optimal policy, which enables the ontology (agent) to take actions according to the current state in a specific environment, so as to obtain the maximum return. The biggest difference between reinforcement learning and supervised learning is that each decision has no right or wrong, but hopes to get the most cumulative rewards.

Furthermore, based on the mission task and method, the Machine learning model can be divided into the regression model [17], the classification model [18], and the structured learning model [19]. From the perspective of methodology, it can be divided into a linear model and a nonlinear model.

The idea of learning is very rich, but in practice, people find it difficult to learn data representation directly from the original form of data. Deep learning is the most successful representation learning method [20, 21]. Deep learning is to divide the task of representation learning into several small goals. First, learn the lower level representation from the original form of data, and then learn the more advanced representation from the previous low-level representation. In this way, each small goal is relatively easy to achieve, so we can complete the task of expressing learning. This is similar to divide and conquer in the idea of algorithm design. In summary, we gradually extract high-level features from low-level features by using the characteristics of processing features layer by layer in the network.

The emotion detection based on deep learning [22] consists of three main processes [11, 23]: 1. Data pre-processing; 2. Feature extraction; 3. the classification. The so-called preprocessing is to eliminate all interference that has nothing to do with the face before calculating the features.

Therefore, there are processes such as face detection, face alignment, normalization, etc. The main processes are face detection, face alignment, data augmentation, face normalization [24].

The most important here is the data augmentation [25]. Mainstream offline data enhancements include random perturbations, transformations (rotation, translation, flip, scaling, alignment), noise additions such as salt and pepper noise, speckle noise, and variations in brightness, saturation, and noise with a 2-dimensional Gaussian random distribution between eyes [26, 27]. At the same time, there are others such as GAN to generate faces and 3D CNN to assist AUs to generate expressions. However, it has not been proved whether GAN-generated faces can improve the performance of network models [28, 29].

Online data enhancement, including Crop, Horizontal Flip, etc. The main meaning is that when predicting, test data can be Crop, Flip, and other operations at one time to generate several similar test maps, and then the output predicted by each test map is averaged, which is mainly based on the model of random perturbation training, and the reason why the mean value needs to be calculated during testing [30, 31].

Face normalization mainly refers to brightness normalization and posture normalization (that is, face alignment correction). Brightness normalization includes not only brightness adjustment, but also contrast adjustment, common ways of contrast adjustment. There are histogram normalization, DCT normalization, DOG normalization, related papers have proved that the effect of histogram normalization is the most stable, suitable for various network models. There are also papers proving that global contrast normalization, local normalization, histogram normalization, global contrast normalization, and histogram normalization are the best of the three methods. Therefore, it is recommended that histogram normalization be combined with brightness normalization. Posture normalization has the greatest impact on face items. At present, most of them are still in small angle, 2D landmark alignment, and the more reliable direction is a

3D landmark, which is estimated from image and camera parameters, or measured and calculated from the depth sensor. Newer estimation models are FF-GAN, TP-GAN, and DR-GAN [32-34].

Facial expression classification is mainly to predict probability (layer of softmax) directly based on the features learned in a deep network. Or the deep learning characteristics can be classified using shallow classifiers such as SVM. Furthermore, most studies now focus on the effects of light, posture, and occlusion on expression recognition. The authors in this paper focused on the effects of individual differences such as age, gender, ethnic background on facial recognition. This paper studies expression recognition from the perspective of "De-expression". Through some facts and literature, the author found that human expressions can be divided into two parts: Neutral Component and Expressive Component. The author's idea was to get a neutral expression corresponding to a face through a GAN network, and then trained and learned residue to further classify the expression [35].

2.2 Image Clustering

Data clustering is a basic problem in many filed such as machine learning, computer vision, and data compression. The goal of data clustering is to categorize similar data into one cluster based on some similarity measures. Although a large number of data clustering methods have been proposed, conventional clustering methods usually have poor performance on high-dimensional data, due to the inefficiency of similarity measures used in these methods.

Traditional clustering methods contain a Partition-based method, a density-based method, and a hierarchical method. Disadvantages of traditional clustering can be listed as the similarity measurement method used is inefficient, traditional clustering methods have poor performance on high-dimensional data, and have high computational complexity on large-scale datasets. People figure out the solutions that dimension reduction and feature transformation map the original data into a new feature space, making the generated data easier to separate from the existing

classifiers. Data conversion methods include linear transformations (such as principal component analysis (PCA)) and nonlinear transformations (such as the kernel and spectral methods) [3].

Automatic Encoder (AE) is one of the important algorithms in unsupervised representation learning. Train the mapping to ensure that the reconstruction loss between the encoder layer and the data layer is minimal. Since the dimensions of the hidden layer are usually smaller than those of the data layer, it can help extract the most significant features of the data.AE is mainly used to find better initialization for parameters in supervised learning, and it can also be combined with unsupervised clustering.AE can be thought of as consisting of two parts: mapping the original data X to an encoder representing h, and a decoder that produces the reconstructions. Where phi and theta represent the parameters of the encoder and decoder, respectively. The reconstructed representation r must be as similar as X [36-39].

The CDNN based algorithm only uses a cluster loss training network [40], in which the network can be FCN, CNN, or DBN [41, 42]. Since there is no rebuilding loss, when data points are simply mapped to a tight cluster, CDNs-based algorithms run the risk of acquiring corrupted feature space, resulting in small but meaningless values for cluster loss. Therefore, setting up the loss of clustering and network initialization is important. Therefore, according to the way of network initialization, deep clustering algorithms based on CDNs are divided into three categories: unsupervised pre-training, supervised pre-training, and random initialization (no pre-training).

Moreover, VAE can be considered as a variant of AE, which combines the variation Bayesian method with the flexibility and scalability of neural networks. It introduces a neural network to accommodate a conditional posterior, so the goal of variation inference can be optimized through random gradient descent and standard backpropagation. It uses the parameterization of the lower bound of the variation to produce a simple, differentiable, unbiased estimator of the lower bound.

This estimator can be used for valid approximate posterior inferences in almost any model with continuous potential variables. Mathematically, it is designed to minimize the lower (changing) bounds of the marginal likelihood of dataset X [43, 44].

GAN based deep clustering is another kind of deep generation model popular in recent years. GAN establishes a minimum-maximum game between two neural networks (Generating Network G and Discriminating Network D). Generating networks attempt to map sample z from the previous distribution p(z) to the data space, while discriminant networks attempt to calculate the probability that the input is a real sample based on the data distribution, rather than generating the network-generated samples. SGD optimizer G and discriminator D can be used alternately. GAN provides a confrontational solution that matches the distribution or representation of data to any prior distribution. GAN-based deep clustering algorithms have GAN-like problems, such as difficult convergence and pattern crashes [45].

CHAPTER III

EMOTION RECOGNITION IN THE WILD

In this chapter, we will talk about emotion recognition in the wild based on the machine learning techniques including the methods employed, as well as the several improvements we made to enhance the quality of results. Machine learning, or more specifically, Convolutional Neutral Network (CNN) is a basic approach to detect facial expressions in an image. The artificial features of the human face include 68 commonly used facial landmarks and other features. In addition to prediction, deep learning often plays the role of Feature Engineering, thus eliminating the steps of manual feature extraction. What we do here is to extract features from images using CNN, thus avoiding the tedious manual feature extraction. In this chapter, there are four sections: (1). The type CNN applied in this chapter; (2). The baseline results to show the performance of CNN in the realistic images. (3). Procedures for improvements. (4). Better results are achieved.

3.1 Convolutional Neural Networks

3.1.1 VGG Network

VGG was proposed by the Visual Geometry Group of Oxford. The main work is to prove that increasing the depth of the network can affect the final performance of the network to a certain extent. VGG net explored the relationship between the depth of the convolutional neural network and its performance. By repeatedly stacking 3 * 3 small convolution cores and 2 * 2 maximum pooling layer, VGG net successfully constructed a convolutional neural network with 16 - 19 layers. Compared with the previous state of the art network structure, the VGG net has significantly reduced the error rate [24, 46].

At the same time, the VGG net has a strong generalization ability, and the data generalization of migrating to other images is very outstanding. The structure of the VGG net is very simple. The convolution kernel size (3 * 3) and maximum pooling size 2 * 2 of the same size are used in the whole network. So far, the VGG net is still used to extract image features. The model parameters after VGG net training are open-source on its official website, which can be used for retraining on specific image classification tasks (equivalent to providing a very good initialization weight), so they are used in many places. There are two structures of the VGG network, VGG-16, and VGG-19. There is no essential difference between them, but the depth of the network is different [24].

One improvement of VGG-16 compared with Alexnet is that several 3x3 convolution kernels are used in VGG-16 to replace the larger convolution kernels (11x11, 7x7, 5x5) in Alexanet. For a given receptive field (the local size of the input image related to the output), the small stacking convolution kernel is better than the large convolution kernel, because the multi-layer nonlinear layers can increase the network depth to ensure the learning is able to handle more complex patterns, and the cost is relatively small with fewer parameters [46].

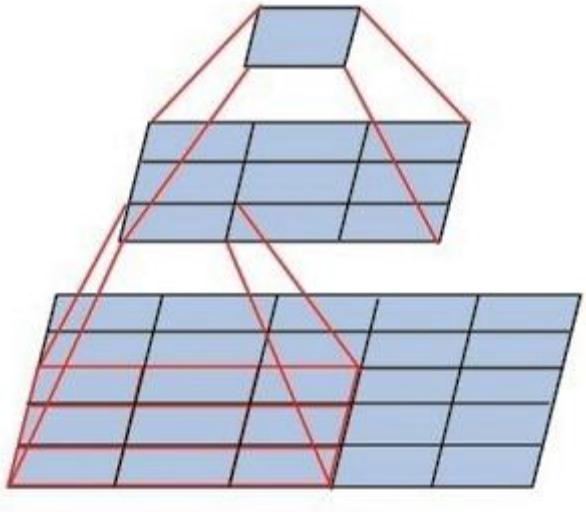

Figure 1. Mini – network replaces the $5 * 5$ convolution with two $3 * 3$ convolutions to reduce the parameters from $25 * C^2$ to $9* C^2$. Same idea is applied to replace the $7 * 7$ convolution with three 3x3 convolutions [46].

In general, in VGG, three 3x3 convolution kernels are applied to replace one 7x7 convolution kernels, and two 3x3 convolution kernels are used to replace one 5x5 convolution kernels. The main purpose is to enhance the depth of the network and improve the effectiveness of a neural network to a certain extent of ensuring the same perceptual field.

Therefore, the 5x5 convolution is regarded as a small fully connected network, which is sliding in the 5x5 area. We can first use a 3x3 convolution filter for convolution, and then use a fully connected layer to connect the 3x3 convolution output. The fully connected layer can also be treated as a 3x3 convolution layer. In this way, we can replace a 5x5 convolution by concatenating two 3x3 convolutions.

Furthermore, the VGG net employed the multi-scale method for data augmentation. The original image is scaled to a different size S, and then randomly cut to 224 x 224. This procedure increases the amount of data and has an optimized impact on preventing the model from

overfitting. In practice, the author took values of S in the range of 256 - 512, and uses the way of multi-scale to obtain data of multiple versions, and combines all of the data for training. Besides, after the trial processes of LRN, VGG authors thought that LRN has little effect on the model and leads to an increase in memory consumption and computing time.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					

Figure 2. The development of the structure of the VGG Network [24].

Although the VGG network is deeper than the Alexnet, the VGG network converges faster than Alexnet in the training process, mainly because:

(1) Regularize the data using a small convolution kernel and deeper network;

(2) The pre-training data is used to initialize the parameters in the specific layer. For shallow networks, such as network A, random numbers can be directly used for random initialization, while for deeper networks, the convolution layer and the last full connection layer can be initialized by using the parameter values of shallow networks that have been trained before.

3.1.2 Residual Network

The emergence of Alexnet brings in deep learning. Its most important point is that the model can learn features automatically through a data-driven system, which saves the steps to search for features manually. However, different models also find features in different quality, the quality of features directly affects the accuracy of classification results, and the features with stronger expression ability also bring stronger classification ability to the model. Therefore, deep networks will learn more features through data [47, 48].

Features can also be roughly divided into three categories, according to the complexity and representation ability. Theoretically speaking, the more complex features have a stronger representation ability. In the deep network, each feature will be continuously calculated through the linear/nonlinear comprehensive model Thus, the deeper the network output is, and the stronger the feature is. Therefore, the depth of the network is very important for learning more complicated features. This behavior is well reflected in the VGG net [47, 49]. In the depth model, the size of the output of each layer changes with the depth of the network, mainly because the height and width become smaller and smaller. The depth of the feature map increases with the depth of the network layers. This design conforms to the concept of Inception v3. On the other hand, the reduction of height and width helps to reduce the amount of calculation, while the increase of the depth of the feature map increases the number of available features in each layer output.

Deep convolutional neural network has made a breakthrough in image classification. Recent evidence shows that the depth of the network is crucial. Leading teams use deep models, with layers of at least 16 or more, regularly between 16 and 30 in the Imagenet (a famous database). Moreover, many special visual recognition tasks also benefit greatly from the deep CNN model.

So can we get a better network by just adding layers? The answer is obviously no. The main reason is the gradient dispersion problem caused by backpropagation (BP). After the gradient backpropagation, the backpropagation gradient will disappear completely after several layers, which leads to some layers are not get trained due to that the gradient cannot reach the layers when the number of network layers increases. Surely, due to the occurrence of RELu and the central normalization layer, the depth of the network becomes larger is necessary, but the problem of gradient dispersion is still not solved fundamentally. The effect of simply stacking layers on a deep network is not as good as that of a shallow network with an appropriate number of layers, which is called network degradation [47, 48]. But, when the depth of the network continues to increase, the error rate becomes higher. Through experiments, the author found that the increase of the error rate is not due to the overfitting caused by the deep of the network, but by the increase of the low error limit of the network structure. As shown in the figure below:

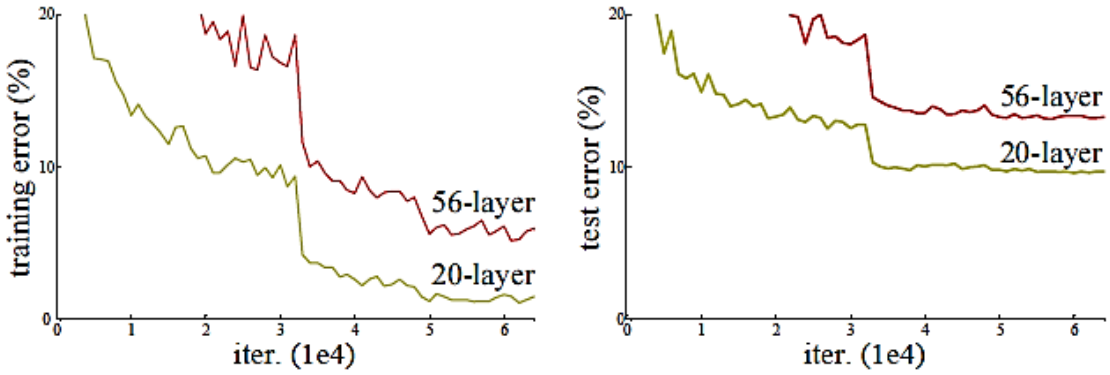

Figure 3. The rate of error increases as the layer stacks which indicated that simply adding layers will not help the training [47].

The author pointed out that if the adding layer can be constructed as identity mapping, the training error rate of a deeper model should not be larger than that of its corresponding shallower model. However, as we said before when we continue to increase the number of layers, the training error will become larger. Therefore, the author proposed to reconstruct the mapping of the network with the residual error. In other words, the input x is introduced into the result again. In this way, the weight of the stack layer will tend to become zero, which will be easier for the model to learn and more convenient to approach identity mapping. If x is mapped to f(x) + x through the network, then the network mapping f(x) naturally tends to f(x) = 0.

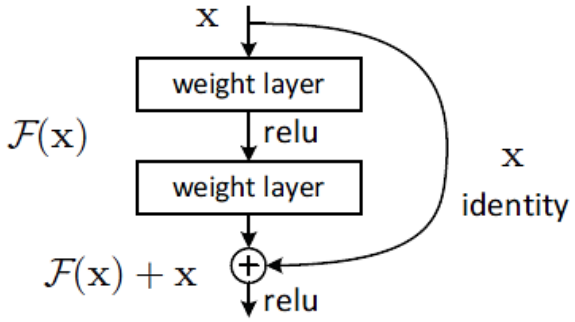

Figure 4. A building block for the residual learning [47].

The single building block of the residual network is not complicated. Suppose that in a single building module (which can be understood as several layers of the plain layer), the final result we need is H(x), because, in theory, a basic mapping matched by several stacked layers (not necessarily the whole network) can fit any function, then naturally it can also fit H(x) - x, and it may be more convenient and accurate to fit H(x) - X. So the author used the stack layer to fit H(x) - x, and then adds x again at the end to get H(x). Here we define H(x) - x as f(x). This module is shown in the following figure 5:

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

Figure 5. The structures of different Residual network with the increase of the layers from 18 to 152 layers [47].

In other words, without using residual networks, the structure of the plain is inspired by VGG. The convolution layer is mainly composed of convolution layers and 3x3 filters. Two simple design rules are followed:

(1) For the same output feature mapping size, the layer has the same number of filters;

(2) If the size of the feature map is halved, the number of filters is doubled to maintain the time complexity of each layer. We do downsampling directly by having a convolution layer with a span of 2. The network ends with a global average pool layer and the mildest 1000 channel full connection layer. The total number of weighted layers in Figure 3 (middle) is 34. It is worth noting that the model in this paper has fewer filters and lower complexity than the VGG network (Fig. 6, left). Our 34 tier baseline has 360 million triggers, which is only 0.18 of VGG-19 (1.96 billion triggers).

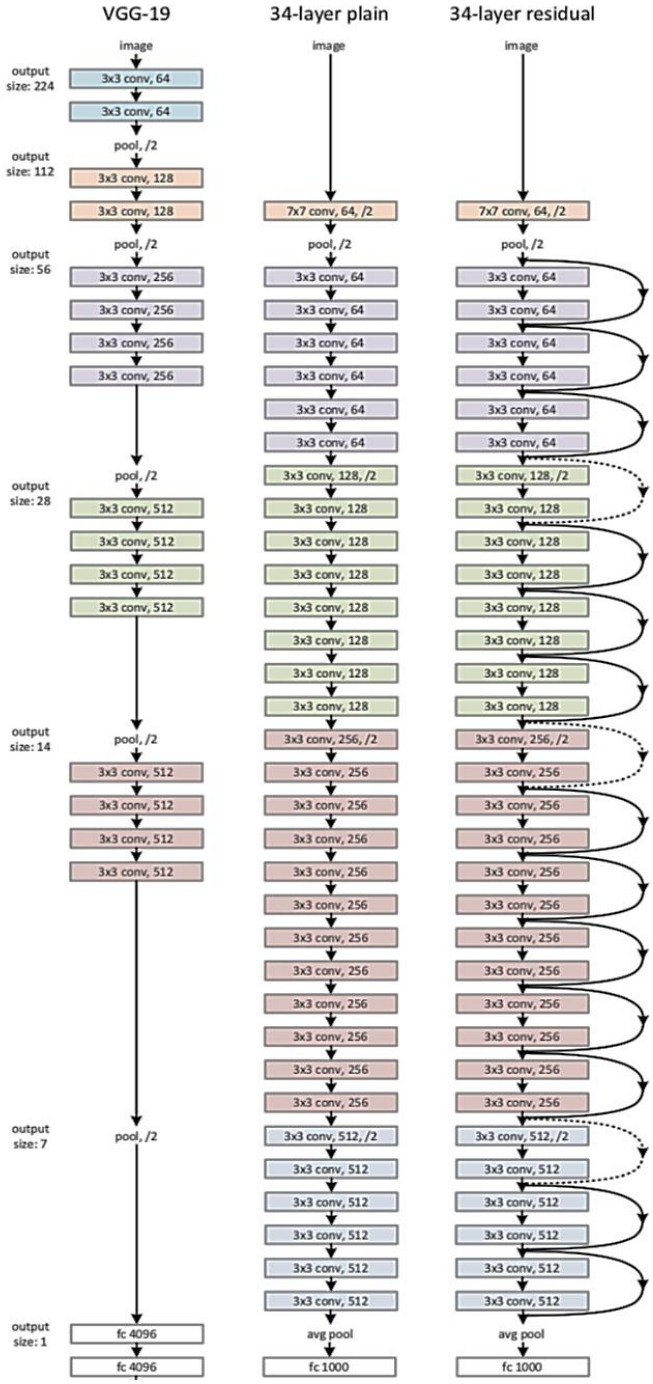

Figure 6. A 34 layers plain network is showed here as baseline to compare VGG network in 19 layers from the Residual network in 34 layers [47].

Therefore, for Residual network: based on the above plain network, the author inserted a quick link (Fig. 6, right), and transforms the network into its corresponding residual version. X is directly connected to the module output. When the input and output sizes are the same (the shortcut of the solid line is shown in Fig. 6), the shortcut connection can be used directly. As the size increases (the dashed shortcut is shown in Fig. 6), we consider two options [47, 49]:

(1). The shortcut still performs identity mapping, filling the entire vector with zeros until it is the same size as the input. This option does not introduce additional parameters;

(2). The projection shortcut in formula (2) is used to match the size (complete 1×1 convolution). And use the same span size as feature maps.

3.1.3 Real-Time Convolutional Neural Networks

At present, the most commonly used models end with a full connection layer, which usually contains most parameters of the CNN model. For example, VGG 16 contains 90 percent of the parameters in the final full connection layer. Google Inception V3 reduces model parameters by adding a global average pooling layer at the end. Xception combines two of the most successful modules: residual and depth separable convolution. The depth separable convolution can further reduce the parameters by separating the feature extraction process and combine them through a convolution layer. For instance, the best performing model on the FER2013 has 5 million parameters, and the final full connection layer accounts for 98 percent of the parameters [50].

In this particular network, two models are trained, and the best results are obtained by balancing the accuracy and the model parameters. Reducing model parameters help us overcome two problems: the slow running speed and generalization. The idea of the first model is to directly remove the full connection layer. In the second model, the full connection layer is removed, and then the convolution and residual modules can be separated by fusion depth. Both models are trained with Adam optimization.

In the first model, the authors used global pooling to replace the full connection layer. The basic model has nine convolution layers, including 600,000 parameters. It is trained on the IMDB (460,000 images) gender data set, and the accuracy rate is 96 percent. At the same time, the training on Fer2013 (35,000 images) with 7 different kinds of expression data sets has an accuracy rate of 66%.

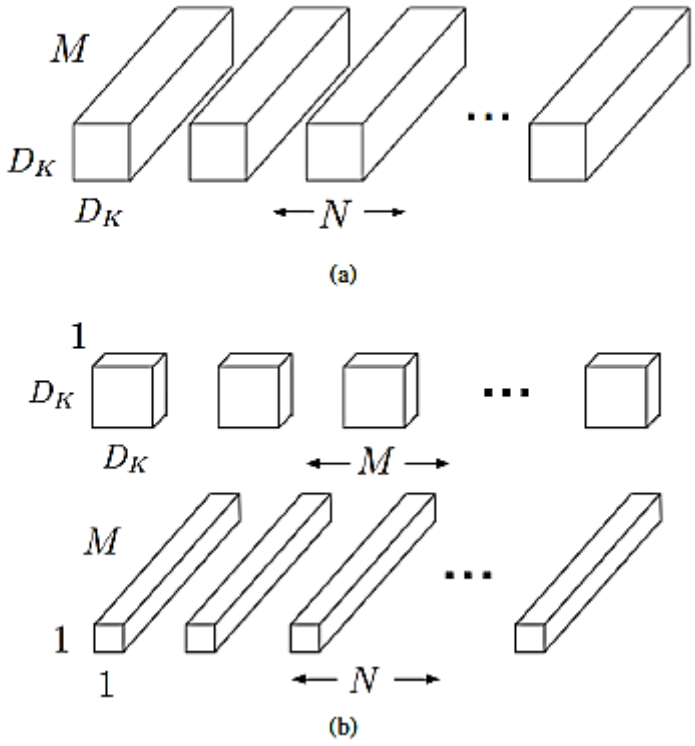

Figure 7. (a) Standard convolution layers. (b) Depth wise convolutions. The standard convolution layers are replaced by depth wise convolutions, which extremely efficient relative to standard convolutions [51].

The second model is inspired by Google Inception. This model uses the residual module and depth separable convolution. Depth separable convolution is composed of depth convolution and point convolution. The purpose is to separate spatial correlation from channel correlation. The final model has four residual blocks. Every convolutional layer has a BN and Relu. A global

average pooling layer and softmax are added after the last convolution layer. This structure has nearly 60,000 parameters, which is 10 times less than the basic model. The model is as follows, which we call Mini Xception.

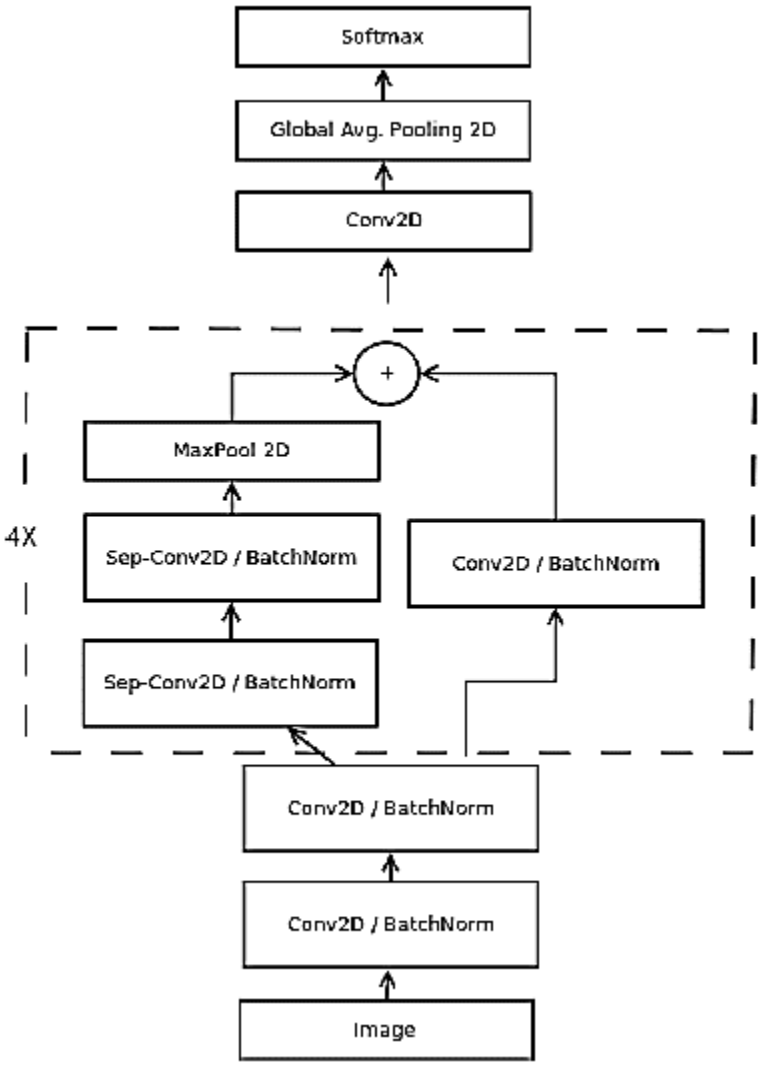

Figure 8. Mini-Xception network [50].

3.2 Baseline Results

In this report, we use a deep convolutional neural network to integrate facial expression feature extraction and expression classification into an end-to-end network. Three convolutional neural networks are discussed here and learned into our facial expression recognition, which are VGG

19, ResNet 18, and the mini Xception. To maintain the accuracy of the baseline result, the model is trained and tested on the famous public FER2013 (Facial Expression Recognition) dataset. The data set of fer2013 consists of 28,709 in the training set, 3,589 in the public test sets, and 3,589 in the private test sets. Each image is a grayscale image with 48 * 48 pixels.

There are seven expressions in Fer 2013 database: anger, disgust, fear, happy, sad, surprised, and neutral. The database comes from the 2013 Kaggle competition. Since most of the images in the database are downloaded from web crawlers, there are some errors. The artificial accuracy of this database is 65 $\pm$ 5 percent. Though the detection of gender is not required in this report, the gender test is still applied with the CK+ dataset. The CK+ database was released in 2010 and extended from the Cohn Kanade dataset. This database includes 123 subjects and 593 image sequences. The last frame of each image sequence has an action unit label. Among the 593 image sequences, 327 sequences have emotion labels. This database is obtained under laboratory conditions, which are rigorous and reliable. CK+ is a standard database in facial expression recognition, many articles use this data for testing.

As mentioned above in chapter 2, our images come from mainly in two locations and downloaded through a web crawler. In this chapter, the images from the park will be discussed in detail. In the park, the key objects are chairs, water or lake, pool, forest, some games, and most importantly, the people. The tested emotions include anger, disgust, fear, happy, sad, surprise, and neutral. The details of the dataset is shown in Table 1. We will use the CNN to test these images and our goal is to check how the convolutional neural networks trained on a regular public dataset performs on the wild large scale images in reality.

Table 1. Top: The details of the emotion dataset that "Happy" occupied the most of the dataset. Bottom: The baseline results is displayed for emotion recognition via different networks. Xception shows the best performance, while the different between each network is not significant.

Happy	5,078
Neutral	1,620
Sad	712
Fear	34
Angry	18
Surprised	61
Disgust	16
Happy + Neutral + Sad	7,410
Total	7,539

	VGG Network	ResNet	xception
Happy	0.79	0.81	0.85
Neutral	0.35	0.33	0.38
Sad	0.55	0.56	0.58
*Fear	0.22	0.21	0.22
*Angry	0.36	0.38	0.41
*Surprised	0.45	0.44	0.48
*Disgust	0.20	0.22	0.26
Average	**0.41**	**0.42**	**0.44**

The tested results are shown in Table 1 as well as the count of each category in the dataset. A test on different emotions is also listed. From the result, we can see that in general, the results have an

approximately 0.4, 0.42, and 0.45 rate accuracy in corresponding to the mini VGG Net, ResNet, and Xception. To be mentioned, since these images are from the park, the negative emotions (with a star) are hardly occurred in people's face, which indicates that these results are not confirmative enough. In Figure 6, we provide some results of our emotion classification through the networks.

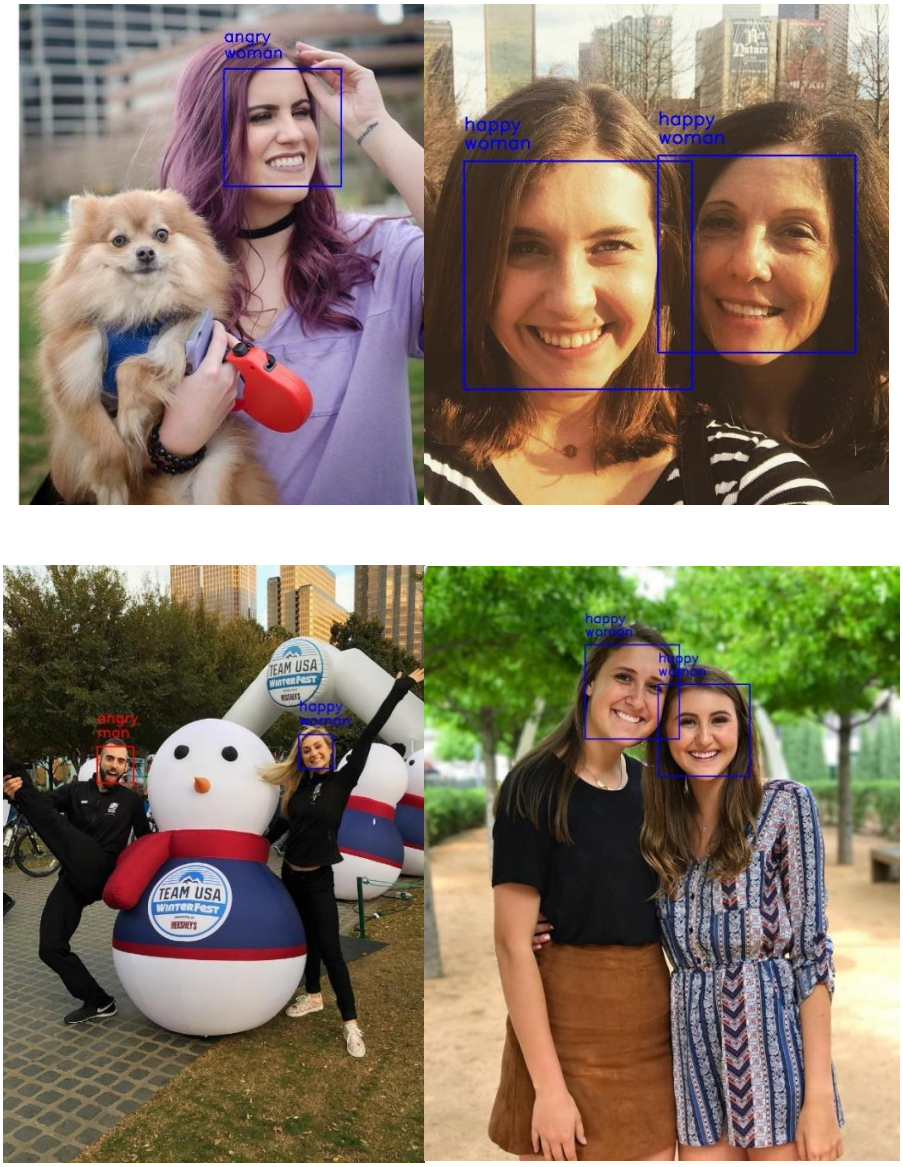

Figure 9. Examples of the some tests on the emotions to show “happy” and “angry”. If there are multiple faces in an image and their emotions are different, this image will be discarded.

To be mentioned, in the test, if there are multiple faces in an image and their emotions are different, this image will be discarded. The reason of this is because in the landscape design, there is just one label for each image no matter how many different emotions for the image. So we erased the images that have multiple emotions. Fortunately, these images are very less which will not have big image on our results. Then, we can observe that there are several common misclassifications such as predicts "angry" instead of "happy". The accuracy rate of happiness and surprise is significantly higher than that of others, but the accuracy rate of fear was very low. The explanation could be that there are 5,478 happy pictures, but only 16 disgusting pictures. This kind of imbalance is enough to increase the rate of the error. The other is that the expression of anger, disgust, fear, and sadness have certain similarities. In real life, people will find it difficult to distinguish these four kinds of expressions, especially when they do not know each other. Moreover, we find that misclassification always appears in the same classes. It may be that some classes are really difficult to distinguish and easy to be confused.

Furthermore, a huge amount of images that are not even detected. Like the following images, we found that certain types of images are not classified due to the face is not detected. The detection of the face is preparing work the emotion recognition, which plays a significant role in the procedures.

Figure 10. The rate of error occurs due to that some faces are not detected for children (mid) and the faces with sunglasses and glasses (left and right).

Considering the misclassification mentioned above, we propose the following steps to improve the accuracy of emotion detection:

1. Most of the convolutional neural networks for emotion recognition utilize a build-in package, such as the "Haarcascades" in OpenCV. The package reduces the time and space complexity, and ease the processing of code at the cost of a relatively low rate of accuracy. Hence, we built a filter to eliminate the images with a small face or the face cannot be detected due to sunlight, sunglasses, and side faces. Here, we use another face recognition program to complete our jobs. This program uses the latest face recognition method based on deep learning of C++ Dlib. The training and test are based on the outdoor face data test Library Labeled Faces in the Wild with an accuracy rate is 99.38%. As a result, we can pick out unique features of the face that we can use to tell it apart from other people— like how big the eyes are, how long the face is, and then find all the faces in it, even if a face is turned in a weird direction or bad lighting.

2. The FER2013 dataset provides seven different kinds of emotions, including anger, disgust, fear, happy, sad, surprised, and neutral. However, when we talk about the people in the park, the major emotions would just be happy, calm, or depress. The other emotions like fear, disgust, or anger rarely appear to the people who enjoy their time in the park. Therefore, we change the number of output classes in the output layer of the existing network and train our network based on the existing network weights. This idea is gained from the experience of transfer learning. In terms of the convolutional neural network, the weight of each node in a layer of the network is transferred from a trained network to a new network, instead of training a neural network for each specific task from the beginning. The benefits of this can be reflected in the following example. Suppose you already have a deep neural network that can distinguish cats and dogs with high accuracy. Later, you want to train a picture model that can distinguish different kinds of dogs. What you need to do is not to train the first few layers of neural networks that are used to distinguish straight lines and angles, but to use the trained network to extract them. After that, only the last few layers of neural networks are trained to distinguish dogs.

3. By observation, the children and the people with sunglasses are the two most features that can be recognized but not. According to this point, we add more images in the training set including the emotional expressions when people wear sunglasses and the facial expression of the children for the network to sufficiently learn the network.

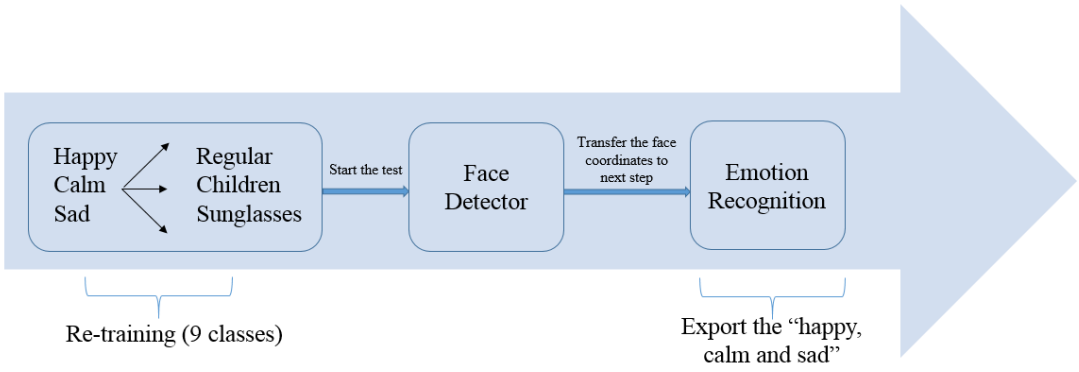

Figure 11. Flow chart of the preprocessing and test of the emotion recognition after the modification.

3.3 Experimental Results

In this section, we present our results of emotion recognition for the images in the park after the changes we made as listed below. In summary, we have more faces detected and the rate of the precision increase to 0.57%, 0.59%, 0.63% for VGG network, ResNet, and Xception.

Table 2. The network is able detect more faces after the modifications, while the dataset has 7,410 images

	Base line	Afterwards
Face detected	4,120	4,500

Table 3. The emotion detection after the processes of improvement.

	VGG Network	ResNet	xception
Happy	0.8	0.82	0.85
Neutral	0.39	0.38	0.41
Sad	0.54	0.58	0.62
Average	**0.57**	**0.59**	**0.62**

Table 4. Comparason of the emotion recognition results between baseline and our algorithms. Our method has a great improvement.

	VGG Network	ResNet	Xception
Baseline	**0.41**	**0.42**	**0.44**
After modification	**0.57**	**0.59**	**0.62**

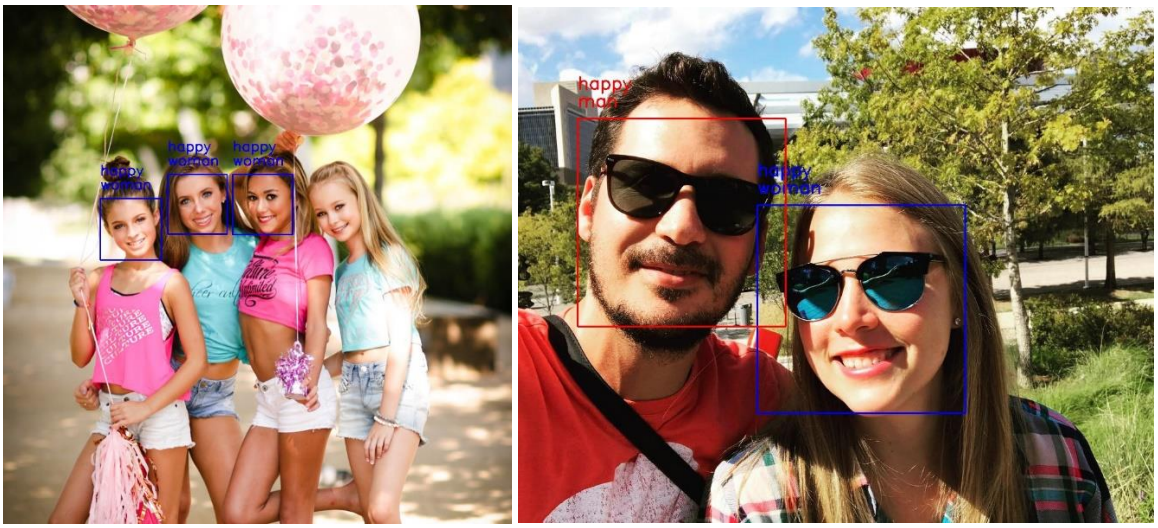

Figure 12. Left: Happy faces of children. Right: Happy faces with sunglasses. Our method can detect more faces compared with the baseline.

CHAPTER IV

LANDSCAPE CHARACTER CONSTRUCTION

This chapter will discuss the preparation we made for the image clustering. The outdoor images we used to contain some labels that have not been defined in the regular database. Moreover, we would like to build a connection between the works of art, landscape art, landscape perception, and photograph properties with deep learning techniques. In this way, the label of the images has a meaningful impact on our results. Based on the points from landscape color, complexity, and preference on viewing behavior, we create our self – designed label to meet the specific requirements. The following section includes 1. Landscape research; 2. Discussion of the new label.

4.1 Visual Analysis of Landscape

Landscape in Ecology in a narrow sense refers to the heterogeneous geographical unit with a repetitive pattern composed of different types of the ecosystem in the range of tens to hundreds of kilometers. The landscape complex which reflects the comprehensive characteristics of climate, geography, biology, economy, society, and culture is called "region". The narrow sense of landscape and region is the macro landscape that people usually refer to. Broadly speaking: it includes spatial units with heterogeneity or patch from micro to macro scales. The concept of generalized landscape emphasizes spatial heterogeneity, and the absolute spatial scale of landscape varies with the research object, method, and purpose.

The landscape is an orderly system of regular combination process of a series of landscape clusters, from small to large, as follows [52-55]:

1. Landscape elements: the smallest hierarchical component, including various landscape elements such as farmland, roads, and villages.

2. Landscape element group: also known as landscape element chain, is composed of a series of landscape elements closely related to the composition structure, which can perform certain landscape functions, such as succession chain, interference chain, and function chain.

3. Landscape unit: also known as landscape functional area, is a relatively independent and complete unit composed of a variety of functionally related landscape elements, such as villages, forest farms, and market towns.

4. Landscape cluster: it is the organic combination of several similar landscape functional areas, including forest landscape cluster, agricultural landscape cluster, suburban landscape cluster, and urban landscape cluster.

5. Landscape system: the regular combination of several landscape clusters, such as the landscape system in New York

Based on the basic characteristics of spatial morphology, contour, and distribution, five spatial types can be distinguished: patch, corridor, matrix, net, and edge.

1. Patch refers to the nonlinear landscape ecosystem unit that is different from the surrounding background;

2. Corridor refers to the spatial type of landscape ecosystem with a line or belt shape;

3. The matrix is the largest and most widely distributed landscape ecosystem in a certain area, and its high-quality type is very prominent;

4. Network refers to a kind of structure that connects different ecosystems in the landscape.

5. Edge, also known as the transition zone, fragile zone, or marginal zone, etc., refers to the part of the landscape ecosystem with significant transitional characteristics.

In our study, one of the major challenges in analyzing landscape clustering is the decision of the correct labels. For many clustering problems, including the clustering for the fruit, biodiversity, vehicles, or specific buildings, we have a strong conceptual base to guide the indicators of the labels. However, in our case, this conceptual base is weak, which is urgently required if we are to be able to compare the difference between images and apply image clustering [55].

The contents of the landscape vary greatly according to different starting points. For large-scale regulation, the planning is mostly from the perspective of geography and ecology; for medium-sized theme park design, the landscape design is often from the perspective of planning and garden; the small areas like green space in the resident or city square are built from detailed planning and location of architecture. Generally speaking, the consideration of landscape factors

in the process of planning and design is usually divided into the hard landscape and soft landscape. As far as the scientist concerned, a hard landscape refers to artificial facilities, usually including pavement, sculpture, awning, seats, lights, fruit boxes, etc.; soft landscape refers to more natural features including vegetation, rivers, and other simulated natural landscapes, such as fountains, pools, or forest [56].

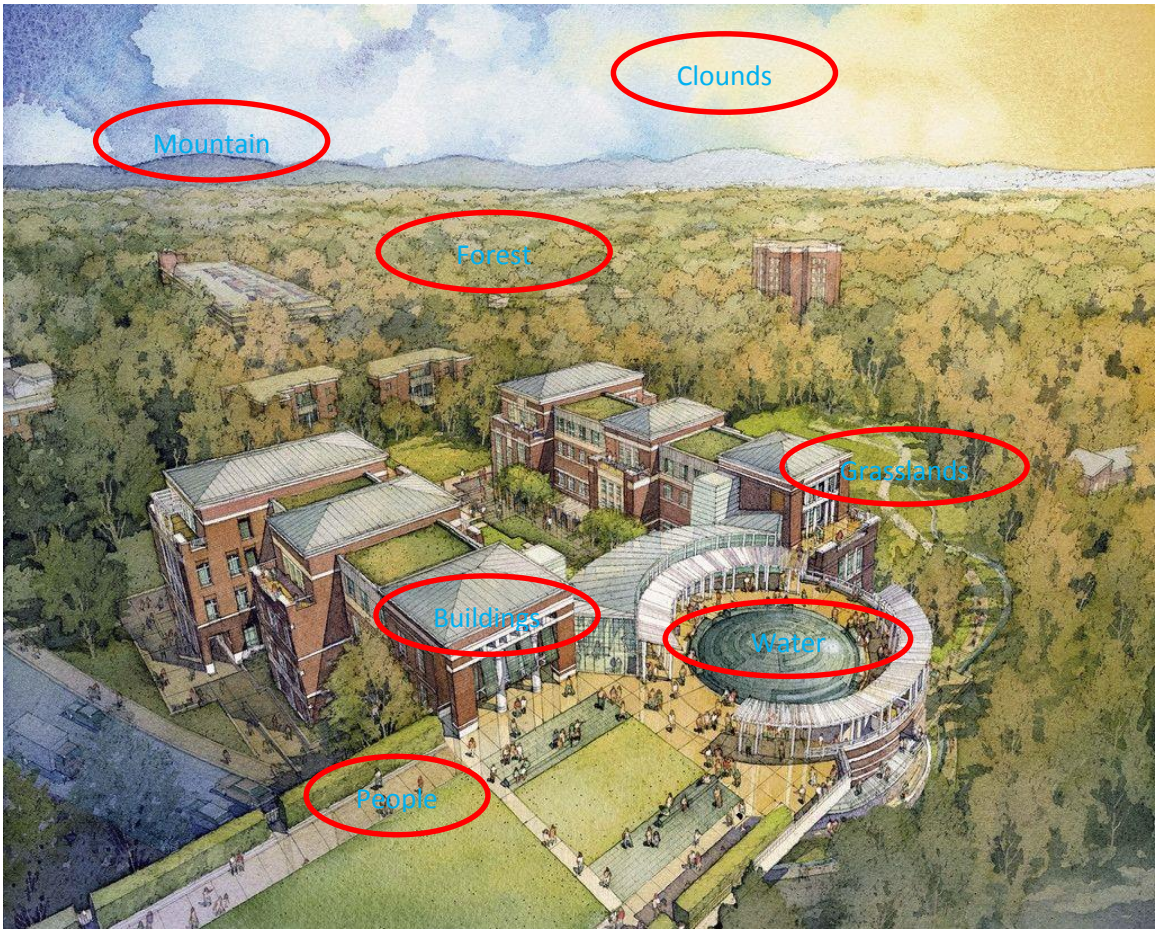

Figure 13. The elements involved in the outdoor scene are complicated from the point of landscape design [57].

As design elements, the landscape of the plant also has such elements as color, size, texture, shape, spatial scale, and so on. The selection of plant landscape is based on these element characteristics of the overall plant landscape type, rather than the individual plants, and follows

the design principles and creative techniques described in plant configuration theory. The plant landscape is composed of many kinds of plants [40]. Although their element characteristics are highly related to individual element characteristics, they are not equal to the element characteristics of an individual plant or the simple superposition of multiple individual plant element characteristics. Sometimes it shows that the essential characteristics are completely different from those of individual plants. It should be noted that the element characteristics of plant landscape types are not only related to the composition of plants within the plant landscape types but also related to the structural arrangement of individual plants. Also, the same plant landscape can have completely different plant structure combinations [58-60].

Furthermore, the mountain has a profound historical significance. Mountain is a landscape mainly composed of natural mountains. It usually has the aesthetic characteristics of magnificence, danger, beauty, and strangeness. Mountain is a regional complex with aesthetic feeling, which takes mountain area as the carrier of tourism resources and landscape elements. Mountain scenic area is the main part of the landscape in China, accounting for about half of 119 national parks. It has attracted numerous tourists and scientific workers from all over the world for sightseeing, scientific investigation, and literary and artistic activities. It has the value of historical culture, art appreciation, and scientific investigation. From the perspective of planning, the purpose of landscape design is usually to provide a comfortable environment and improve the commercial, cultural, ecological value of the area. Therefore, we should grasp the key factors in the design and put forward the basic ideas [61, 62].

Here, we introduce the composition of a place [63-65]. The place is composed of space and characteristics, which can also be understood as resources. For the analysis of the structure of space, one is the point node line surface model, the most typical is Lynch's node – mark - path – edge - region model, and the "outside-inside" mode. The latter can be analyzed through the elements of the bottom, top, enclosure, notch, and boundary, and strengthen the sense of space

through centripetally, direction, and rhythm. In the landscape cognitive model, the space of place phenomenon is more like a box. Whether it is the geomantic pattern, the spatial composition in the paintings, and the sky and earth in religious mythology, all reflect the existence of this spatial pattern. The point, line, and surface model can be combined, which will be more conducive to our grasp of space.

The characteristics of space are determined by more specific material elements and their states, which specifically describe the elements or components of space, the texture of objects, light, color, form, etc., forming the atmosphere of a local place, such as the blue sky, white clouds, black land, dark green forest, mossy roof, the fragrance of bamboo rice. All of these together create the characteristics and atmosphere of a place. These all form the local or geographical character of the landscape. Based on the discussion of the landscape, we set up the labels in corresponding to describe the structure of the landscape and relationship with people.

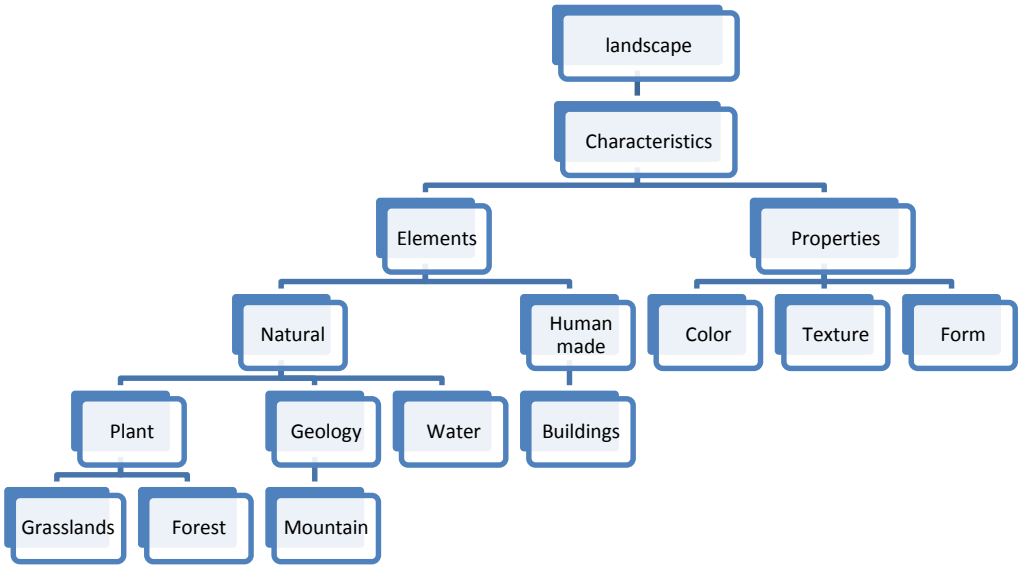

Figure 14. From the different perspective, the landscape can be classified from characteristics and space. In this figure, characteristics are further classified to elements and the properties.

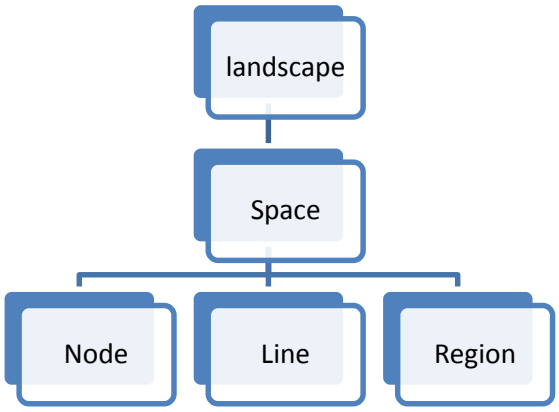

Figure 15. The space considers a big picture with the structure organization.

4.2 Labeling of Landscape Images

Here are some examples: in the following six images downloaded from the website for the location we study – NaLaTi Grassland (downloaded from *http://www.mafengwo.cn/*), we can tell more than 10 different features: forest, animal, car, grassland, mountain, a single tree, river, people, status, flower, and buildings. The problem is how to choose our labels in which not only has strong conceptual support but also is applicable for image clustering.

Figure 16. From the examples of the images on the grassland, we conclude that NaLaTi grassland is a complex outdoor scene that contains multiple features. We need some key elements for the image clustering.

Considering the key components of the grassland, we summarized the following features to be labelled as 1.The mountain; 2. The single element compared with the whole part (tree vs forest); 3. Grassland.

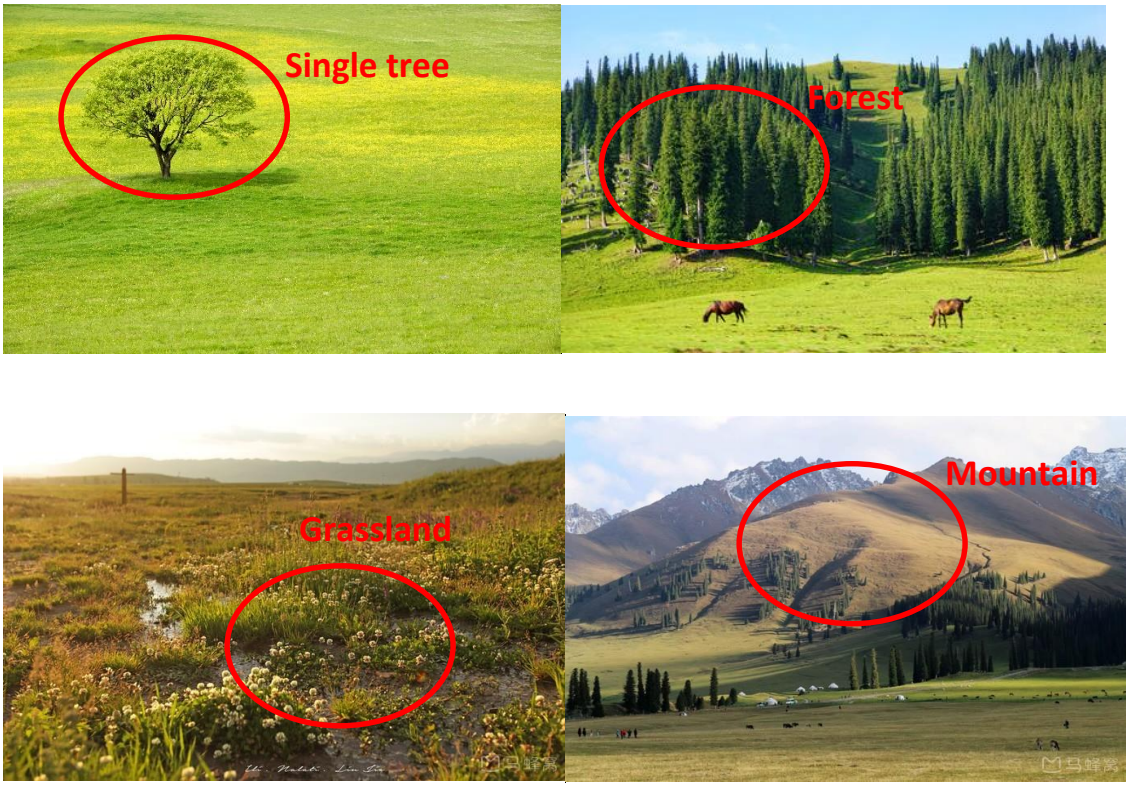

Figure 17. Key elements of the landscape including a tree, forest, grassland, and mountain.

It has been observed that these elements occupied a magnificent part in most of the images, as well as the sky. The ratio of the sky area is another consideration when the landscape is discussed. People can observe meteorological or astronomical phenomena in the sky, to know whether changes, the pass of time, or their position. Sunrise and sunset can tell the time of the day, and the moon's changing tells people the time of a month at night. The big dipper can indicate the north. The thickness and shape of the cloud can tell whether it will rain or not. You can enjoy many beautiful phenomena in the sky, such as rainbow, Aurora, and meteor shower. Birds can fly in the sky. So in conclusion, here is the detailed table of the new label that will help the training of the weakly supervised image clustering.

Table 5. Top: The details of the dataset corresponding to 16 classes for image clustering. Bottom: The new label of the weakly image clustering not only study the four key elements but analyze the impact of the area of the sky. Some typical images are shown here as examples.

	Tree	Forest	Grassland	Mountain
Small sky	70	64	80	120
Medium sky	75	120	100	150
Large sky	100	146	125	120
Major sky	70	80	125	55

	Tree	Forest	Grassland	Mountain
Small sky				
Medium sky				
Large sky				
Major sky				

CHAPTER V

WEAKLY SUPERVISED IMAGE CLUSTERING

In the field of machine learning, learning tasks can be divided into two categories: supervised learning and unsupervised learning. Usually, both of them need to learn the prediction model from the training data set containing a large number of training samples, and each training sample corresponds to the event/object. Classification problems and regression problems are the representatives of supervised learning, while the clustering problem is the representative of unsupervised learning. Although the current supervised learning technology has achieved great success, it is worth noting that due to the high cost of the data labeling process, it is difficult for many tasks to obtain strong indicative information such as all truth labels. However, unsupervised learning develops slowly because the learning process is too difficult. Therefore, we hope that machine learning technology can work under weak supervision.

In this chapter, we will introduce Curriculum Net, which is a weakly supervised method of image clustering, to our dataset. The images in this part come from the NaLaTi grasslands with so many noisy and complex labels. As we introduced in the last chapter, specific labels are designed in order to meet the requirements of image clustering. There are four sections: (1). The Curriculum Net; (2). Hierarchical Density-Based Spatial Clustering of Applications with Noise. (3). Experimental Results.

5.1 Curriculum Net

Curriculum Net takes into account how the learning of people works. Knowledge is well organized in order to learn faster. So, for machine learning, can we improve the performance by changing the order of learning (simply organizing knowledge)? In this report, through experiments, it is found that this is a feasible idea. Changing the order of learning can improve the speed and quality of learning for the following reasons [66-68]:

1. Easy first, then difficult, is an effective and efficient human learning strategy. Can machine learning use this strategy to achieve the same effect? This is what this paper will explore.

2. Deep neural network has a limitation, that is, it often falls into a local minimum value and cannot extricate itself. The method of curriculum learning proposed in this paper can help to find a better local minimum value to improve the results. This statement can be verified by experiments [66, 69].

3. In fact, the layer structure of a deep neural network is a bit similar to our human learning process. For example, the first layer carries out simple abstraction (learning simple things). With the increase of layers, the abstraction gradually deepens. Learning things become more and more difficult, which is consistent with the curriculum viewpoint of this report.

Based on the idea of curriculum learning, which is similar to the human learning process, the model starts with simple problems, and then gradually learns more and more complex task problems.

There are three methods of data monitoring in curriculum net [70]:

1. Initial features generation: Firstly, all the training data are used to learn the initial model; then, the depth feature representation of each image in the training data set is calculated by using the training model (e.g., FC layer output feature)

2. Curriculum design: The initial training model aims to roughly map the training image to the feature space, to mine the potential structure and relationship of the images in each category. It provides an effective method to define the image complexity. The defined image complexity is analyzed to design the learning course. All images in each category are divided into several subsets according to the order of complexity.

3. Curriculum learning: Firstly, the CNN model is trained from the simple data subset containing all categories. Here, we assume that more clean images with accurate tags are included in the simple data subset. Then, in the training process, more and more complex data are added continuously to improve the recognition ability of the model.

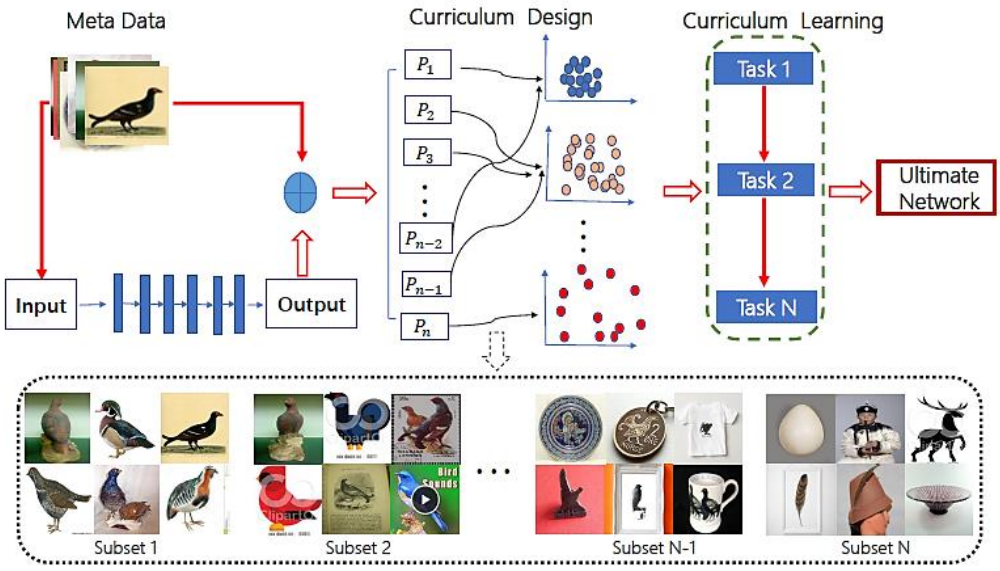

Figure 18. The structure of the Curriculum Net. The process includes feature generation, curriculum design, and curriculum learning [70]. After extract the features from the full connection layer using Inception – V2, we use density-based clustering method to divide the training set to N subset. The 1 to N subset indicate the clean to noisy subsets.

The goal of design course learning is to sort training images from simple to complex in an unsupervised way. In this paper, a density-based clustering algorithm is used to evaluate the complexity of training samples according to data distribution density.

Specifically, the whole training data set is divided into several data subsets, and the data subsets are ranked from simple to complex, in which the simple data subset contains more clean images with more labels and more reliable features, while the complex data subset contains more and more noise tags.

According to the density-based clustering algorithm, each category of the image data set is processed

1. First, train all training datasets on the Inception - V2 model, as the initial model;

2. Then, based on the FC layer features of the initial model, all images in each category are projected into the depth feature space: $P_i - f(P_j)$.

3. Then calculate the Euclidean distance matrix as:

$$D_{ij} = \left\| f(P_i) - f\left(P_j\right) \right\|^2 . \tag{5.1}$$

In here, n is the number of the images in each category. D_{ij} represents the similarity between each image P_i and P_j. The smaller the D_{ij}, the more similar P_i and P_j is.

4. For each image, we calculate their local density ρ_i:

$$\rho_i = \sum_j X\left(D_{ij} - d_c\right). \tag{5.2}$$

In which,

$$X(d) = \begin{cases} 1 & d < 0 \\ 0 & other \end{cases}. \tag{5.3}$$

In here, d_c is the cutting distance by sorting n^2 distance in $D \in R^{n*n}$ in ascending order and select from a certain percent of rank k ranging from 50 – 70, and 60 is set for all the experiments to have suitable value. ρ_i is the number of samples whose distances to i is smaller than d_c.

It is easy to know that clean images composed of correct labels usually have a relatively similar visual representation, and they have larger local density values in the feature projection space. On the other hand, the noise image often has an obvious visual difference, which is sparse distribution in the feature projection space and has a small density value. For each image, we define the distance δ_i:

$$\delta_i = \begin{cases} min_{j:\rho_j > \rho_i}(D_{ij}) & if\ \exists j\ s.t.\ \rho_j > \rho_i \\ max(D_{ij}) & else. \end{cases} \quad (5.4)$$

If there is an image having $\rho_j > \rho_i$, then δ_i is D_{ij} where the j is the sample closest to the i among the data. Or j is the distance between i and the data point which is farthest from i. Therefore, the data points with the maximum local density value have the largest value δ_i and are regarded as the clustering center of the class image in that category.

For each class of image data set, three clusters are generated, and each cluster image is taken as a data subset. Since each cluster contains a density value to measure its data distribution and the relationship between different clusters, it is easy to define the complexity of each data subset and give the design rules of course learning. In the data subset with high-density values, all the images are closer to each other in the feature space, which indicates that these images have a stronger similarity.

In the data subset with low-density values, all images have greater visual representation differences, which may contain more irrelevant images with incorrect labels. Therefore, the data subset is defined as noisy. At this time, we can get three different complexity data subsets: clean,

noise, high noise. Each image category contains the same number of data subsets, and all the image classes are combined into the final course of learning data set as follows:

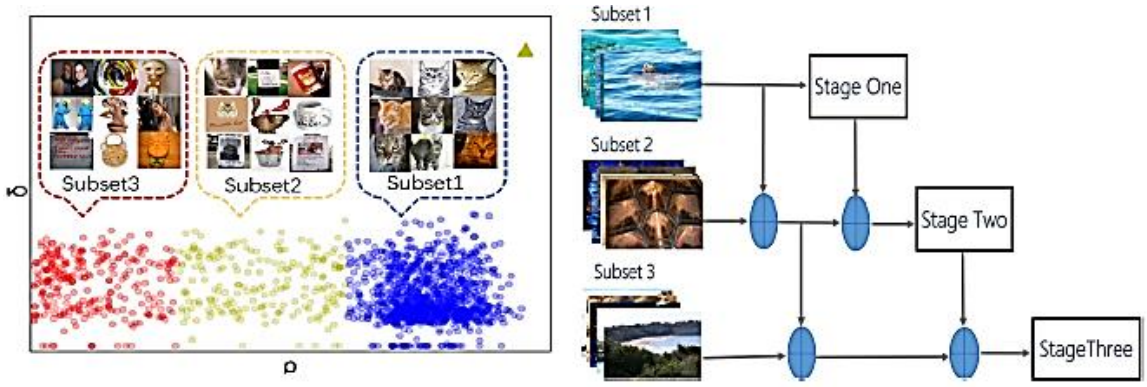

Figure 19. Left: the subsets are divided into three parts from easy to hard. Right: The mode is trained at different stages based on the import of subsets [70].

The designed course can dig up the potential data structure based on the visual representation of images in an unsupervised method. Here, a multi-stage learning scheme is designed, as shown in Fig. (right). CNN model is trained by continuously mixing three stages from a clean data subset to a highly noisy data subset [70].

1. First, train the concept only based on a clean subset of data in the Inception V2 model, each category of the image has a similar visual representation in the data subset, which helps the model to learn the basic and clean visual information of the image as the basic feature of subsequent processing.

2. After the training convergence of the Inception V2 model, noise data is added to continue the learning process. At this time, the image contains more obvious visual differences, which enables the model to learn more meaningful and discriminative features of difficult samples. Although the noise data contains incorrect class labels, it still roughly maintains the main structure of the data.

3. Continue to add noise data of high noise to further train the model. The data set of high noise contains a large number of visual irrelevant images with incorrect class labels. The deep features of the first two stages 1 and 2 can maintain the main potential structure of data. The authors found that the data subset of high noise does not hurt the data structure of learning. On the contrary, it can improve the generalization ability of the model and provide a regularization method to avoid overfitting for clean data.

When the final training model converges, all three data subsets are used. In addition, for samples of different data subsets, different loss weights are set in 2 and 3 stages. For clean, noisy, and highly noisy data subsets, the weights are 1, 0.5, and 0.5, respectively.

5.2 Density Clustering

We can tell that the key to the success of the Curriculum Net is the clustering algorithm employed. The clustering method used in the Curriculum Net is a density-based algorithm by comparing the density peak of each point. However, there are multiple shortcomings for this method, such as that the method is sensitive to the parameters. Furthermore, the density-based algorithm employed certain parameters for image clustering. Obviously, when the density of clustering is different, the clustering effect is also very different. When the data density is not uniform, it is difficult to use the algorithm. Also, if the sample set is large, the convergence will be slow[71, 72].

Hence, we are going to apply an improved clustering method into the Curriculum Net, which is the Hierarchical Density-Based Spatial Clustering of Applications with Noise. "*Performs DBSCAN over varying epsilon values and integrates the result to find a clustering that gives the best stability over epsilon. This allows HDBSCAN to find clusters of varying densities (unlike DBSCAN), and be more robust to parameter selection*" [73].

First, we need to transform the space according to the density. The core distance is introduced here, denoted as $core_k(x)$. The simple way to do this is to define a new distance metric between points which we will call (again following the literature) the mutual reachability distance. We define mutual reachability distance as follows:

$$d_{mreach-k}(a,b) = max\{core_k(a), core_k(b), d(a,b)\}, \qquad (5.4)$$

where $d(a,b)$ is the original distance between a and b. In this formula, dense points (with lower core distance) keep the same distance from each other, but sparse points are pushed away so that their core distance is at least away from any other point. This actually "lowers the sea level", and push the sparse "ocean" points to the outside world, while the "land" is not affected. It should be noted here that this depends on the choice of k, and a larger k value interprets more points as being in the "ocean." All of these are easy to understand in a single picture. We use the k value of 5, and then for a given point, we can draw a circle with the core distance as the circle contacting the sixth nearest neighbor (including the point itself), as shown below:

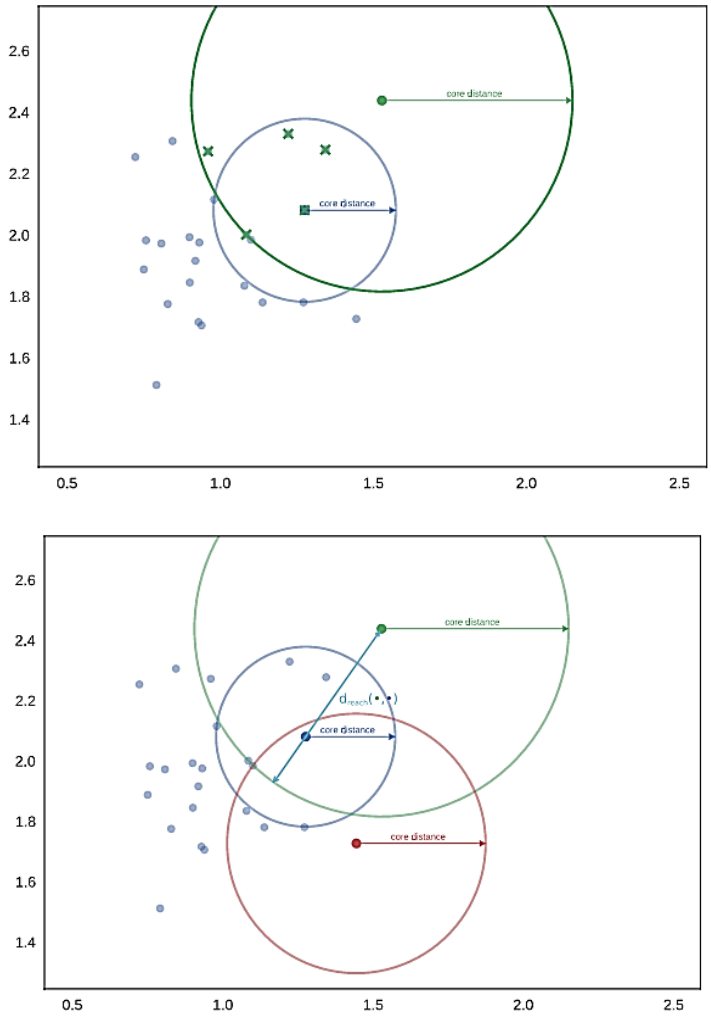

Figure 20. The core distance here is the radius for the circle [74].

Now let us consider the mutual reachability distance between the blue and green centers. The core distance between the green and blue points is their Euclidean distance, while this distance is smaller than the radius of the green cycle. Therefore, we need to mark the mutual reachability distance between blue and green as the radius of green circle. In a similar way, the mutual reachability distance from red to green is simply the pure distance from red to green, because this distance is the largest compared with the other two core distances..

There is a theory that as a kind of transformation, mutual reachability distance can allow single link clustering to be closer to the hierarchical structure of the level set, no matter what the actual density distribution of the sampled points is.

Now we build the minimum spanning tree efficiently through Prim's Algorithm. We build one edge at a time, always adding the smallest weight edge to connect the current tree to vertices that are not yet in the tree.

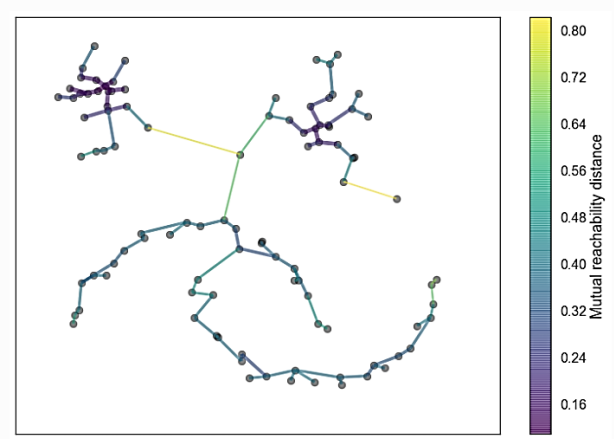

Figure 21. The minimum spanning tree for mutual reachability distance while $k = 5$ [74].

Given the minimum spanning tree, the next step is to convert it into a hierarchy of connected components. It's easy to do this in reverse order: sort the edges of the tree by distance (in ascending order), then traverse, creating a new merged cluster for each edge. The only difficult part here is to determine the edge that joins two clusters together, but it can be easily achieved by union finding data structures. We can think of the results as a tree view. However, what we need is a set of flat clusters. It is simple to just draw a solid line to cut off all the data, which is done in the regular DBSCAN method. What we want to deal with is the variable density cluster. Any pruning choice is a choice based on mutual reachability distance, so it is a single fixed density level. Ideally, we want to be able to choose our clusters by pruning in different places. That's

what HDNSCAN is going to do next, which will create content that is different from the robust single link.

The first step of cluster extraction is to merge the cluster into a smaller hierarchy tree. As you can see in the hierarchy above, cluster splitting is usually the separation of one or two points from a cluster. First, we define a concept of minimum cluster size, which is the minimum points that a cluster cycle required to cover so it can be considered as cluster center. Once we determine the minimum cluster size, we can now decide if the hierarchy tree that has fewer points than the minimum cluster size. If we have a point that is less than the minimum cluster size, we say it as the parent cluster. On the other hand, if it is split into two clusters, each cluster contains more points than the minimum cluster size, then we remain the cluster tree unchanged. In this smaller structure, each node relates to that the point decreases with different distances. We can visualize it as a tree view, similar to the above tree view, which uses the width of the merged line to represent the count of points. However, when points are removed, the width varies with the length of the line. In here, we use data with a minimum cluster size of 5, and the results are as follows:

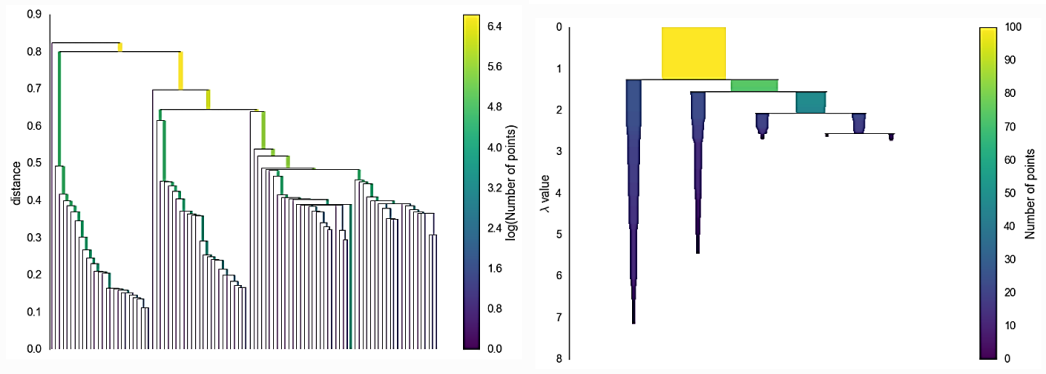

Figure 22. Left: The draft of the minimum spanning tree. Right: The cluster tree after the condensing [74].

In the end, we declare a variable $\lambda = \frac{1}{distance}$ to calculate the cluster stability. For a given cluster, λ_{birth} and λ_{death} represents when cluster is split and became a parent cluster. In turn, for a given cluster, for each point in that cluster we can define the value λ_p as the lambda value at which that point is not a cluster center anymore which is a value between λ_{birth} and λ_{death} since the point either becomes a parent cluster at some point or leaves the cluster when the cluster becomes two smaller clusters.

Now, for each cluster we compute the stability as:

$$\sum_{p \in cluster}(\lambda_p - \lambda_{birth}). \quad (5.5)$$

Now by traversing the tree in reverse sort order, if the sum of the stability of the smaller clusters is greater than the parent cluster, then we set the stability of the parent clusters as the sum value. On the other hand, if the stability of the parent cluster is greater than the sum of its sub-clusters, the parent cluster is selected without considering its sub-clusters. When we reach the end, we return the current selected clusters.

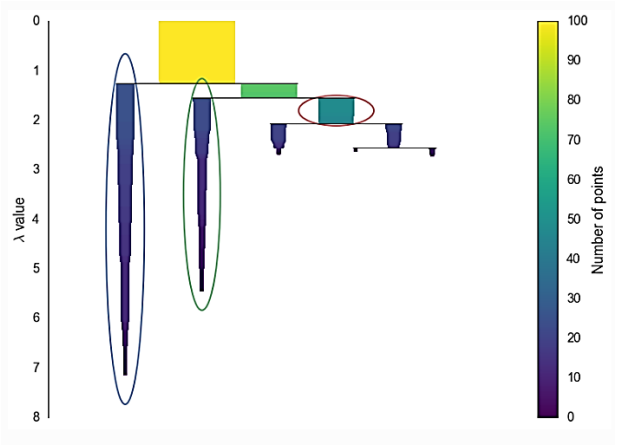

Figure 23. The extraction of the clustering is done by selecting the clusters in the condensed tree. This map can be converted to the clustering labels [74].

5.3 Experimental Results

The data set of NaLaTi Grassland Park is downloaded from the internet by a web crawler. Meta information is along with those web images including tags, descriptions, or dates. The dataset for the park contains 16 object categories (4 elements x 4 various sky area). The training data contains 1,200 images in total with correct human annotations but it includes many noisy features that cannot be labeled precisely, as shown in the Figure 22.

Figure 24. Both images are labelled as single tree, while the right image contains more non-relevant features than the left one.

There are 600 manually labeled images that are used as the validation set. The evaluation is based on the rate of precision by comparing the clustering results with the ground truth.

We conducted extensive experiments to evaluate the efficiency of the regular Curriculum Net and the revised method. We implement the Curriculum Net with Inception V-2 and focus on two parts of the clustering: the 4 classes contain the key elements include a tree, forest, mountain, and grassland versus the 16 classes with the sky area involved. The full results are presented in Tables 2 and 3.

Table 6. Comparison between the density peak and the HDNSCAN. It showed the HDBSCAN performs better than the density peak method.

Clustering Algothrim	4 classes	16 classes
Density Peak	36.7%	24.8%
HDBSCAN	47%	28.3%

Table 7. Comparison between the regular Curriculum Net and the revised version.

	4 classes	16 classes
K - means	40%	21.57%
Curriculum Net	48.4%	31.2%
Revised C - Net	63%	37.1%

It is found that (i), the performance is generally improved by Curriculum Net, such as 40% → 48.4% for 4 classes and 21.57% → 31.2% for 16 classes; (ii); The learning strategy is important for the results. By improving the clustering algorithm, it boots the performance from a rate of precision of 48.4% → 63% for 4 classes and 31.2% → 37.1% for 16 classes. This performance gain is due to that more clean data is imported into the mode which confirms the strong capability of Curriculum Net on learning from noisy data.

CHAPTER VI

CONCLUSION

In this study, emotion recognition and image clustering in the wild under a realistic environment have been investigated. To address the challenges in our research, we applied various deep learning approaches and tried to improve them with some additional schemes. As for emotion recognition, we upgraded the face detector to detect more face reliably and accurately. The training and test sets were expanded to include faces occluded by sunglasses and children's faces which are very rare in other public databases. In this way, the network would be fully trained and have more case-specific features to learn. Regarding scene-based image clustering, we found that proper preparation of the dataset is essential to the clustering performance. We downloaded the images from online sources and carefully erased the ones that were not relevant. As a result, we designed 16 new classes of the images by according a systematic visual analysis of landscape. Then we introduced Curriculum Net into for weakly supervised image clustering that simulates the human learning strategy - easy first, then difficult, was an effective and efficient learning strategy. In Curriculum Net, the training dataset is split into two parts by the employment of Density Peak Clustering algorithm: clean and noisy datasets, and then the CNN network can be initialized by learning from the lean dataset and then fine-tuned using more noisy data. The efficiency of the training and final clustering results is highly dependent on the selected clustering algorithms and its clustering datasets. However, there are shortcomings for the Density Peak

Clustering algorithm such as the large time of calculation and the high sensitivity to the parameters. To address the limitation of the previous clustering method, we adopted a better density-based clustering technique with less parameters and low running space, called Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), to create the clean and noisy training subsets that allows Curriculum Net to be sufficiently trained. In this way, the image clustering performance was improved. In conclusion, this research has studied the two most popular aspects in emotion recognition and image clustering. For each aspect, we tested the powerful algorithms on our real – world case images. Then we introduce the approaches to conquer the challenges of the error of the small face, occlusion and child face for emotion recognition. After the labeling of the landscape images, we further analyzed the impact of the existing clustering algorithm on the Curriculum Net. By the changing, we enhanced the performance of both emotion recognition and image clustering and provide the work for the future research to further reveal the relationship between the two aspects in real world scenarios.

REFERENCES

[1]. Baltrušaitis, T., C. Ahuja, and L.-P. Morency, *Multimodal machine learning: A survey and taxonomy*. IEEE transactions on pattern analysis and machine intelligence, 2018. **41**(2): p. 423-443.

[2]. Li, S. and W. Deng, *Deep facial expression recognition: A survey.* IEEE Transactions on Affective Computing, 2020.

[3]. Min, E., et al., *A survey of clustering with deep learning: From the perspective of network architecture.* IEEE Access, 2018. **6**: p. 39501-39514.

[4]. Burgos-Artizzu, X.P., P. Perona, and P. Doll ár. *Robust face landmark estimation under occlusion*. in *Proceedings of the IEEE international conference on computer vision*. 2013.

[5]. Ekman, P. and W.V. Friesen, *Constants across cultures in the face and emotion.* Journal of personality and social psychology, 1971. **17**(2): p. 124.

[6]. Ekman, P., *Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique.* 1994.

[7]. Jack, R.E., et al., *Facial expressions of emotion are not culturally universal.* Proceedings of the National Academy of Sciences, 2012. **109**(19): p. 7241-7244.

[8]. Shan, C., S. Gong, and P.W. McOwan, *Facial expression recognition based on local binary patterns: A comprehensive study.* Image and vision Computing, 2009. **27**(6): p. 803-816.

[9]. Martinez, B. and M.F. Valstar, *Advances, challenges, and opportunities in automatic facial expression recognition*, in *Advances in face detection and facial image analysis*. 2016, Springer. p. 63-100.

[10]. Liu, P., et al. *Facial expression recognition via a boosted deep belief network*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[11]. Mollahosseini, A., D. Chan, and M.H. Mahoor. *Going deeper in facial expression recognition using deep neural networks*. in *2016 IEEE Winter conference on applications of computer vision (WACV)*. 2016. IEEE.

[12]. Mitchell, T.M., R.M. Keller, and S.T. Kedar-Cabelli, *Explanation-based generalization: A unifying view.* Machine learning, 1986. **1**(1): p. 47-80.

[13]. Barlow, H.B., *Unsupervised learning.* Neural computation, 1989. **1**(3): p. 295-311.

[14]. Zhu, X. and A.B. Goldberg, *Introduction to semi-supervised learning.* Synthesis lectures on artificial intelligence and machine learning, 2009. **3**(1): p. 1-130.

[15]. Pan, S.J. and Q. Yang, *A survey on transfer learning.* IEEE Transactions on knowledge and data engineering, 2009. **22**(10): p. 1345-1359.
[16]. Sutton, R.S. and A.G. Barto, *Reinforcement learning: An introduction*. 2018: MIT press.
[17]. Schoenfeld, D., *Partial residuals for the proportional hazards regression model.* Biometrika, 1982. **69**(1): p. 239-241..
[18]. Ren, X. and J. Malik. *Learning a classification model for segmentation*. in *null*. 2003. IEEE.
[19]. Tello, G., et al., *Deep-structured machine learning model for the recognition of mixed-defect patterns in semiconductor fabrication processes.* IEEE Transactions on Semiconductor Manufacturing, 2018. **31**(2): p. 315-322.
[20]. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436-444.
[21]. Goodfellow, I., et al., *Deep learning*. Vol. 1. 2016: MIT press Cambridge.
[22]. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *Imagenet classification with deep convolutional neural networks.* Communications of the ACM, 2017. **60**(6): p. 84-90.
[23]. Szegedy, C., et al. *Going deeper with convolutions*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
[24]. Simonyan, K. and A. Zisserman, *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014.
[25]. Kim, B.-K., et al. *Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition*. in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 2015.
[26]. Yu, Z. and C. Zhang. *Image based static facial expression recognition with multiple deep network learning*. in *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 2015.
[27]. Meng, Z., et al. *Identity-aware convolutional neural network for facial expression recognition*. in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. 2017. IEEE.
[28]. Abbasnejad, I., Sridharan, S., Nguyen, D., Denman, S., Fookes, C., & Lucey, S. *(2017). Using synthetic data to improve facial expression analysis with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 1609-1618)..*
[29]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. *(2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).*
[30]. Levi, G. and T. Hassner. *Emotion recognition in the wild via convolutional neural networks and mapped binary patterns*. in *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 2015.
[31]. Liu, X., et al. *Adaptive deep metric learning for identity-aware facial expression recognition*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017.
[32]. Huang, R., et al. *Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis*. in *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[33]. Liao, K., et al., *DR-GAN: Automatic radial distortion rectification using conditional GAN in real-time.* IEEE Transactions on Circuits and Systems for Video Technology, 2019. **30**(3): p. 725-733.
[34]. Jia, R., T. Li, and F. Yuan. *FF-GAN: Feature Fusion GAN for Monocular Depth Estimation*. in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. 2020. Springer.
[35]. Yang, H., U. Ciftci, and L. Yin. *Facial expression recognition by de-expression residue learning*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
[36]. Ren, Y., et al., *Almost unsupervised text to speech and automatic speech recognition.* arXiv preprint arXiv:1905.06791, 2019.
[37]. Huang, F., et al., *Automatic classification of single-molecule charge transport data with an unsupervised machine-learning algorithm.* Physical Chemistry Chemical Physics, 2020. **22**(3): p. 1674-1681.
[38]. Alami, N., et al., *Using unsupervised deep learning for automatic summarization of Arabic documents.* Arabian Journal for Science and Engineering, 2018. **43**(12): p. 7803-7815.
[39]. Li, X., et al., *Deep learning-based unsupervised representation clustering methodology for automatic nuclear reactor operating transient identification.* Knowledge-Based Systems, 2020. **204**: p. 106178.
[40]. Le, D.T. and L.M. Le. *CDNN Model for Insect Classification Based on Deep Neural Network Approach*. in *Context-Aware Systems and Applications, and Nature of Computation and Communication: 8th EAI International Conference, ICCASA 2019, and 5th EAI International Conference, ICTCC 2019, My Tho City, Vietnam, November 28-29, 2019, Proceedings*. 2019. Springer Nature.
[41]. Yang, Y., *Medical Multimedia Big Data Analysis Modeling Based on DBN Algorithm.* IEEE Access, 2020. **8**: p. 16350-16361.
[42]. Zhao, W., et al., *A semantic segmentation algorithm using FCN with combination of BSLIC.* Applied Sciences, 2018. **8**(4): p. 500.
[43]. Bottou, L., *Large-scale machine learning with stochastic gradient descent*, in *Proceedings of COMPSTAT'2010*. 2010, Springer. p. 177-186.
[44]. Hecht-Nielsen, R., *Theory of the backpropagation neural network*, in *Neural networks for perception*. 1992, Elsevier. p. 65-93.
[45]. Goodfellow, I.J., et al., *Generative adversarial nets*, in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. 2014, MIT Press: Montreal, Canada. p. 2672–2680.
[46]. Tetteh, G., et al. *Deep-FExt: Deep feature extraction for vessel segmentation and centerline prediction*. in *International Workshop on Machine Learning in Medical Imaging*. 2017. Springer.
[47]. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[48]. Szegedy, C., et al., *Inception-v4, inception-resnet and the impact of residual connections on learning.* arXiv preprint arXiv:1602.07261, 2016.
[49]. He, K., et al. *Identity mappings in deep residual networks*. in *European conference on computer vision*. 2016. Springer.

[50]. Arriaga, O., M. Valdenegro-Toro, and P. Plöger, *Real-time convolutional neural networks for emotion and gender classification.* arXiv preprint arXiv:1710.07557, 2017.
[51]. Howard, A.G., et al., *Mobilenets: Efficient convolutional neural networks for mobile vision applications.* arXiv preprint arXiv:1704.04861, 2017.
[52]. Yang, B.-E. and T.J. Brown, *A cross-cultural comparison of preferences for landscape styles and landscape elements.* Environment and behavior, 1992. **24**(4): p. 471-507.
[53]. Dupont, L., M. Antrop, and V. Van Eetvelde, *Eye-tracking analysis in landscape perception research: Influence of photograph properties and landscape characteristics.* Landscape Research, 2014. **39**(4): p. 417-432.
[54]. Forman, R.T., *Interaction among landscape elements: a core of landscape ecology.* Perspectives in landscape ecology, 1981: p. 35-48.
[55]. Tveit, M., Å. Ode, and G. Fry, *Key concepts in a framework for analysing visual landscape character.* Landscape research, 2006. **31**(3): p. 229-255.
[56]. Ja'afar, N.H., A.B. Sulaiman, and S. Shamsuddin, *landscape features and traditional streets character in Malaysia.* Asian Journal of Environment-Behaviour Studies, 2018. **3**(8): p. 121-131.
[57]. TuXingTianXia, *https://www.photophoto.cn/pic/10373294.html.* 2018.
[58]. Kim, K.-S., et al., *Development of flower color changed landscape plant through interspecific and intergeneric crosses of several Cruciferae crops.* Korean Journal of Plant Resources, 2018. **31**(1): p. 77-85.
[59]. Chu, S., et al., *Effects of landscape plant species and concentration of sewage sludge compost on plant growth, nutrient uptake, and heavy metal removal.* Environmental Science and Pollution Research, 2018. **25**(35): p. 35184-35199.
[60]. Green, B.J., et al., *Landscape plant selection criteria for the allergic patient.* The Journal of Allergy and Clinical Immunology: In Practice, 2018. **6**(6): p. 1869-1876.
[61]. Antonelli, A., et al., *Geological and climatic influences on mountain biodiversity.* Nature Geoscience, 2018. **11**(10): p. 718-725.
[62]. Rumpf, S.B., et al., *Range dynamics of mountain plants decrease with elevation.* Proceedings of the National Academy of Sciences, 2018. **115**(8): p. 1848-1853.
[63]. Thornett, L., *Landscape, Place and the Gothic in Contemporary Australian Scenography.* 2016.
[64]. Nairn, K., P. Kraftl, and T. Skelton, *Space, place and environment*. 2016: Springer.
[65]. Duarte, F., *Space, place and territory: A critical review on spatialities*. 2017: Taylor & Francis.
[66]. Miyazaki, K. and M. Ida. *Consistency Assessment between Diploma Policy and Curriculum Policy using Character-level CNN*. in *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)*. 2018. IEEE.
[67]. Weinshall, D., G. Cohen, and D. Amir, *Curriculum learning by transfer learning: Theory and experiments with deep networks.* arXiv preprint arXiv:1802.03796, 2018.

[68]. Appalaraju, S. and V. Chaoji, *Image similarity using deep CNN and curriculum learning.* arXiv preprint arXiv:1709.08761, 2017.
[69]. Wang, Q. and T.P. Breckon, *Segmentation Guided Attention Network for Crowd Counting via Curriculum Learning.* arXiv preprint arXiv:1911.07990, 2019.
[70]. Guo, S., et al. *Curriculumnet: Weakly supervised learning from large-scale web images*. in *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
[71]. McInnes, L., J. Healy, and S. Astels, *hdbscan: Hierarchical density based clustering.* Journal of Open Source Software, 2017. **2**(11): p. 205.
[72]. McInnes, L. and J. Healy. *Accelerated hierarchical density based clustering*. in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. 2017. IEEE.
[73]. Ghamarian, I. and E. Marquis, *Hierarchical density-based cluster analysis framework for atom probe tomography data.* Ultramicroscopy, 2019. **200**: p. 28-38.
[74]. Leland, M., John, H, Steve, A. *hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html*, 2018

VITA

Yanyao Li

Candidate for the Degree of

Master of Science

Thesis: SEMATIC UNDERSTANDING OF LARGE – SCALE OURDOOR WEB IMAGES: FROM EMOTION RECOGNITION TO SCENE CLASSIFICATION

Major Field: Electrical Engineering

Biographical:

Education:

Completed the requirements for the Master of Science in Electrical Engineering at Oklahoma State University, Stillwater, Oklahoma in December, 2020.

Completed the requirements for the Bachelor of Science in Petroleum Engineering at The University of Tulsa, Tulsa, Oklahoma, in 2016.

Experience:

Visual Computing and Image Processing Lab (VCIPL) at Oklahoma State University, Stillwater, Oklahoma, Sep 2019 – Present.