감사의 글

딥러닝 전문가라는 꿈을 이루기 위하여 석사과정을 시작하면서 많은 분들의 도움을 받아 졸업하게 되었습니다. 2년이라는 길면서도 짧은 기간 동안 연구실 생활을 하면서 저를 이끌어주시고 응원해주신 모든 분들께 감사의 마음을 전하고자 합니다.

먼저, 석사 과정 동안 지도해주신 김병규 교수님께 감사드립니다. 제가 연구하고자 하는 분야에 대하여 많은 지지를 해주시고, 방황하지 않도록 이끌어주심에 감사합니다. 졸업 논문을 준비하는 동안 자신감을 잃고 소극적으로 연구하는 저에게 야단치기 보단 묵묵히 기다려주셔서 좋은 성과를 낼 수 있었습니다. 또한 열정적으로 학문을 탐구하는 자세와 연구자가 가져야할 마음가짐을 배울 수 있었습니다. 항상 학생들의 입장에서 생각하시면서 향후 어떤 연구자가 되어야 할지 진지하게 함께 고민해주셔서 앞으로 나아가야 할 자리를 잡을 수 있었습니다.

석사 학위 논문 심사를 맡아주신 박영호 교수님과 강지우 교수님께도 감사의 말씀을 드립니다. 심사 과정 중 논문의 부족한 점을 지적해 주시고 더 나은 방향으로 이끌어 주셔서 논문을 완성할 수 있었습니다. 귀중한 시간을 내어 미래에 대한 격려와 따뜻한 조언을 해 주셔서 정말 감사드립니다.

저의 마음의 안식처이자 언제나 활기찬 연구실 생활을 만들어준 IVPL 연구실 동료분들께 정말 감사하다는 말을 전하고 싶습니다. 연구실에 어려움이 있을 때 마다 앞장서서 도와주시겠다고 하는 영운님, 항상 연구에 대하여 조언을 아끼지 않아주고 슬럼프에 빠질 때 마다 도와주는 영주언니,

연구실의 새벽과 주말을 함께한 메이트이자 본 받을 점이 많은 혜진언니, 묵묵히 연구하면서 재치있게 분위기를 풀어주는 정인이, 연구실 생활동안 함께 과제를 진행하며 성장할 수 있도록 도와준 예지언니, 연구실의 막내이지만 언니들 보다 더 언니같아 배울 점이 많던 한림이까지 모두들 감사합니다. 비록 제가 석사 생활을 시작했을 때 바로 졸업을 하여 함께 연구실 생활을 한 기간은 짧지만, 졸업 이후에도 간간히 연락을 하면서 항상 응원해준 영진이와 운하에게도 감사하다는 말을 전합니다.

언제나 아낌없는 지원과 응원을 해준 가족과 친구들에게도 감사를 전합니다. 석사 과정을 하면서 항상 멋있다고 격려해주시는 아버지, 어려움이 있을 때 실질적 해결책을 주시려는 어머니, 친구같은 둘째 동생 예린이, 그리고 가장 사랑하는 막내 리나 언제나 응원하고 위로해 주어서 감사합니다. 나의 가장 친한 친구이자 힘들때 마다 격려해준 채린, 존재만으로 든든한 살리고 멤버들 주현, 소연 그리고 수진이에게 감사드립니다. 제주도 메이트이자 본격적인 프로그래밍 공부를 시작하는데 도움을 많이 준 소희, 유진, 민지 그리고 수빈에게도 감사드립니다. 그리고 졸업프로젝트할때 카메라를 빌려주고 대학원 생활에 대하여 소소한 팁들을 남겨준 강타님께도 감사합니다.

많은 사람들의 지지가 있었기에 학위과정을 포기하지 않고 해낼 수 있었다고 생각합니다. 석사 학위 과정 동안 배운 지식과 터득한 지혜로 사회에 도움이 되는 사람이 되도록 노력하겠습니다. 감사합니다.

Contents

Contents

List of Tables iv

List of Figures v

List of Abbreviations vii

Abstract viii

I Introduction 1

- 1.1 Motivation . . . 1
- 1.2 Related Works . . . 5
  - 1.2.1 Relation Network . . . 6
- 1.3 Contributions . . . 7
- 1.4 Outline . . . 8

II Preliminaries 10

2.1 Overview of FER 10

2.1.1 FER issues related to methods and datasets . . . 13

2.2 Overview of FER with FSL 14

2.2.1 Generalization on novel data 17

2.2.2 Domain adaptation 18

III Proposed Method 20

3.1 Data Preprocessing 20

3.2 Network architecture 23

3.2.1 Feature embedding 25

3.2.2 Channel Selection 27

3.2.3 Emotion similarity learning 31

3.3 Overall training process 32

IV Experimental Results and Discussion 35

4.1 Dataset 35

4.1.1 Dataset division 35

4.1.2 Dataset Construction 37

4.2 Training Details 39

4.3 Performance Analysis 40

4.4 Ablation study 46

4.4.1 Ablation study comparing the channel information of sample feature and average feature 46

4.4.2 Ablation study on the impact of feature size in CS module 48
4.4.3 Ablation study on fine-tuning weights on individual modules 49

V Conclusion 52

References 54

Abstract in Korean 66

List of Tables

2.1 Various issues existing in facial datasets 15
2.2 Various issues related to methods 16

4.1 The number of images of each class for experiments. . . . 36
4.2 Value of nsample in one training episode. 38
4.3 Environment condition for experiments. 39
4.4 The comparison of additional lip and eye data accuracy (%). 39
4.5 The performance comparison of FER accuracy (%). . . . 42
4.6 The performance comparison based on the CS module's channel size (%). 48
4.7 The comparison of accuracy (%) in different weights on individual modules. 50
4.8 The comparison accuracy (%) of various combinations of weights on each individual module. 50

List of Figures

1.1 *In-the-lab* vs *in-the-wild*. 2

1.2 Deep learning-based SOTA model accuracy for FER datasets. 3

1.3 Relation Network architecture. 7

2.1 The general pipeline of Facial Expression Recognition task. 13

3.1 A process of getting aligned faces from video dataset. . . 21

3.2 Overview of MAXDIST peak frame selection method [1]. 22

3.3 An example of selecting a peak frame. 24

3.4 A process of cropping eye and lip region by landmark information. 24

3.5 The overview architecture of the proposed model. 26

3.6 The overview of the Channel Selection process. 29

3.7 The whole pipeline of Channel Selective Spatial Relation Network. 34

4.1 An accuracy of the proposed method with the ratios of the chosen parameter λ of channel selection on the RAFDB. 40

4.2 Visual comparison between the sample channel features and averaged channel features. 47

† List of Abbreviations

- FER : Facial expression recognition.
- HCI : Human-computer interaction.
- CV : Computer vision.
- FSL : Few-shot learning.
- CNN : Convolutional neural network.
- FPS : Frames per second.
- CE : Cross entropy.
- MSE : Mean squared error.
- ReLU : Rectified linear unit.
- FC : Fully connected.
- StepLR : Step learning rate.
- GPU : Graphics processing unit.
- CPU : Central processing unit.
- OS : Operating system.

ABSTRACT

Efficient Few-shot Learning based on Channel Selective Spatial Relation Network for Facial Expression Recognition

Chae-Lin Kim
Department of IT Engineering
The Graduate School
Sookmyung Women's University

Facial expression recognition (FER) is one of the essential tasks in both computer vision and human-computer interaction (HCI) fields. It has been widely used in applications such as autonomous driving, robotics, and e-learning enhancement by recognizing emotion through facial expressions. Though its practicality, Convolution Neural Network (CNN) -based FER have fallen into the overfitting problem due to the few numbers of samples available in the FER dataset.

To address this issue, we propose to a few-shot learning (FSL) method for FER. FSL is a training mechanism that can predict new categories of samples with only a few data. It learns the relation between data by similarity learning and inference test data by way of learning. In this way, FSL can help to solve the overfitting problem in FER.

This thesis proposes a method using the relationNet, which learns relation similarity among datasets. Based on the relationNet, we design a channel selection module and additional spatial data construction. To effectively exploits the best from a few datasets, we make a representative feature as an averaged feature of sample features. Then this representative feature of each channel is compared with each channel information of sample features to find which sample channel feature is the most similar channel information. By comparing channel information, the channel from a selected sample is extracted as an optimal channel of the corresponding sample feature. Therefore, one reconstructed feature is composed of each sample's channel information by the designed module. Focusing on fine-grained features, we figure out that facial expressions have significant information on eyes and lip area. We generate eyes and lip image patches and set this additional data as support and query sets.

We prove that the selected optimal feature and additional spatial information can improve the generalization performance. Comparing to the existing method, the average performances on RAFDB, FER2013, SFEW, and AFEW datasets are increased by 3.5%, 3.68%, 5.58%, and 2.31% of accuracy, respectively.

keywords: Facial expression recognition (FER), Few-shot learning (FSL), Computer vision (CV), Deep learning, Channel selective spatial information

student number: 2131610

Chapter I

Introduction

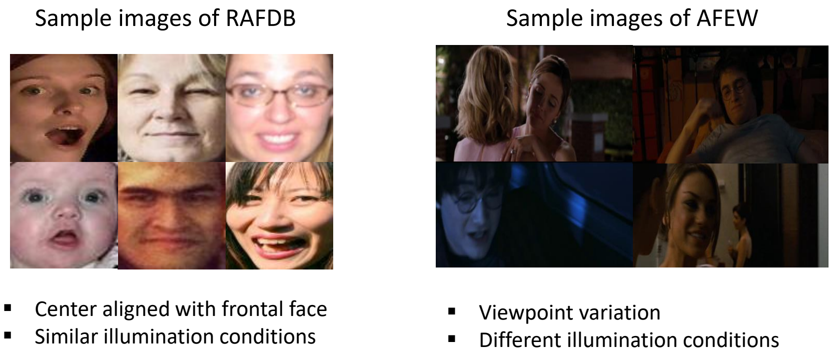
1.1. Motivation

Machine learning has been researched with the development of large-scale datasets and resources. Various technology fields, especially, computer vision (CV) and human-computer interaction (HCI) fields have been developed along with deep learning. Both fields have been actively studied since it directly relates to human convenient life. HCI focuses on human senses and CV substitutes human vision. Therefore, the computer vision field has been significantly studied and broadens possibilities in various tasks such as autonomous driving, robotics, e-learning enhancement, cashier-less store, and more.

Facial Expression Recognition (FER) is one of the significant research fields in both CV and HCI. FER widely studied with deep learning networks [2, 3, 4, 5]. The usage of FER is prevalent in various systems such as feedback for e-learning enhancement, driver fatigue surveillance, and robotics.

However, FER based on deep learning has a generalization problem. Deep learning SOTA models have been often renewed due to high-quality datasets like ImageNet [6] and MNIST [7]. However, it is hard to gather large and high-quality datasets in a real industry condition. The situation is similar in FER tasks. Collecting actual human face images is hard and limited due to privacy issue.

Facial images can be classified as *in-the-lab* dataset and *in-the-wild* dataset. *In-the-lab* dataset usually provides clean and high-quality data that is collected in a controlled environment. On the other hand, *in-the-wild* dataset usually contains spontaneous facial expressions captured in various environmental conditions [8].



Sample images of RAFDB

- Center aligned with frontal face
- Similar illumination conditions

Sample images of AFEW

- Viewpoint variation
- Different illumination conditions

Fig. 1.1: *In-the-lab* vs *in-the-wild*.

In Figure 1.1, we set sample images of RAFDB and AFEW to compare differences. Faces in RAFDB aligned in frontal faces with similar illumination conditions. However, AFEW images contain large space backgrounds and each data has different ranges of brightness.

Due to the data differences, the performance gap between datasets is large. In Figure. 1.2, the State-of-the-Art (SOTA) accuracy of each FER dataset is illustrated: CK+ (Extended Cohn-Kanade), RAFDB (Real-world Affective Face Database), FER+, FER2013, AFEW (Acted Facial Expression in the Wild), SFEW (Static Facial Expressions in the Wild). The performance gets lower as the aspect of the dataset gets close to real-world conditions. This is the overfitting problem of FER, which have low generalization in the real world.

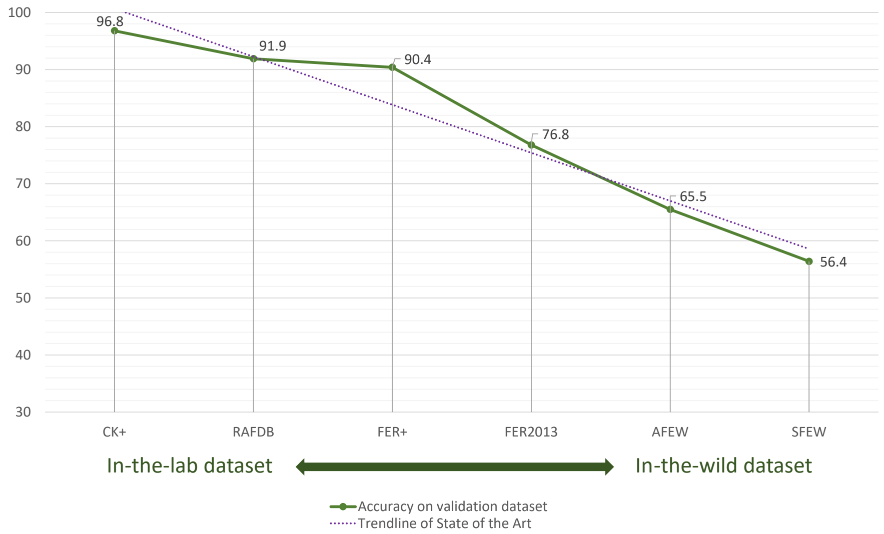


Fig. 1.2: Deep learning-based SOTA model accuracy for FER datasets.

To address the problem of insufficient facial data and the generalization problem of FER, models applying few-shot learning have been researched [9, 10, 11, 12, 13, 14]. Few-shot learning on FER can be classified into two scenarios: generalization on novel data and cross-domain adaptation [15].

Few-shot FER for generalization on novel data aims at generalization with insufficient data using the same domain dataset. It can be classified into two training strategies. The first strategy is to train on half of the basic emotion category and use the rest of the emotion category for the test to see its performance on unseen categories [14]. The second strategy is to use all the 7 basic emotion categories for the train and test like supervised learning [9, 10]. On the other hand, few-shot FER for cross-domain adaptation aims to classify compound emotion categories. To see the generalization performance on the unseen domain dataset, train on 7 basic emotion categories and test on compound emotion, which has more than 11 emotion categories [11, 12, 13].

Focusing on generalization issue, we figure out class imbalance is one of the major problem of few-shot FER and set our model on the first scenario to show the generalization performance. Zhu *et al.* also focuses on class imbalance as the main problem of generalization [14]. They propose Convolutional Relation Network (CRN) and set basic emotion classes with larger samples as the train dataset and classes with fewer samples as the test dataset. They designed depth average pooling (DAP) to fed averaged feature of samples from the same category into relationNet. After that, they computed Jensen-Shannon Divergence (JSD) between the distribution of averaged feature and query feature to train general information [14].

However, there is information loss from the original data while making the average feature. To address this problem, we propose a channel selection (CS) module instead of DAP and this could prevent information loss. The CS compares the similarity of sample features and average feature for each channel. A feature of the sample with the most similar channel is selected and composed of a reconstructed feature. Also, we use additional spatial information on facial regions to enhance the generalization performance. Through these approaches, we could utilize optimal features and significant spatial information, and shows improved general performance.

Before presenting our approach, we will discuss few-shot learning as related works.

1.2. Related Works

Few-shot classification aims to recognize novel categories with a limited number of labeled examples in each class. Unlike supervised learning, few-shot learning (FSL) models define the feature distribution of unseen classes by learning from given training domain datasets. Basically, few-shot learning is also known as n-way k-shot learning. “n” is The number of classes (or categories) that model will classify, and “k” is the number of samples for each class. Therefore, training on the example of data is called one-shot learning [16], and training example’s sub-information without example is called zero-shot learning [17].

In few-shot scenarios, support and query are needed. The support is a training dataset with labels and the query is testing data without labels. In training, the support set learns its relationship while predicting where the query belongs. The following three networks have been mainly used in FER with FSL: Prototypical network [18], Matching network [19], and Relation network [20]. In this paper, we discuss the Relation network since we designed our model based on Relation network.

1.2.1. Relation Network

Inspired by [19, 18], relation network propose two-branch training: embedding module f_φ and relation module g_ϕ. The network performs by learning to compare unlabeled query set against few-shot labeled sample images. The embedding module generates representations from given support and query data. The produced feature map of each support data pairs with the feature map of query data by depth-wise concatenation. Finally, the combined feature map of the support and query are fed into the relation module g_ϕ to determine if they are from matching categories or not. Through this learning strategy, it encompasses both few and zero-shot learning.

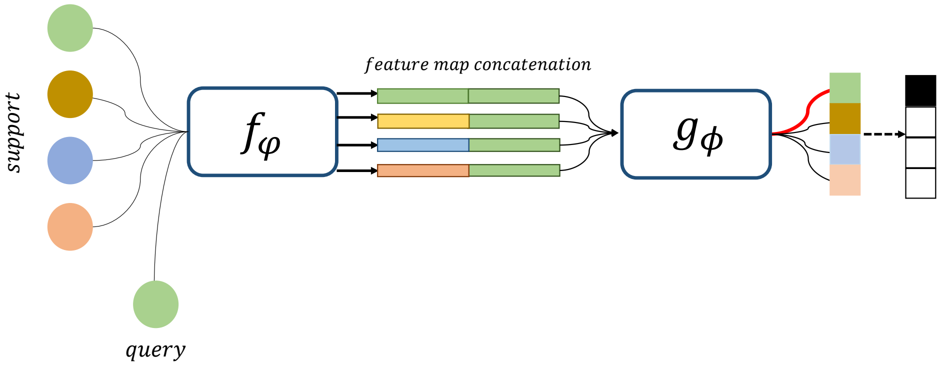


Fig. 1.3: Relation Network architecture.

The embedding and relation modules of relation network are meta-learned end-to-end to support few-shot learning. In this context, relation network is an extended strategy from prototypical network and Matching network. The designed architecture can extract transferrable knowledge that performs better in few-shot learning on support and query.

1.3. Contributions

The main contributions of this study are presented as follows:

- We propose channel selection module to generate the optimal feature for learning similarity without information loss. By comparing the representative channel feature with average channel feature, it is possible to select an optimal channel from support features and improve the generalization performance of few-shot FER.

- We design additional spatial data construction to alleviate the skewed performance of each category. The overall accuracy improved by utilizing eye and lip spatial images and facial images. At the same time, the performance deviation of each category is reduced.

- We redefine the loss function by adding the channel selection feature score and each relation score of spatial information. We achieve performance enhancements on the SFEW and AFEW dataset despite the challenges of improving performance on the *in-the-wild* dataset.

1.4. Outline

The remainder of the thesis is divided into four chapters. The following summary provides an overview of each chapter.

Chapter 2: This chapter presents preliminaries of the FER and FSL. Primarily, we introduced FER and the history of the development of deep learning-based FER. While focusing on issues of deep-learning based FER by given training dataset and its methods, we could get the insight of using other training methodologies. We proposed few-shot learning method as the solution for generalization issues and discussed FER on FSL with two scenarios: generalization on novel data and domain adaptation.

Chapter 3: This chapter provides a detailed description of the proposed module and data construction. First, we introduce a two-stage framework: feature embedding and similarity learning. Second, we propose a channel selection module that we designed to enhance generalization performance. Then, we demonstrate a specific training process of the whole pipeline with lip and eye patches included as support and query set.

Chapter 4: All experimental results for the proposed algorithm are reported in this chapter. In particular, data construction and proposed method for training CNN are specified. The performance of the proposed model is presented by datasets. Then some discussions are made here based on the result analysis.

Chapter 5: Finally, the thesis closes with a conclusion and a preview of future work in this chapter.

Chapter II

Preliminaries

This chapter provides a brief background of the FER and FER's issues on deep neural network.

2.1. Overview of FER

FER systems have been deeply explored in the computer vision field and have been applied in various applications. According to Ekman and Friesen [21], the facial expression is a universal signal across the world. They defined the 6 basic emotions: anger, anger, disgust, fear, happiness, sadness, and surprise. Neutral emotion was added and total of 7 emotion labels have been utilized for classification tasks. Subsequently, contempt was added as one of the basic emotions [22]. However, those 6 basic emotions were dominant [23] in FER field. Many studies also have been conducted on compound expressions recognition with the idea that human emotions can not be justified as 6 basic emotions [24, 25, 26, 27, 28].

The three main modules in the FER system are as follows: face detection, feature extraction, and classification. Face detectors such as MTCNN [29], Dlib [30], Retinaface [31], and FAN [32] have been applied to detect and align the faces. After that, facial expression features are captured by a feature extractor and these features are classified into specific categories by classifier. Traditionally, texture on the image and shape of the face played a significant role in FER. Previous work, HOG [33], Gabor Wavelet [34], LBP [35], LTP [36], and NMF [37] were mainly researched. These methods have been usually experimented with under lab-controlled dataset such as CK+, MMI [38], Oulu-CASIA [39], CFEE [40], and other constrained dataset [2, 4]. Extracted features from those algorithms were classified by Support Vector Machine (SVM) [41], Adaptive Boosting [42], and other classification models. Especially, the combination of LPB which extract fine features regardless of the image's overall brightness, and SVM which shows high classification performance from the extracted features have been widely researched.

Large-scale facial datasets such as RAF-DB, AffectNet, and EmotioNet [43] were gathered from the web as the EmotiW challenge started.

However, training on large-scale unconstrained dataset degrades the performance. To address this problem, convolution neural network (CNN) -based models have been employed for FER tasks. It is because CNN [7] can extract deeper and more spatial inductive biases information. Moreover, other approaches like network ensemble, cascade network, and GAN-based models are also widely researched in FER tasks.

As shown in Figure 2.1, the recent FER process has three stages: Pre-processing, Feature learning, and Classification. In the pre-processing stage, face alignment, data augmentation, and normalization are proposed. Some models use alignment and augmentation without normalization depending on their necessity. In the feature learning stage, various deep learning-based models like CNN, DNN, and RNN have been applied to extract effective features without losing prior information. In classification, the network predicts extracted features into one of the emotion categories. In recent work, however, feature extraction and feature classification have been separated. This FER pipeline can do end-to-end learning from pre-processing to classification. Due to this process, regularizing learning is possible by the loss function.

Many FER tasks have presented promising results. However, still, there are limitations. FER with deep learning networks often falls into the overfitting problem. Although facial images are trained well on the training stage, their accuracy drops on test images. One way to solve this issue is by applying a pre-trained model trained on ImageNet or

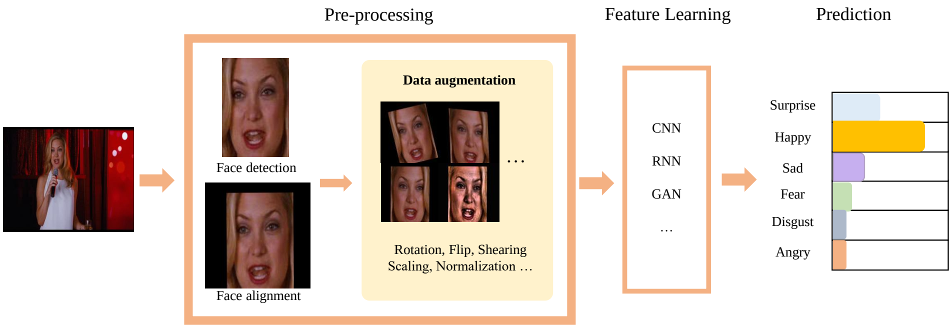


Fig. 2.1: The general pipeline of Facial Expression Recognition task.

other large-scale datasets. Images like ImageNet are coarse-grained images that contain categories: dogs, cats, persons, and other objects. On the other hand, images such as facial expressions images are fine-grained images. Examples of fine-grained images are species of dogs and species of flowers. However, fine-grained images (facial expression images) have smaller variances than coarse-grained images (ImageNet). As applying pre-trained models may not be appropriate, utilizing available datasets is needed. Therefore, few-shot learning can play a key role in this situation.

2.1.1. FER issues related to methods and datasets

Extracting information from a small area of an image and classifying its emotion is challenging. Table 2.1 illustrates various approaches that have been proposed to address this problem. Researchers usually studied attention mechanisms to maximize the learning effect [44, 45, 46]. Also, augmentation strategy and ensemble learning has been proposed [47, 48, 49, 50].

On the other hand, training facial expression datasets are challenging. Many types of research have been studied focusing on FER datasets. Facial expression datasets have issues including class imbalance, tiny variation among classes, occlusion, etc. Especially, in-the-wild datasets have differences in illumination, occlusion, and pose. As shown in Table 2.2, we discuss the recent studies dealing with the facial data problem and enhancing performance in FER tasks.

2.2. Overview of FER with FSL

As the solutions to FER problems aligned with FSL scenarios, Wang *et al.* and Psaroudakis *et al.* proposed that generalization on a few datasets is required in future FER tasks [62, 47]. Therefore, the FSL technique can be reasonably applied for FER tasks.

In this section, we discuss an overview of FER systems using FSL with respect to generalization on novel data and domain adaptation. Research that targets generalizing novel data usually focuses on inferencing unseen class or unseen face images. On the other hand, research that targets domain adaptation focuses on generalizing compound expression recognition by training on basic emotions and inference on compound emotions.

TABLE 2.1: Various issues existing in facial datasets

Method	**Issues**	**Reference**
Augment strategy	overfitting problem on unconstrained dataset	[47]
Ensemble learning	high computational cost and redundancy	[51]
Attention Mechanism	difficult to capture global information fromfacial images	[52], [48], [49], [50]
	degradation of performance owing to low-level feature extraction	[46]
	FER-related attention method is required	[44]
	difficulty in learning from video frames	[45]
Loss function	similar variation in facial expressions	[53]
	requires robust optimization	[54]
Other approaches	uncerntainty learning problem	[55], [56], [57]

Table 2.2: Various issues related to methods

	dataset (size, type, class)	**issues**	**ref**
image	FER+ (0.03M, image, 7 basic)	class imabalance	[58]
		class imbalance with vast variations on intra-class features, rotations, and occlusions	[51]
	RAF-DB (0.03M, image, 7 basic + 11 compound)	large variance in feature embedding and data uncertainty due to ambiguous annotation	[59]
		occlusion and pose variation	[60]
	AffectNet (0.45M, image, 8 basic)	input image size variation	[61]
video	AFEW (957, image, 7 basic)	few examples	[44]
		important frame is not fixed among frames	[45]

2.2.1. Generalization on novel data

FER tasks using FSL generally focused on two techniques: training with few examples and generalizing on unseen datasets. Early FER tasks had suffered from the lack of examples until the appearance of facial datasets obtained from the Internet. Therefore, utilizing as few examples as possible for training and achieving reasonable performance were the main issues [63]. Another approach to generalizing new faces has been studied. One problem is that there are insufficient examples of a specific person to model his or her emotion. To summarize, people have different faces with different expressions; satisfactorily generalizing faces is difficult if the model encounters a new face that has not been previously trained. This study also indicated the problem of insufficient examples. To address these issues, Cruz et al. proposed a model that matches a face video to references of emotion without requiring fine registration [9]. Shome *et al.* suggested research that focused on the generalization problem to deploy a system in a real-world environment [10]. They proposed few-shot federated learning for FER (FedAffect), thus tackling the problem of generalization on unseen data. Zhu *et al.* insisted that class imbalance and great intra-class variation in facial datasets cause poor performance in FER [14]. Certain emotion categories such as fear, disgust, and anger cannot meet the training needs of deep learning models owing to their scarcity when compared to other basic emotions such as surprise, happiness, and sadness. Zhu *et al.* considered the emotion classes with sufficient training samples as the training set, and the emotion classes with limited samples as the testing set [14]. To address the problem of intra-class variation, they proposed a convolutional relation network (CRN) that included emotion similarity learning with salient discriminative feature learning. The implementation setting is similar to that of k-shot n-way learning, where support and query sets are exploited as input for the relation network. Then, the feature extracted from the relation network is calculated with depth attention pooling and Jensen-Shannon (JS) divergence. The depth attention pooling and query vectors are concatenated, and then multiplied with the relation score. With this architecture, the intra-class distance is reduced and the inter-class distance is penalized.

2.2.2. Domain adaptation

Ciubotaru *et al.* explored the generalization ability of few-shot classification algorithms on recognizing unseen categories with limited training examples [11]. Starting from this review study, many approaches for generalizing domain shift problems in FER have been studied, especially compound emotion recognition. Zou *et al.* proposed emotion guided similarity network (EGS-Net) for compound emotion recognition with joint learning [12]. Through joint learning, it prevents the model from overfitting to highly overlapped sampled base classes. Finally, an alternate learning stage is set to further improve the inference ability of the model for generalizing to the unseen task. Zou *et al.* used basic emotion datasets on training for compound expression recognition [12]. Dai *et al.* proposed cross-domain few-shot learning for micro-expression recognition. They used compound emotion datasets on both training and testing [13]. Typically, cross-domain few-shot learning focuses on the data scarcity problem of the new task. Two methods (fine-tuning and metric-based few-shot learning method) are adopted to enable the model to acquire knowledge from datasets available in other scenarios (source domain), and then transfer the knowledge to the scenario where it works (target domain), recognizing novel classes with only a few labeled samples.

However, Zou *et al.* proposed a novel cascaded decomposition network (CDNet), which was trained to obtain transferable expression feature region with cascaded learn-to-decompose (LD) modules with a shared parameter [64]. With this approach, Zou *et al.* pointed out the burden of collecting large-scale labeled compound expression data and solved the limitations of CRN and EGS-Net.

Chapter III

Proposed Method

3.1. Data Preprocessing

In this section, we introduce all the data preprocessing algorithms. We utilized RAF-DB [65] , FER2013 [66], SFEW [67] and AFEW [68] for the experiments. RAF-DB, FER2013, and SFEW provide aligned facial images that can directly use as training data without data preprocessing. On the other hand, AFEW is a movie-based video dataset and provides only grey-scale face images as preprocessed data. Therefore, we did the whole data preprocessing for AFEW dataset.

The overview of the process is illustrated in Figure 3.1. First, frames are extracted from the video according to the frames per second (fps). After that, we detect faces from the frame using multi-task cascaded

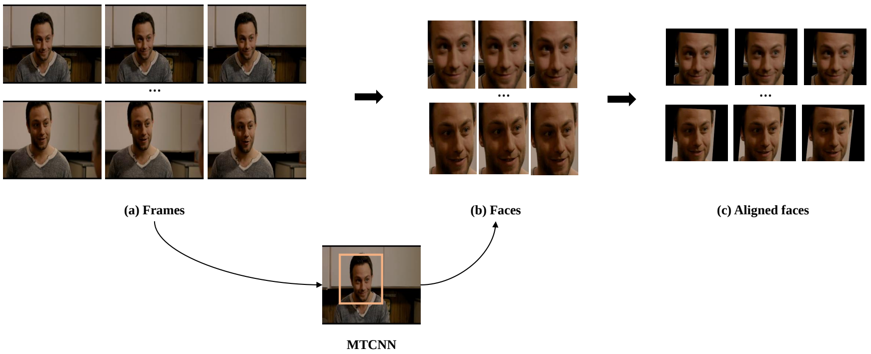


Fig. 3.1: A process of getting aligned faces from video dataset.

CNN (MTCNN) [29] and crop facial regions with 84 × 84 size. Although the MTCNN is slow compared to other face detection models, we chose it since the MTCNN provides various calculation information that is useful for aligning slanted faces.

Inspired by MAXDIST peak frame selection method [1], we selected peak emotion frames among all the frames. Figure 3.2 gives an overview of the peak frame selection process and the algorithm is based on calculating differences between frame features. We calculated and compared all the features to figure out which frame has the largest difference. The selection process is illustrated in Figure 3.3 and we chose the highest peak frame.

To train with spatial information, we generate patch-wise images from given facial images. We crop the eye area and lip area of each aligned face image. In our study, we employ the RetinaFacePredictor [31] to detect faces in an image, and based on this information, the FAN Predictor generates landmarks on the key points of the detected faces [32]. For

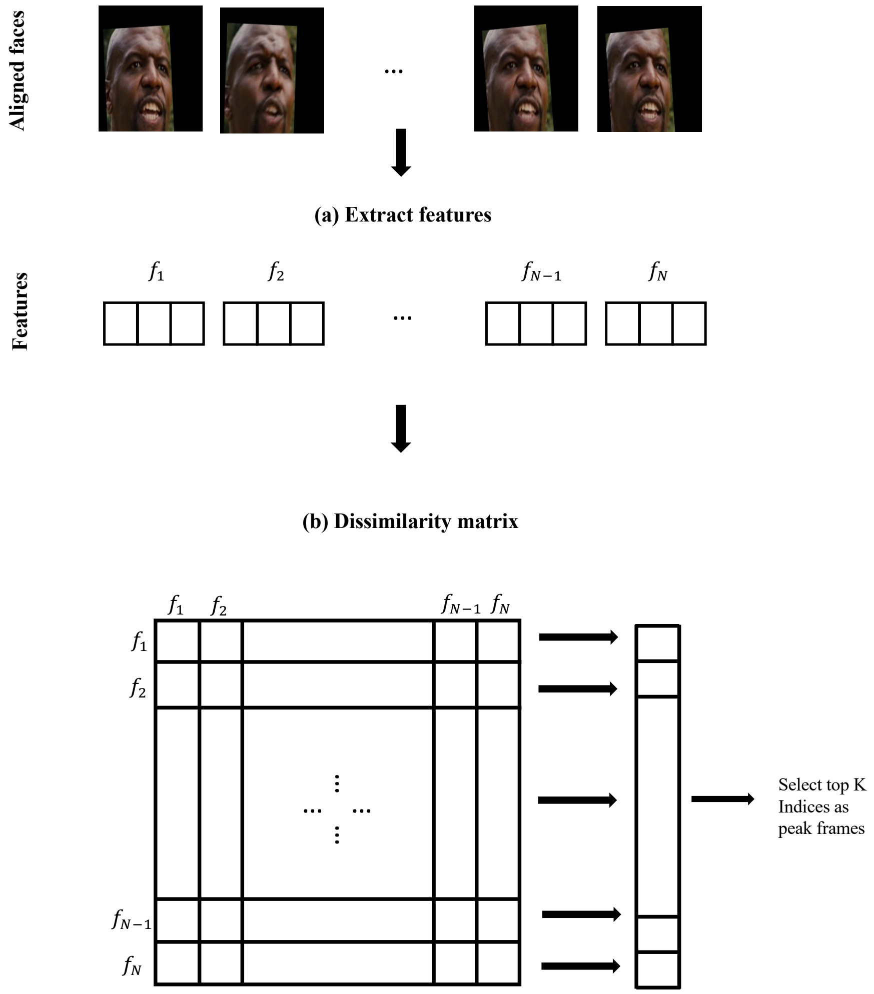


Fig. 3.2: Overview of MAXDIST peak frame selection method [1].

the RetinaFacePredictor, we utilize the model trained on resnet50. Additionally, we utilize the publicly available 2dfan2 alt weights of FAN, as released by Bulat *et al.* [32].

The facial landmark is composed of 68 dots and each dot has a position number. We crop the eye using 46, 20, 30, and 37 in 84 × 84 images as shown in Figure 3.4. We denote each landmark number as x_1, x_2, y_1, and y_2 point in the face image. First, we get the center point from the width and height of a given image. We calculate the length and height of the eye area by subtracting each eye edge's landmark point. Finally, we set the cropping point by subtracting the center point with calculated eye width and height. We denote the center point as $cx = width/2$, $cy = height/2$. The final cropping x point is $x_1 - \frac{cx-(x_2-x_1)}{2}$ and $x_2 + \frac{cx-(x_2-x_1)}{2}$, and the final cropping y point is $y_1 - \frac{cy-(y_2-y_1)}{2}$ and $y_2 - \frac{cy-(y_2-y_1)}{2}$.

While training, images are normalized, flipped, rotated, and resized to 64 × 64.

3.2. Network architecture

Figure 3.5 shows the overall pipeline of the proposed model, which comprises a two-stage framework based on CNN with channel selection module. In the first stage, features extracted from given input images and the emotion similarities are calculated in the emotion similarity learning in the second stage. We denote images in the support set $S = \{s_k^{(i)}\}_{k=1}^{\frac{N_{train}}{2}}$

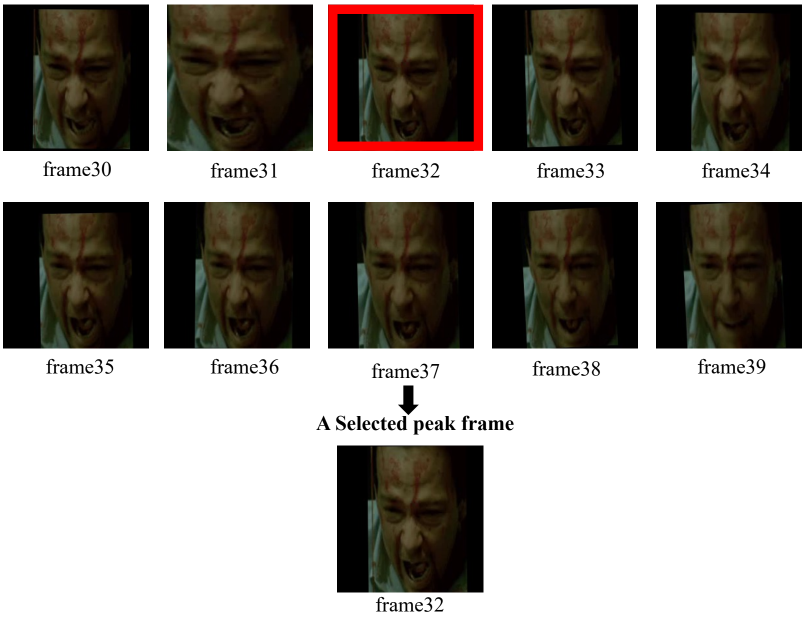


Fig. 3.3: An example of selecting a peak frame.

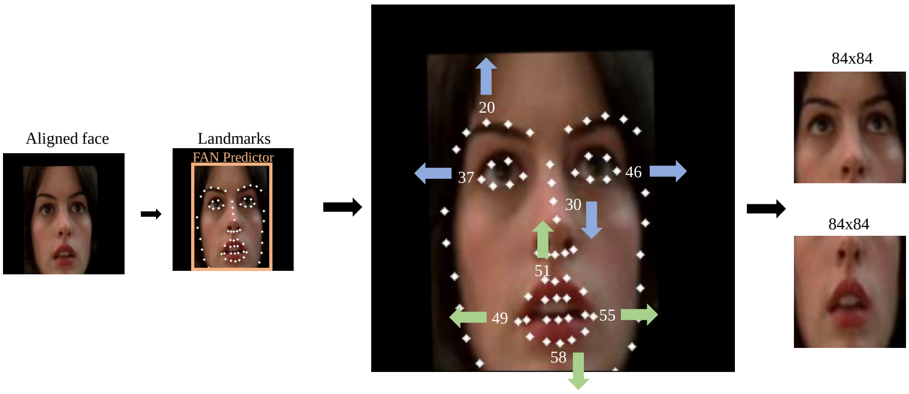


Fig. 3.4: A process of cropping eye and lip region by landmark information.

and $Q = \{q_k^{(j)}\}_{k=\frac{N_{train}}{2}}^{N_{train}}$, where i and j represents class and k represents k_{th} image among total samples. The N_{train} represents total number of images for training. The reason why we divide total number by 2 is to set support and query set ratio 1:1 during training.

The embedding module extracts feature from support and query sets. We set face, eye, and lip as input and extract each support and query feature. The paired support and query features are fed into the relation module to learn emotion similarity.

We propose Channel Selection (CS) module to train the best from support feature information. CS generates the optimal feature from face support features. The generated feature is paired with the face query feature. Therefore, the paired face feature, paired eye feature, paired lip feature, and the paired reconstructed feature are fed into the relation module together.

The training process of the whole pipeline is described in subsections 3.2.1, 3.2.2, and 3.2.3.

3.2.1. Feature embedding

In this stage, the samples $s^{(i)}$ in the support set S and $q^{(j)}$ in the query set Q are fed into the feature embedding module f_θ to generate feature maps $f_\theta(s^{(i)})$ and $f_\theta(q^{(j)})$. Inspired by the architecture setting in relation network [20], we apply the four convolutional blocks for the

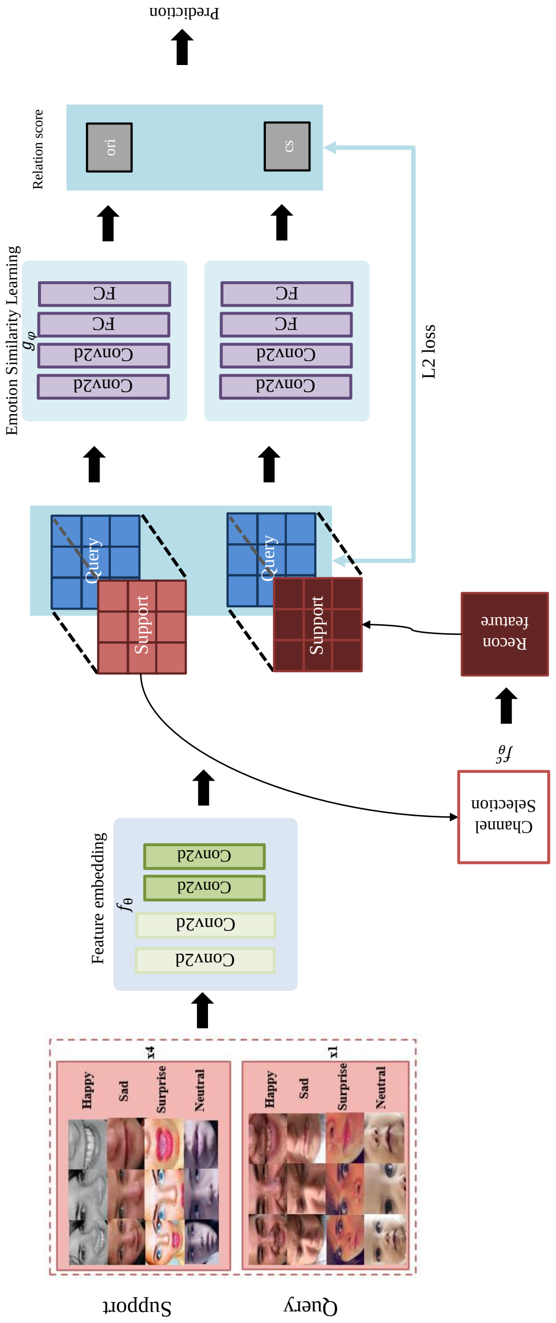


Fig. 3.5: The overview architecture of the proposed model.

feature extraction module. Each convolution layer with batch normalization, ReLu activation function with Maxpooling. With these simple convolutional layers, minimum features are extracted and its training is efficient, which is reasonable for the few-shot learning method.

3.2.2. Channel Selection

Channel selection is a module that generates the optimal feature for similarity learning. It functions after feature embedding and compares each feature $f_\theta(S)$ extracted from the feature embedding module f_θ. The overview of the CS process is shown in Figure 3.6. At first, one 14×14×64 average feature S_{avg} is obtained through the depth average pooling (DAP) of the incoming 14×14×64 support features $S_1 = f_\theta(s_{fa_1}) \sim S_i = f_\theta(s_{fa_i})$. The DAP feature is expressed as follows:

$$S_{avg} = \frac{1}{N}\sum_{i=1}^{N} S_i, \tag{3.1}$$

where N is a number of sample features of an emotion class. After that, pool 14×14×64 support features and the average feature into 1×1×64. The pooled feature of i^{th} sample feature is defined as:

$$a_i = AP(S_i), \tag{3.2}$$

where AP means average pooling. The pooled feature of S_{avg} is also defined as:

$$a_{avg} = AP(S_{avg}). \tag{3.3}$$

In this way, the features have their representative information in every 64 channels. We choose pooled 1×1 instead of 14×14 to compare channel-wise information while preventing comparing ambiguous information. The average feature consists of a representative value of each feature. Therefore, comparing the distribution of features and the average feature would include vague information which fails to select optimal channel features. To prevent this, we set representative values (1×1) to compare and generate the best channel features by selecting the most similar channel. Each 1×1 channel of sample feature and average feature are defined as:

$$a_{i\_c}, a_{avg\_c}, \tag{3.4}$$

where c represents c^{th} channel and i is i^{th} sample. The channel (14×14×1) is taken from the 14×14×64 feature whose 1×1 information is most similar to the 1×1 average feature. The channel selection module is defined as:

$$R = f_{\theta}^{c}(f_{\theta}(s_{fa})), \tag{3.5}$$

where s_{fa} represents sample of facial image and f_{θ} is feature embedding module that $f_{\theta}(s_{fa})$ is defined as extracted feature of face image.

The whole process of channel selection is illustrated in Algorithm 1.

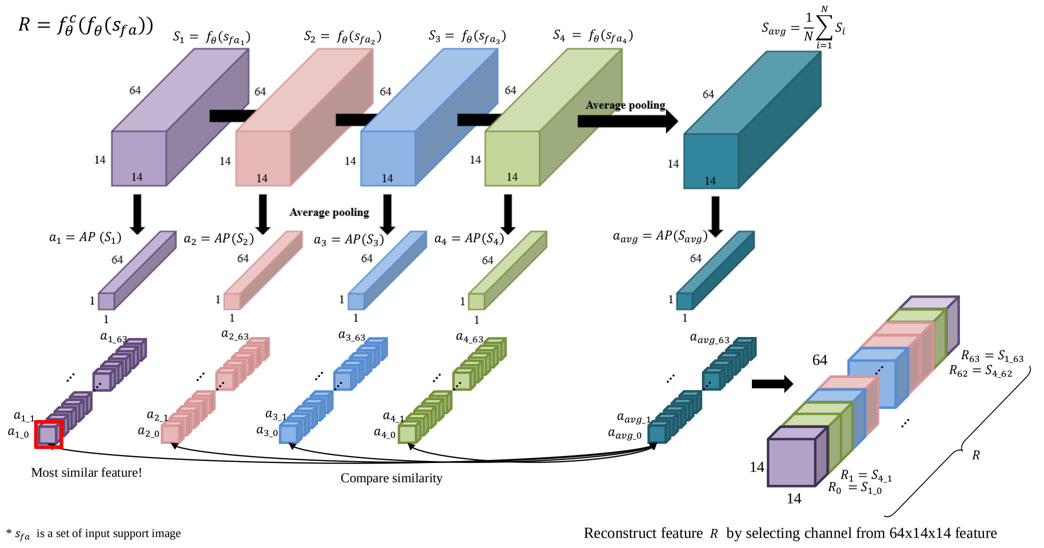


Fig. 3.6: The overview of the Channel Selection process.

Algorithm 1 Channel Selection (f_θ^c)

```
$S_{avg} = DAP(S_1, S_2, \dots, S_i)$
Set $a_{avg} = AP(S_{avg})$
for $i = 1, 2, \dots, I$ do
    Set $a_i = AP(S_i)$
end for
for $c = 0, 1, \dots, C-1$ do
    Set $idx = 1$
    Set $min = \mathrm{abs}(\mathrm{sub}(a_{1\_c}, a_{avg\_c}))$
    for $i = 2, 3, \dots, I$ do
        if $sub(a_{i\_c}, a_{avg\_c}) \leq min$ then
            $min \leftarrow \mathrm{abs}(\mathrm{sub}(a_{i\_c}, a_{avg\_c}))$
            $idx \leftarrow i$
        end if
    end for
    $R_c \leftarrow S_{idx\_c}$
    if $n \geq 1$ then
        $R_c = [R_{c-1}, R_c]$
    end if
end for
$R = R_{C-1}$
```

According to the designed algorithm, the $sub(A, B)$ represents subtracting B from A. To calculate the difference, we took the absolute value of the subtracted result which was illustrated as $abs()$. The equation from lines 3 to 14 can be also defined as follows:

$$min \leftarrow \underset{x}{\operatorname{argmin}} f(x) = \{x|f(x) = \mathrm{abs}(\mathrm{sub}(A, B))\}, \tag{3.6}$$

which can be illustrated without for loop. At the same time, C represents the number of channels and I represents a number of sample features. The reconstructed feature by selecting channels from 14×14×64 is R.

Therefore, the optimal 14×14×64 feature is made. In this way, we could generate the optimal feature while minimizing information loss.

3.2.3. Emotion similarity learning

The extracted query and support feature are concatenated together and fed into the relation module is defined as:

$$g_{\varphi}(C(f_{\theta}(s^{(i)}), f_{\theta}(q^{(j)}))), \tag{3.7}$$

where g_{φ} represents the relation module. This module follows the same architecture setting that Sung *et al.* proposed [20]. It is composed of two convolutional blocks and two fully-connected (FC) layers. The two convolutional blocks have batch normalization, ReLU activation function, and max pooling.

The final FC layer with a sigmoid activation function generates relation scores, which represent the similarity between samples in the support and query set. The mean square error (MSE) loss is computed as:

$$MSE = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (s^{(i)} - q^{(j)})^2, \tag{3.8}$$

where N represents class number while $s^{(i)}$ and $q^{(j)}$ represents i_{th} class support set and j_{th} class query set. Our model can classify unlabeled facial expression images into new classes based on the learned similarity.

3.3. Overall training process

This section introduces a new form of few-shot learning process. We aim to learn the similarity using spatial information to enhance the generalization power to predict unseen labels. We proposed to utilize lip and eye patch images for support and query to learn the similarity so that it can focus on important facial regions.

Figure 3.7 shows a more specific training process with face-lip-eye data construction. Each facial image's eye and lip patches are combined together with the original input and set as a support and query. We denote the support set and the query set of each class as $S^{(i)} = s_{fa}^{(i)} + s_e^{(i)} + s_l^{(i)}$ and $Q^{(j)} = q_{fa}^{(j)} + q_e^{(j)} + q_l^{(j)}$. In the feature embedding stage, it generates features $f_\theta(S)$ and $f_\theta(Q)$. We mapped each paired patch to feed into the emotion module to learn the similarity of both spatial and global information. The paired feature map is defined as $C(f_\theta(s^{(i)}_{fa}), f_\theta(q^{(j)}_{fa}))$, $C(f_\theta(s^{(i)}_{e}), f_\theta(q^{(j)}_{e}))$, $C(f_\theta(s^{(i)}_{l}), f_\theta(q^{(j)}_{l})$, and $C(f^c_\theta(f_\theta(s^{(i)}_{fa})), f_\theta(q^{(j)}_{fa}))$.

Each features expand dimension to each other's batch size to see all the combinations of support and query images. Therefore, the relation network can calculate all the relations between the face, eye, lip, and generated features. The relation scores are generated from each concatenated feature map. We denote the relation score as $r^{(i,j)}$ and each score is calculated with each query feature regressively. The loss L_r in training for emotion similarity is defined to be an MSE classification loss as the following:

$$L_r = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}[y^{(i,j)} - r^{(i,j)}(g_\varphi[C(f_\theta(s^{(i)}), f_\theta(q^{(j)}))])]^2. \tag{3.9}$$

The relation score $r^{(i,j)}$ is updated by comparing it with the ground truth. The three relation scores are: $r^{(i,j)}_{fa}$, $r^{(i,j)}_{e}$, $r^{(i,j)}_{e}$, and $r^{(i,j)}_{c}$ and these are added together for joint learning. The final loss function is defined as:

$$\begin{aligned} L = {} & \lambda_f L_2(r^{(i,j)}_{fa}, f_\theta(q^{(j)}_{fa})) + \lambda_e L_2(r^{(i,j)}_{e}, f_\theta(q^{(j)}_{e})) \\ & + \lambda_l L_2(r^{(i,j)}_{l}, f_\theta(q^{(j)}_{l})) + \lambda_c L_2(r^{(i,j)}_{c}, f_\theta(q^{(j)}_{fa})). \end{aligned} \tag{3.10}$$

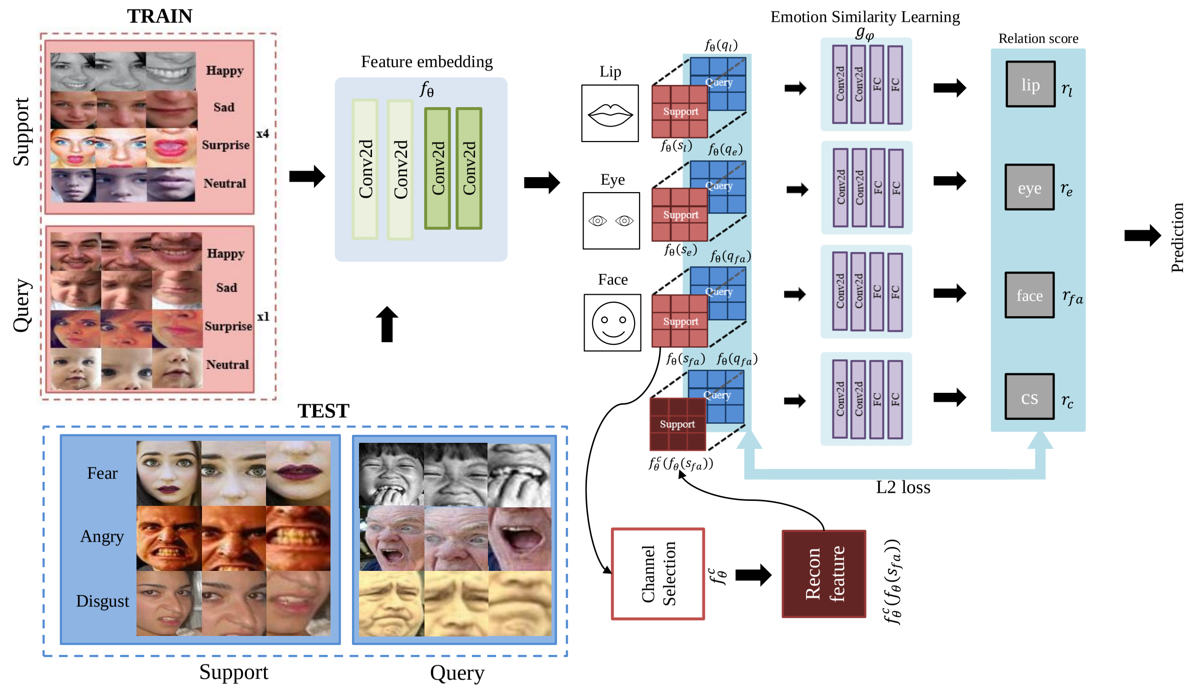


Fig. 3.7: The whole pipeline of Channel Selective Spatial Relation Network.

Chapter IV

Experimental Results and Discussion

4.1. Dataset

In this section, we introduce datasets and explain dataset construction. The data division and construction are different from the supervised learning, to see the generalization capability of untrained classes.

4.1.1. Dataset division

All experiments are conducted on RAF-DB, FER2013, SFEW, and AFEW. The details of the sample distribution among the whole datasets are shown in table 4.1. The distribution of each class in train and test samples are imbalanced. To alleviate training in imbalanced conditions, we select classes with sufficient samples as training sets, and the rest of the categories with few samples as the testing set.

We selected 4 emotion categories for training (i.e. Happy, Sad, Surprise, and Neutral) and 3 emotion categories for testing (i.e. Angry, Disgust, and Fear) for RAF-DB, SFEW, and AFEW. On the other hand, 4 emotion categories for training (i.e. Angry, Fear, Happy, and Sad) and 2 emotion categories for testing (i.e. Disgust and Surprise) are selected for FER2013 dataset. A total of 11,239 images were trained and a total of 396 images were tested for RAF-DB, and a total of 20,137 images were trained and 471 images were tested for FER2013. SFEW and AFEW both have much fewer samples and the number of sample gap between the test and train set is smaller than other datasets. A total of 195 images were trained and 130 images were tested for SFEW, while a total of 451 were used in training and 142 were tested for AFEW.

Table 4.1: The number of images of each class for experiments.

		Happy	Sad	Surprise	Neutral	Angry	Disgust	Fear	Total
RAF-DB	Train	4772	1982	1290	3195	-	-	-	11239
	Test	-	-	-	-	162	160	74	396
FER2013	Train	7215	4830	-	-	3995	-	4097	20137
	Test	-	-	415	-	-	56	-	471
SFEW	Train	46	40	31	78	-	-	-	195
	Test	-	-	-	-	47	43	40	130
AFEW	Train	146	105	66	134	-	-	-	451
	Test	-	-	-	-	61	39	42	142

4.1.2. Dataset Construction

To learn spatial information of facial images more effectively, we included patch images of lip and eyes as support and query set during training. The input images are aligned, normalized, and resized in 64 × 64. In the training process, we divide each dataset into the support set and query set in a ratio of 1:1. According to Snell *et al.* and Vinyals *et al.* an effective training strategy in few-shot learning is to exploit the training set via episode (mini-batch) [18, 19]. Also we trained a specific number of samples (nsample) on each dataset and the samples were randomly selected for an episode, to compare our results with CRN [14]. Zhu *et al.* set their nsample as 80, 150, and 60 for RAF-DB, FER2013, and SFEW [14]. To compare with the CRN performance, we allocated the nsample for RAF-DB, FER2013, SFEW, and AFEW as 25, 50, 20, and 20, respectively as we have additional lip and eyes patch samples.

Table 4.2 shows the nsample construction for one episodic training comparing with CRN [14]. The total $(2 \times nsamples \times \mathrm{C}_{train}) \times 3$ samples were used during training. Therefore, $(2 \times nsamples \times \mathrm{C}_{train}) \times 3$ samples were set in one training episode/mini-batch for FER task. The number '2' represents 1:1 support and query, C_{train} is the number of classes. The number '3' represents three input samples: face, eye, and lip images.

Furthermore, in order to compare our model with other models fairly and conform to the few-shot learning setting, we have guaranteed the value of nsamples is consistent in whole experiments on the same dataset.

Table 4.2: Value of nsample in one training episode.

	dataset	nsample	(bh, n-way, k-shot)	support samples	query samples	Total samples
CRN*	RAFDB	80	(80, 4, 4)	1280	1280	2560
	FER2013	150	(150, 4, 4)	2400	2400	4800
	SFEW	60	(60, 4, 4)	960	960	1920
	dataset	nsample	(bh, n-way, k-shot)×3	support samples	query samples	Total samples
ours	RAFDB	25	(25, 4, 4) ×3	1200	1200	2400
	FER2013	50	(50, 4, 4) ×3	2400	2400	4800
	SFEW	20	(20, 4, 4) ×3	960	960	1920
	AFEW	20	(20, 4, 4) ×3	960	960	1920

4.2. Training Details

All models were trained using the PyTorch deep learning framework, and used single GPU for training. Table 4.3 shows the specifications of the experimental environment. In all experiments, we used Adam optimizer [69] with $\beta_1 = 0.9, \beta_2 = 0.99$; initial learning rate 0.001 that decays with the stepLR function. We determine each score's weight by experiments. First, we explore the proposed model training with only eye and only lip data to see if there is a performance difference. However, there was no significant difference between eye and lip data according to the test result. The experimental result is shown in Table 4.4. Therefore we set each eye, lip and face relation score weight 1:1:1.

Table 4.3: Environment condition for experiments.

ITEMS	spec
OS	Ubuntu 20.04.5 LTS
CPU	AMD Ryzen Threadripper PRO 3955WX 16-Cores
GPU	GeForce RTX 3090 (24GB)
Mem	128GB

Table 4.4: The comparison of additional lip and eye data accuracy (%).

Dataset	Method	data	AN	DI	FE	Average
RAFDB	CSSR	eye	44.27	32.76	31.06	36.03
	CSSR	lip	29.82	46.35	36.86	37.67

We took an experiment on the effect of parameter λ of CS on the performance with the value of λ in [0,1]. As shown in Figure 4.1, we

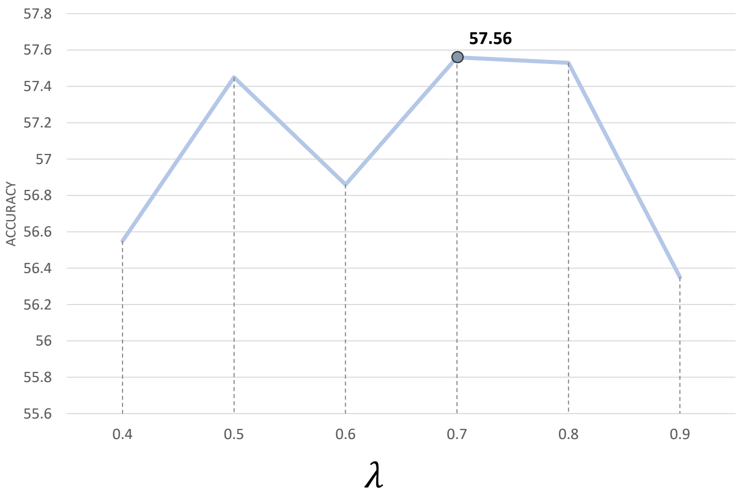


Fig. 4.1: An accuracy of the proposed method with the ratios of the chosen parameter λ of channel selection on the RAFDB.

can select 0.7 as the best parameter for CS. For episodic training, we trained each episode (mini-batch) for 10000 epochs. To prove the generalization performance, we did not use pre-trained weight for both feature embedding network and relation network.

4.3. Performance Analysis

The performance result is calculated in accuracy by computing total corrected query samples: $\frac{Q_{correct}}{Q_{total}}$. Snell *et al.* proposed that the data construction should be the same while inferencing to learn the similarity. They predicted the result with test dataset with the same episode scenario 1000 times and computed the average value [18]. Therefore, the average accuracy is computed as:

$$Accuracy = \frac{1}{N} \sum_{e=1}^{1000} \frac{Q^{e}_{correct}}{Q^{e}_{total}}. \tag{4.1}$$

where $Q^{e}_{correct}$ and Q^{e}_{total} represent the total number of corrected query samples and the total query samples of e^{th} episode, respectively. N represents the total number of episodes, 1000. We took the sum of the accuracy in every episode and divided it by the number of episodes.

All the FER accuracies are arranged in Table 4.5. The recognition accuracy is reported in % and the mean recognition rates showed in the last column. FEL represents face, eye and lip data construction and CSSR represents the proposed model Channel Selective Spatial Relation Network.

We conduct experiments with four basic emotion dataset and compare each performance in four different methods. We choose CRN [14] as the comparative reference as it is the only and most recent paper that addresses the task of generalizing to unseen emotion classes by separating the train class and test class within a single dataset. We implemented CRN ourselves and documented its performance. We also implemented relation network which is the baseline model of CRN to check its performance.

To compare, we conducted experiments to examine the performance of our proposed data composition (FEL) method in CRN. Utilizing FEL data construction on CRN (CRN+FEL) shows generalized performance on each emotion class of all the four dataset. Especially, CRN+FEL demonstrates strong generalization performance on SFEW dataset which has significant variation among emotion classes in CRN method.

Furthermore, using both our proposed CS module and data construction method (CSSR+FEL) outperforms over all the methods. Also, our proposed model outperforms without JS-Divergence in CRN [14].

The overall performance has improved, and each class's accuracy has also increased. By employing the FEL data construction, we have effectively mitigated deviations within each class while achieving performance

Table 4.5: The performance comparison of FER accuracy (%).

Dataset	Method	AN	DI	FE	SU	Average
RAF-DB	RelationNet	43.07	40.17	56.43	-	46.55
	CRN	48.14	53.89	60.16	-	54.06
	CRN +FEL	50.78	57.77	58.16	-	55.57
	CSSR +FEL	54.54	63.14	55.02	-	57.56
SFEW	RelationNet	42.53	42.45	42.46	-	42.51
	CRN	28.85	53.91	68.75	-	50.50
	CRN +FEL	39.95	62.49	56.08	-	52.84
	CSSR +FEL	36.89	64.78	66.57	-	56.08
AFEW	RelationNet	29.17	47.70	27.77	-	34.88
	CRN	21.15	37.21	43.59	-	33.98
	CRN +FEL	34.76	38.97	31.05	-	34.92
	CSSR +FEL	36.36	48.07	24.46	-	36.29
FER2013	RelationNet	-	57.18	-	63.37	60.27
	CRN	-	73.22	-	61.37	67.29
	CRN +FEL	-	71.50	-	68.18	69.84
	CSSR +FEL	-	71.67	-	70.33	71.73

improvement of +2.64%, +3.88%, and -2% for the Angry, Disgust, and Fear classes in RAFDB, respectively. Furthermore, the proposed model of CSSR+FEL, we have achieved the performance gain of +3.76%, +5.37%, and -3.14% compared to CRN+FEL in the respective category order. Finally, we have achieved the performance gain of +6.4%, +9.25%, and -5.14% compare to CRN.

The increase in performance for Angry and Disgust classes and the decrease for the Fear class can be attributed to several factors. Firstly, the Fear class is known to be the most challenging to distinguish in the RAFDB dataset [70]. Additionally, it could be a result of the generalization of the performance, where reduced deviations within each class lead to more balanced performance across all classes. The average performance for RAFDB has increased by +1.99% compare to CRN+FEL and +3.5% compare to CRN.

For the SFEW dataset, the proposed model demonstrates significant performance variations among different classes. However, through the FEL construction, we could mitigate these variations, resulting in the performance improvement of +11.1%, +8.58%, and -12.67% for the Angry, Disgust, and Fear classes, respectively.

Furthermore, when comparing CSSR+FEL with CRN+FEL, we observed the performance gains of -3.06%, +2.29%, and +10.49% in the same order. The performance decline in the Fear class when applying FEL to CRN can be considered as the significant reduction in extreme variations. On the other hand, the decrease in the performance for the Angry class in CSSR+FEL can be explained by the presence of blurred images mixed with the Angry samples in the SFEW dataset. .

Compare to CRN, CSSR+FEL have achieved the performance gain of +8.04%, +10.87%, and -2.18% in each category order. To analyze performance decrease in Fear class, the reason can be a similarity confusion between Fear and Angry images. Both the Fear and Angry classes exhibit similar characteristics, such as open mouths, which can lead to confusion. Finally, the average performance of the CSSR+FEL has increased by +3.24% compare to CRN+FEL, and gain +5.58% compare to CRN.

For the AFEW dataset, utilizing FEL data construction to CRN also alleviates performance variations. Compare CRN to CRN+FEL, there are performance gain of +13.61%, +1.76%, and -12.54%. Each emotion class performance increase and decrease to similar level. The average performance is enhanced to +0.94%.

When comparing CSSR+FEL with CRN+FEL, we observed the performance gain of +1.6%, +9.1% and decrease of -6.59% in the category order. The proposed model CSSR+FEL shows +15.21%, +10.86%, and -19.13% compare to CRN, in a similar pattern.

Focusing on the problem of continued performance decline in Fear class, we researched in dataset level. We analyzed AFEW dataset and found out Fear images consist minimal variations in facial expressions.

Capturing facial muscle movements is crucial for feature extraction in the field of FER. Therefore, the lack of significant expression differences in Fear data may hinder performance improvement. Finally, the overall performance of our proposed model has enhanced compare to CRN+FEL, with +1.37% and compare to CRN, with +2.31%.

FER2013 dataset also shows similar performance variations as seen in previous datasets. The difference between Disgust and Surprise classes was initially 11.85%, but through the application of FEL, the deviation reduced significantly to approximately 3.32%. Moreover, each emotion class showed the performance of -1.72% and +6.81%. When comparing CSSR+FEL with CRN+FEL, the CSSR+FEL model demonstrated the performance enhancement in all classes, with an overall improvement of +0.17% and +2.15% in respective order. To compare with CRN, the CSSR+FEL demonstrates performance of -1.55% and +12.36%. The performance of Disgust class decreased. However, the both Disgust and Surprise class reached similar performance level. This implies a general improvements on each emotion categories. We obtained the average performance of +1.16% compare to CRN+FEL and +3.68% compare to CRN.

By employing the CS module, we have achieved promising results without the need for DAP and JS-Divergence. This means that the proposed CS module and their inclusion in the loss function are essential for the performance improvement.

4.4. Ablation study

4.4.1. Ablation study comparing the channel information of sample feature and average feature

In this section, we analyze the limitations of DAP and the motivation behind proposing the Channel Selective module through the visualized data. In order to compare and visualize the 14x14 features for each channel, we have arranged them in Figure 4.2.

The images corresponding to the same channels are vertically aligned. In the last row, the visualization includes the S_{avg} feature, which represents the average of the channels for four sample features ($S_1 \sim S_4$).

Examining $S_1 \sim S_4$ features, it is apparent that important facial regions are highlighted in brighter colors. Also, it is enable to recognize facial shapes. However, when observing the average feature s_{avg}, it is evident that the distinct features that were prominently captured in the individual samples are diminished or blended together.

The model in CRN has limitations as it keeps losing important features and blurring the contours of the face. On the other hand, the CS overcomes this limitation by selectively utilizing features from the original channels with minimizing the loss. This allows for the preservation of facial contour features while avoiding any degradation of important details.

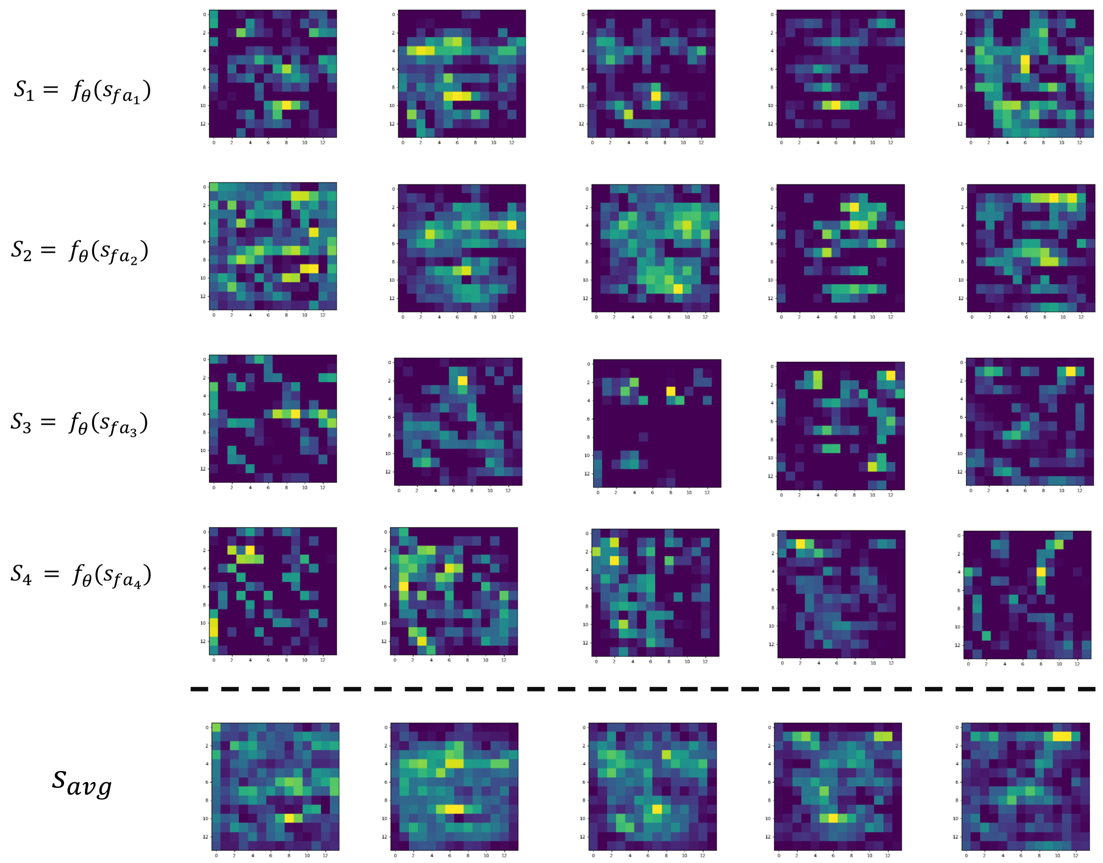


Fig. 4.2: Visual comparison between the sample channel features and averaged channel features.

4.4.2. Ablation study on the impact of feature size in CS module

The CS module compares each individual sample feature with the feature obtained by averaging all the sample features. All the features are subjected to 1×1 average pooling before the comparison. The reason for performing 1×1 average pooling is as follows.

The averaged feature contains a general representation of the sample features. On the other hand, individual features have different importance variations. In this way, comparing the similarity of their distribution over a wide area may include less significant details. We find that selecting the optimal channel is challenging when comparing larger distribution areas and propose using 1×1 average pooling.

To prove our hypothesis, we experimented and compared the performance of 1×1 pooling with 14×14 pooling. The performance result is shown in Table 4.6. 1×1 pooled channel feature is more promising than

Table 4.6: The performance comparison based on the CS module's channel size (%).

Dataset	Method	Avg Acc (%)
RAFDB	CRN	54.06
	CRN + FEL	55.57
	CSSR + FEL + CS (14×14)	56.89
	CSSR + FEL + CS (1×1)	57.56

14×14 in terms of performance. We observed the performance improvement of 0.67% by using 1×1 average pooling.

4.4.3. Ablation study on fine-tuning weights on individual modules

According to the results of the experiments above in Table 4.4, we figure out that the individual face, eye, lip, and CS module have similar impact strengths. We found that adjusting the cs module would have an impact on the overall performance. Therefore, we conducted all experiments with a weight ratio of 1:1:1:0.7. In this section, however, we conduct additional experiments by fine-tuning the lambda values for each module while varying their proportions to investigate the most effective combination of weights.

We investigate the influence of the four modules (face, eye, lip, and CS) by gradually increasing or decreasing their lambda values from 0.1 to 0.4. This allows us to understand the impact of each module. Additionally, we conduct experiments using the same weight ratio of 0.25 for all four modules to see which combination would show progress.

The experiments were conducted using the RAF-DB. The average results represent the performance obtained by training the model on the happy, sad, surprise, and neutral emotion classes and then testing it on the angry, disgust, and, fear emotion classes. According to the results presented in Table 4.7, it is evident that the same weight of each module

Table 4.7: The comparison of accuracy (%) in different weights on individual modules.

Dataset	Method	Weights on each module				Average accuracy(%)
		face	eye	lip	CS	
RAFDB	CSSR	0.1	0.2	0.3	0.4	55.82
		0.4	0.3	0.2	0.1	56.37
		0.25	0.25	0.25	0.25	57.15

yielded the best performance of 57.15%.

We found that when the face module was given the highest weight, the performance reached 56.37%. Therefore, we conduct further experiments by gradually increasing the weight on the face module and decreasing the weight on the rest of the modules to explore if it would lead to even better performance. However, the average performance decreased.

Also, we conduct additional experiments in an opposite manner. The result is shown in Table 4.8. In this experiment, although the performance gradually improved, it could not surpass the best performance of 57.56%.

Table 4.8: The comparison accuracy (%) of various combinations of weights on each individual module.

Dataset	Method	Weights on each module				Average accuracy(%)
		face	eye	lip	CS	
RAFDB	CSSR	0.5	0.3	0.1	0.1	54.39
		0.7	0.1	0.1	0.1	55.02
		0.1	0.1	0.3	0.5	52.13
		0.1	0.1	0.1	0.7	57.53

Therefore, considering different combinations of the modules is not useful since each module has a similar significance. Finally, maintaining the same weight ratio of the face, eye, and lip modules while fine-tuning the weight of the CS module resulted in the best performance.

CHAPTER V

Conclusion

In this thesis, we proposed a Channel Selection module and additional spatial data construction. To effectively exploits the best from a few datasets, we set an averaged feature of sample features as a representative feature. The representative feature of each channel was compared with each channel information of sample features to find the most similar channel information. By comparing channel information, the channel from a selected sample is extracted as an optimal channel of the corresponding sample feature. Therefore, one reconstructed feature is composed of each sample's channel information by the designed module.

To maximize the spatial information of facial images, we generated eyes and lip image patches and set as a data construction. This proposed data construction demonstrated the significance of learning the relation between partial information while training few-shot FER. Furthermore, utilizing spatial information enhanced the generalization capability.

From the experiments, we have proved the selected optimal feature and additional spatial information can achieve the generalization performance. The average performances of proposed data construction on RAFDB, FER2013, SFEW, and AFEW are increased by 1.51%, 2.55%, 2.34%, and 0.95%. Moreover, applying CS module with data construction, the performance has been improved by 3.5%, 3.68%, 5.58%, and 2.31% of accuracy, respectively when compared to the CRN [14].

As future work, a method to train and test on seven emotion categories to see generalization on few-shot FER tasks is needed. Furthermore, various research on obtaining optimal information from facial datasets is needed.

References

[1] S. Zhalehpour, Z. Akhtar, and C. Eroglu Erdem, “Multimodal emotion recognition based on peak frame selection from video,” *Signal, Image and Video Processing*, vol. 10, pp. 827–834, 2016.

[2] J. Kim, B. Kim, P. Roy, and D. Jeong, “Efficient facial expression recognition algorithm based on hierarchical deep neural network structure,” *IEEE access*, vol. 7, pp. 41273–41285, 2019.

[3] D. Jeong, B. Kim, and S. Dong, “Deep joint spatiotemporal network (djstn) for efficient facial expression recognition,” *Sensors*, vol. 20, no. 7, p. 1936, 2020.

[4] S. Park, B. Kim, and N. Chilamkurti, “A robust facial expression recognition algorithm based on multi-rate feature fusion scheme,” *Sensors*, vol. 21, no. 21, p. 6954, 2021.

[5] Y. Heo, W. Yeo, and B. Kim, "Deepfake detection algorithm based on improved vision transformer," *Applied Intelligence*, vol. 53, no. 7, pp. 7512–7527, 2023.

[6] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Feifei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

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

[8] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[9] A. C. Cruz, B. Bhanu, and N. S. Thakoor, "One shot emotion scores for facial emotion recognition," in *Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP)*, pp. 1376–1380, 2014.

[10] D. Shome and T. Kar, "Fedaffect: Few-shot federated learning for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4168–4175, 2021.

[11] A.-N. Ciubotaru, A. Devos, B. Bozorgtabar, J.-P. Thiran, and M. Gabrani, "Revisiting few-shot learning for facial expression recognition," *arXiv preprint arXiv:1912.02751*, 2019.

[12] X. Zou, Y. Yan, J. Xue, S. Chen, and H. Wang, "When facial expression recognition meets few-shot learning: A joint and alternate learning framework," *arXiv preprint arXiv:2201.06781*, 2022.

[13] Y. Dai and L. Feng, "Cross-domain few-shot micro-expression recognition incorporating action units," *IEEE Access*, vol. 9, pp. 142071–142083, 2021.

[14] Q. Zhu, Q. Mao, H. Jia, O. E. N. Noi, and J. Tu, "Convolutional relation network for facial expression recognition in the wild with few-shot learning," *Expert Systems with Applications*, vol. 189, p. 116046, 2022.

[15] C.-L. Kim and B.-G. Kim, "Few-shot learning for facial expression recognition: a comprehensive survey," *Journal of Real-Time Image Processing*, vol. 20, no. 3, p. 52, 2023.

[16] L. Fefei, "A bayesian approach to unsupervised one-shot learning of object categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1134–1141, 2003.

[17] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proceedings of the 2009 IEEE conference on computer vision and pattern recognition*, pp. 951–958, 2009.

[18] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, vol. 30, pp. 4080–4090, 2017.

[19] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, vol. 29, pp. 3637–3645, 2016.

[20] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. Torr, and T. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1199–1208, 2018.

[21] Y. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[22] D. Matsumoto, "More evidence for the universality of a contempt expression," *Motivation and Emotion*, vol. 16, no. 4, pp. 363–368, 1992.

[23] P. Ekman, “An argument for basic emotions,” *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[24] S. K. Jarraya, M. Masmoudi, and M. Hammami, “Compound emotion recognition of autistic children during meltdown crisis based on deep spatio-temporal analysis of facial geometric features,” *IEEE Access*, vol. 8, pp. 69311–69326, 2020.

[25] J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, B. Knyazev, A. Kuharenko, C. Julio, X. Baró, H. Demirel, J. Allik, and G. Anbarijafari, “Dominant and complementary emotion recognition from still images of faces,” *IEEE Access*, vol. 6, pp. 26391–26403, 2018.

[26] R. E. Haamer, E. Rusadze, I. Lsi, T. Ahmed, S. Escalera, and G. Anbarjafari, “Review on emotion recognition databases,” *Human-Robot Interaction - Theory and Application*, vol. 3, pp. 39–63, 2017.

[27] K. Slimani, Y. Ruichek, and R. Messoussi, “Compound facial emotional expression recognition using cnn deep features,” *Engineering Letters*, vol. 30, no. 4, 2022.

[28] D. Kamińska, K. Aktas, D. Rizhinashvili, D. Kuklyanov, A. H. Sham, S. Escalera, K. Nasrollahi, T. B. Moeslund, and G. Anbarjafari, “Two-stage recognition and beyond for compound facial emotion recognition,” *Electronics*, vol. 10, no. 22, p. 2847, 2021.

[29] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[30] D. E. King, “Dlib-ml: A machine learning toolkit,” *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[31] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-shot multi-level face localisation in the wild,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5203–5212, 2020.

[32] A. Bulat and G. Tzimiropoulos, “How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks),” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1021–1030, 2017.

[33] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, pp. 886–893, 2005.

[34] D. Gabor, “Theory of communication. part 1: The analysis of information,” *Journal of the Institution of Electrical Engineers-part III: radio and communication engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[35] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[36] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE transactions on image processing*, vol. 19, no. 6, pp. 1635–1650, 2010.

[37] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 38–52, 2010.

[38] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proceedings 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, p. 65, 2010.

[39] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image and vision computing*, vol. 29, no. 9, pp. 607–619, 2011.

[40] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the national academy of sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[41] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[42] Y. Freund and R. E. Schapire, “Experiments with a new boosting algorithm,” in *Proceedings of the International Conference On Machine Learning*, vol. 96, pp. 148–156, 1996.

[43] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, “Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5562–5570, 2016.

[44] V. Kumar, S. Rao, and L. Yu, “Noisy student training using body language dataset improves facial expression recognition,” in *Proceedings of the European Conference on Computer Vision*, pp. 756–773, 2020.

[45] D. Meng, X. Peng, K. Wang, and Y. Qiao, “Frame attention networks for facial expression recognition in videos,” in *Proceedings of the 2019 IEEE international conference on image processing (ICIP)*, pp. 3866–3870, 2019.

[46] Q. Huang, C. Huang, X. Wang, and F. Jiang, “Facial expression recognition with grid-wise attention and visual transformer,” *Information Sciences*, vol. 580, pp. 35–54, 2021.

[47] A. Psaroudakis and D. Kollias, “Mixaugment & mixup: Augmentation methods for facial expression recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2367–2375, 2022.

[48] Y. Liu, J. Peng, J. Zeng, and S. Shan, “Pose-adaptive hierarchical attention network for facial expression recognition,” *arXiv preprint arXiv:1905.10059*, 2019.

[49] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, “Region attention networks for pose and occlusion robust facial expression recognition,” *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[50] S. Minaee, M. Minaei, and A. Abdolrashidi, “Deep-emotion: Facial expression recognition using attentional convolutional network,” *Sensors*, vol. 21, no. 9, p. 3046, 2021.

[51] H. Siqueira, S. Magg, and S. Wermter, “Efficient facial feature learning with wide ensemble-based convolutional neural networks,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 5800–5809, 2020.

[52] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, “Attention mechanism-based cnn for facial expression recognition,” *Neurocomputing*, vol. 411, pp. 340–350, 2020.

[53] A. V. Savchenko, L. V. Savchenko, and I. Makarov, “Classifying emotions and engagement in online learning based on a single facial expression recognition neural network,” *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2132–2143, 2022.

[54] A. P. Fard and M. H. Mahoor, “Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild,” *IEEE Access*, vol. 10, pp. 26756–26768, 2022.

[55] Y. Shi and A. K. Jain, “Probabilistic face embeddings,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6902–6911, 2019.

[56] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, “Suppressing uncertainties for large-scale facial expression recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6897–6906, 2020.

[57] P. Terhorst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, “Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5651–5660, 2020.

[58] H. Li, N. Wang, X. Ding, X. Yang, and X. Gao, “Adaptively learning facial expression representation via cf labels and distillation,” *IEEE Transactions on Image Processing*, vol. 30, pp. 2016–2028, 2021.

[59] Y. Zhang, C. Wang, and W. Deng, "Relative uncertainty learning for facial expression recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17616–17627, 2021.

[60] K. Wang and Peng, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[61] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access*, vol. 8, pp. 131988–132001, 2020.

[62] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

[63] Y. Yang, I. Saleemi, and M. Shah, "Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1635–1648, 2012.

[64] X. Zou, Y. Yan, J.-H. Xue, S. Chen, and H. Wang, "Learn-to-decompose: Cascaded decomposition network for cross-domain few-shot facial expression recognition," in *Proceedings of the European Conference on Computer Vision*, pp. 683–700, 2022.

[65] S. Li, W. Deng, and J. Du, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2852–2861, 2017.

[66] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, and D. Lee, “Challenges in representation learning: A report on three machine learning contests,” in *Proceedings of the International conference on neural information processing*, pp. 117–124, 2013.

[67] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, “Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark,” in *Proceedings of the 2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pp. 2106–2112, 2011.

[68] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, “Collecting large, richly annotated facial-expression databases from movies,” *IEEE multimedia*, vol. 19, no. 03, pp. 34–41, 2012.

[69] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[70] Z. Wen, W. Lin, T. Wang, and G. Xu, “Distract your attention: multi-head cross attention network for facial expression recognition,” *Biomimetics*, vol. 8, no. 2, p. 199, 2023.

ABSTRACT IN KOREAN

얼굴감정 인식을 위한 효율적 Few-shot 학습기법 기반 선택적 채널 공간관계 네트워크

김 채 린
숙명여자대학교 대학원 IT공학과 IT공학 전공

얼굴 감정인식 (FER) 은 컴퓨터 비전 및 인간-컴퓨터 상호작용 (HCI) 분야에서 중요한 작업 중 하나이다. FER은 얼굴 표정을 통해 감정을 인식하면서 자율 주행, 로봇공학 및 e-러닝 향상과 같은 응용프로그램들에서 널리 사용되고 있다. 하지만 그 실용성에도 불구하고, 기존의 컨볼루션 신경망 (CNN) 기반 FER은 FER 데이터셋에서 제한된 수의 샘플로 인해 과적합 문제에 직면하고 있다.

이 문제를 해결하기 위해 우리는 FER에 대한 퓨샷러닝 (FSL) 방법을 제안했다. FSL은 단 몇 개의 데이터만으로도 새로운 범주의 샘플을 예측할 수 있는 학습 메커니즘이다. FSL은 유사도 학습을 통해 데이터 간의 관계를 학습하고 테스트 데이터 또한 학습을 하는 방식과 같이 추론하며 작동한다. 이렇게 함으로써, FSL은 FER의 과적합 문제 해결에 도움을 줄 수 있다.

본 연구에서는 데이터셋 간의 관계 유사성을 학습하는 relationNet을

사용하는 방법을 제안한다. RelationNet을 기반으로 채널 선택 모듈과 추가적인 공간 데이터 구성을 설계했다. 몇 개의 데이터셋으로부터 최적의 정보를 효과적으로 활용하기 위해 샘플 피쳐들의 평균 피쳐를 대표 피쳐로 선정하였다. 다음으로는 각 채널의 대표 피쳐를 샘플 피쳐의 각 채널 정보와 비교하여 어떤 샘플의 채널 피쳐가 가장 유사한 채널 정보를 갖는지 찾게된다. 채널 정보를 비교함으로써 선택된 샘플에서의 채널이 해당 샘플 피쳐의 최적의 채널로 간주되어 추출된다. 따라서 설계된 모듈에 의해 하나의 재구성된 피쳐는 각 샘플의 채널 정보들로 구성되었다. 또한, 세밀하다는 특징에 중점을 두어, 얼굴 표정이 눈과 입술 영역에서 중요한 정보를 가지고 있다는 것을 알아내었다. 따라서 눈과 입술 이미지 패치를 생성하고 이 추가 데이터를 서포트와 쿼리 셋으로 설정했다.

선택된 최적의 피쳐와 추가적인 공간 정보가 일반화 성능을 향상시킬 수 있다는 것을 입증했다. 기존 방법과 비교하였을 때, RAFDB, FER2013, SFEW 및 AFEW 데이터셋에서의 평균 성능은 각각 3.5%, 3.68%, 5.58%, 2.31%의 정확도가 향상된다.

주제어: 얼굴감정인식 (FER), 퓨샷러닝 (FSL), 컴퓨터 비전 (CV), 딥 러닝, 채널 선택적 공간정보

학번: 2131610

Curriculum Vitae

† General Information

- Name: Chae-Lin Kim
- Date of Birth: July 22, 1996
- Place of Birth: Seoul, Korea
- Office address:
  - Department of IT Engineering,
    Sookmyung Women's University,
    100, Cheongpa-ro 47-gil, Yongsan-gu,
    Seoul, Republic of Korea
  - Tel: +82-10-4917-4468
  - E-mail: ccaa9697@sookmyung.ac.kr

† Education

- Sookmyung Women's University, *M.S.* in IT Engineering, Aug 2023.
- Sookmyung Women's University, *B.S.* in IT Engineering, Aug 2021.

† Research Interests

- ▷ Deep learning for Facial Expression Recognition.
- ▷ Deep learning for image classification.
- ▷ Machine learning for computer vision.

† Funded Research Career

- ▷ Evaluation of Speech Recognition on Foreign Language Education Services in Online Video Platform Environment, Sep. 2022 ∼ Aug. 2023, Voice Print
- ▷ Development of deep Learning based vision inspection system for bamboo toothbrush, July. 2021 ∼ Aug. 2022, Dr. NOAH

† List of Publications

▷ International Journal ◁

- C. L. Kim, B. G. Kim*, “Few-shot Learning on Facial Expression Recognition: A Comprehensive Survey,” *Journal of Real-Time Image Processing (Springer Nature)*, vol. 20:52 (Article number: 52), pp. 1-18.

▷ International Conference ◁

- Y. J. Kim, C. L. Kim, H. L. Lee and B. G. Kim*, “Recent Trend for Monocular Depth Estimation Based on Deep Learning,” *The 18th International Conference on Multimedia Information Technology and Applications (MITA)*, pp. 103-105, July. 2022 (*Best Paper Award*).

▷ Domestic Conference ◁

- H. L. Lee, C. L. Kim, B. G. Kim*, and G. H Kim, “Integrated System for Facial Alignment and Lip Generation”, KMMS Spring Conference (KMMS), vol. 26, no. 1, pp. 77-78, May. 2023.

- C. L. Kim, Y. J. Choi, B. G. Kim*, E. H. Park, K. T. Lee “A Study of the Fault Detection Technique for the Bamboo Toothbrush Manufacturing Automation”, KMMS Autumn Conference (KMMS), vol. 24, no. 2, pp. 14-17, Nov. 2021.

- C. L. Kim, J. B. Seo, and B. G. Kim* “Development of Real-time Dangerous Situation Detection Algorithm Based on Object Detection and Depth Information Fusion”, KMMS Spring Conference (KMMS), vol. 24, no. 1, pp. 565-567, April. 2021.