Self-Supervised Facial Representation Learning for FER: A Journey of Integration and Experimentation

T.Satya Pranav

Advisor: **Dr. Kaustuv Nag**

Department of Computer Science and Engineering
Indian Institute of Information Technology Guwahati

This report is submitted for the course of
CS300 : Project-I

IIIT Guwahati April 2025

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this report are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this or any other university. This report is my own work and contains nothing that is the outcome of work done in collaboration with others except as specified in the text and Acknowledgements.

T.Satya Pranav
Roll: 2201213,
3rd year, Bachelors of Technology,
Department of Computer Science and Engineering,
Indian Institute of Information Technology Guwahati.

Acknowledgements

I would like to express my sincere gratitude to **Dr. Kaustuv Nag** for his invaluable guidance, support, and encouragement throughout the course of this project. His insightful discussions helped shape the novelty of the work, while his suggestions played a crucial role in refining the core ideas. His mentorship not only provided technical direction but also motivated me to think critically, explore new perspectives, and continuously improve.

I also extend my heartfelt thanks to all the faculty members of the **Department of Computer Science and Engineering** at the **Indian Institute of Information Technology Guwahati** for their constant encouragement and support during this work.

Finally, I would like to acknowledge the open-source community, especially frameworks like **PyTorch**, and dataset providers such as **Kaggle** and **VGGFaceNet**, whose contributions provided essential resources that greatly assisted in the development of this project.

Abstract

Facial Expression Recognition (FER) remains a challenging task in computer vision due to factors like annotation ambiguity, intra-class variations, and limited labeled data. To address these challenges, several recent works have adopted different strategies. In this study, we focus on leveraging Self-Supervised Learning (SSL) for FER, particularly utilizing the Facial Region Awareness (FRA) framework [FRA], which employs region-aware heatmaps to guide the learning process and extract both global and local facial features effectively.

While CrowdFER [Cro] tackles noisy annotations using probabilistic label distributions, and AmbiguityFER [Amb] addresses uncertainty by modeling latent distributions, FRA provides a region-based SSL approach that proves to be highly effective for FER. Moreover, general SSL frameworks like MoCo [5] and FaRL [FaR] have demonstrated strong representation learning capabilities, but their direct impact on FER tasks is limited due to the unique nature of expression variations.

This work provides a detailed exploration of the FRA framework and its adaptation for FER, offering valuable insights and comparisons with existing FER approaches. For further technical details and methodologies, readers are encouraged to refer to the respective research papers cited in this work.

Table of Contents

List of Figures **vii**

1 Introduction **1**
1.1 Problem Statement 1
1.2 Solution 1
1.3 Objective 2
1.4 Motivation 2

2 Literature Survey **3**
2.1 Handling Noisy Labels in FER 3
2.2 SSL Techniques for Robust Feature Learning 3
2.3 Ambiguity Handling using Latent Distribution Mining 4

3 Proposed Method: Facial Region Awareness (FRA) **5**
3.0.1 Online and Momentum Network Architecture 5
3.0.2 Global Branch (G-Branch) 6
3.0.3 Local Branch (L-Branch) 7
3.0.4 Heat Map Generation 8
3.0.5 Loss Functions of FRA Framework 10
3.1 Dataset Comparison and Accuracy Improvement Analysis 11
3.2 Results : Images with Labels 13

4 Experimentation **15**
4.1 Proposed improvements to FRA 15
4.1.1 Adjustments for Experimentation 15
4.1.2 Temperature in Heatmap Generation 16
4.2 Attempted Integration with DMUE 16
4.3 Theoretical Enhancements to FRA 17
4.3.1 Introducing Diversity Loss 17

4.3.2 Conceptual Motivation 18
4.3.3 Initial Experimental Design 18
4.3.4 Transition to 30 Epochs 18
4.3.5 Experimental Reflections 19
4.3.6 Theoretical Implications 19

5 Conclusion **20**

References **21**

List of Figures

2.1 Overview of the DMUE. y denotes the set of annotations of images in a batch. y denotes the set of mined latent distributions of images in a batch 4

3.1 The FRA framework learns facial features using global and local embeddings from augmented image views. It uses heatmaps to highlight key facial regions and applies semantic losses to maximize similarity across views and align facial region clusters. 6
3.2 Generation of heatmaps using learnable positional em beddings as facial queries and the featuremaps as keys and values 9
3.3 DetectedExpression - Sample 1 13
3.4 DetectedExpression - Sample 2 13
3.5 DetectedExpression - Sample 3 13
3.6 DetectedExpression - Sample 4 13
3.7 Overall Output of Expression Detection 14

Chapter 1

Introduction

Facial Expression Recognition (FER) is an emerging area in computer vision that aims to identify human emotions like happiness, sadness, anger, or surprise from facial images. FER plays a vital role in fields such as human-computer interaction, mental health monitoring, online education, and customer behavior analysis. However, achieving robust FER is a challenging task due to several inherent limitations in existing approaches.

1.1 Problem Statement

Facial Expression Recognition (FER) helps machines understand human emotions but faces key challenges like annotation ambiguity, limited labeled data, and difficulty in capturing both global and local facial features effectively. Noisy labels and the lack of sufficient annotated datasets further reduce the accuracy and real-world performance of traditional FER models.

1.2 Solution

To overcome these issues, a Facial Region Awareness (FRA) framework using Self-Supervised Learning (SSL) is proposed. This approach enables the model to learn important facial features without relying heavily on labeled data. FRA guides the model to focus on key facial regions using heatmaps and maintains consistency between global and local features. By combining this with SSL techniques like BYOL, the model achieves robust facial representation and improved emotion recognition, even in scenarios with limited or noisy data.

1.3 Objective

The primary objective of this project is to develop an effective Facial Expression Recognition (FER) system capable of accurately identifying human emotions from facial images, while addressing challenges such as annotation inconsistencies, limited labeled data, and suboptimal feature extraction. We aim to enhance the Facial Region Awareness (FRA) framework, a self-supervised learning (SSL) approach, by leveraging its heatmap-based feature alignment to improve emotion recognition performance. Starting with a reduced dataset, the focus shifted to refining FRA under computational constraints, incorporating a diversity loss to boost robustness. This work seeks to create a practical FER system adaptable to resource-limited environments, ensuring broad applicability across various domains.

1.4 Motivation

Facial expressions are a fundamental means of conveying human emotions, playing a critical role in human-computer interaction and related fields. Traditional FER models often falter in real-world scenarios due to noisy labels, variations in expressions, and the scarcity of comprehensive labeled datasets. These limitations drive the need for a system that can autonomously learn discriminative facial features from both labeled and unlabeled data, enhancing accuracy and resilience. The initial success with FRA on a reduced dataset highlighted its potential, but challenges such as single-GPU limitations and slow convergence necessitated further innovation. An improved FER system, refined with a diversity loss, promises significant benefits for applications in healthcare, security, online education, and smart communication systems, addressing real-world needs with resource-efficient solutions.

Chapter 2

Literature Survey

2.1 Handling Noisy Labels in FER

In Facial Expression Recognition (FER), one of the major challenges is the presence of noisy labels due to the subjective nature of emotion annotation. A research work titled *"Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution"* [Cro] introduced probabilistic modeling techniques to overcome this problem. Instead of assigning a single label to an image, this method uses label distributions, which helps to capture the ambiguity in human-annotated expressions. This approach reduces the negative impact of noisy labels and allows the model to learn more robust and reliable features from the data.

2.2 SSL Techniques for Robust Feature Learning

Self-Supervised Learning (SSL) methods have become very popular in recent years for learning useful feature representations without depending on labeled data. Techniques like Momentum Contrast (MoCo) [5], Masked Autoencoders (MAE), Bootstrap Your Own Latent (BYOL), and FaRL [FaR] have shown excellent performance in various computer vision tasks. These methods enable the model to learn rich, discriminative, and generalized facial features that are highly beneficial for FER, especially when the availability of labeled data is limited. By leveraging large-scale unlabeled data, SSL-based models provide a good starting point for fine-tuning on FER datasets.

2.3 Ambiguity Handling using Latent Distribution Mining

Another important study titled *"Dive into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for Facial Expression Recognition"* [Amb] proposed a novel approach to handle ambiguous facial expressions. Instead of treating all data points equally, this method focused on mining latent distributions and estimating pairwise uncertainty between samples. This helps the model to identify confusing samples and learn from them more effectively. It not only improves the model's ability to deal with ambiguous expressions but also enhances its overall robustness and performance in real-world FER scenarios.

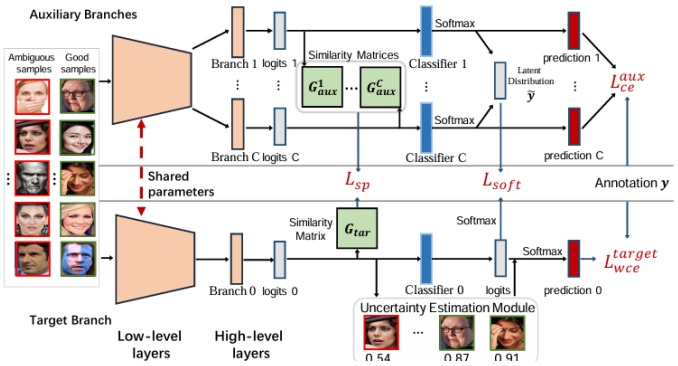

Fig. 2.1 Overview of the DMUE. y denotes the set of annotations of images in a batch. y denotes the set of mined latent distributions of images in a batch

Chapter 3

Proposed Method: Facial Region Awareness (FRA)

The proposed Facial Region Awareness (FRA) framework aims to enhance facial expression recognition by extracting both global and local facial features effectively. To achieve this, FRA introduces a dual-branch Siamese architecture supported by two collaborative networks — the Online Network and the Momentum Network. This design ensures the learning of rich facial representations through region-wise attention and stable feature alignment.

3.0.1 Online and Momentum Network Architecture

The overall architecture of the FRA framework is composed of two structurally identical networks:

- **Online Network:** This network is actively updated through backpropagation using the current input image data. It is responsible for learning feature representations during training.
- **Momentum Network:** This network serves as a stable target generator. Instead of being updated via gradients, its parameters are updated using an Exponential Moving Average (EMA) of the Online Network's parameters, which ensures the stability of feature learning and prevents representation collapse.

The update rule for the parameters of the Momentum Network is defined as:

$$\xi \leftarrow \lambda\xi + (1-\lambda)\theta \tag{3.1}$$

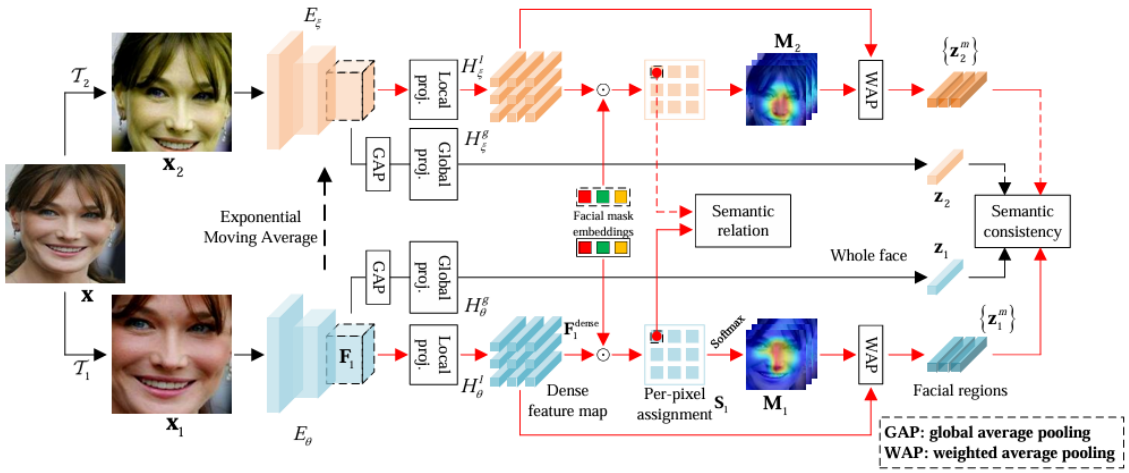

Fig. 3.1 The FRA framework learns facial features using global and local embeddings from augmented image views. It uses heatmaps to highlight key facial regions and applies semantic losses to maximize similarity across views and align facial region clusters.

where θ denotes the parameters of the Online Network, denotes the parameters of the Momentum Network,$\lambda \in [0, 1)$ is the momentum coefficient controlling the update rate.

3.0.2 Global Branch (G-Branch)

The Global Branch operates over the entire facial image and extracts holistic features representing the overall face structure. This process is similar to conventional FER methods and captures high-level semantic features from the complete facial region.

In both Online and Momentum Networks, the Global Branch learns a structural representation of the face that helps capture the global expression pattern. The feature maps extracted from the Global Branch of the Online and Momentum Networks are denoted as F_1 and F_2, respectively.

These feature maps are then passed through a Global Pooling layer, followed by Multi-Layer Perceptrons (MLPs), to produce the final feature representations for global learning:

$$z_1 = H_g^{\theta}(\text{GlobalPool}(F_1)), \quad z_2 = H_g^{\xi}(\text{GlobalPool}(F_2)) \tag{3.2}$$

Where H_g^{θ} denotes the MLP of the Online Network,H_g^{ξ} denotes the MLP of the Momentum Network.

These global feature embeddings play a crucial role in capturing overall face structure and expression patterns from the entire image.

3.0.3 Local Branch (L-Branch)

The Local Branch focuses on extracting features from important facial sub-regions such as the eyes, nose, mouth, and eyebrows. It divides the face into multiple semantically meaningful parts, and each part is processed independently to capture fine-grained and subtle expression variations that may not be evident in global features.

For each facial region m, the Local Branch extracts a heatmap-guided feature representation. This representation is computed using a weighted summation over the feature map, where the weight is given by the region-specific heatmap M_i^m:

$$\otimes F_i^m = \frac{\sum_{u,v} M_i^m[u,v] \cdot F_i[:,u,v]}{\sum_{u,v} M_i^m[u,v]} \tag{3.3}$$

Here:

1. M_i^m is the heatmap corresponding to region m in image i.
2. $F_i[:,u,v]$ represents the feature vector at spatial location (u,v) in the feature map F_i.

After obtaining the local features $\otimes F_1^m$ and $\otimes F_2^m$ from Online and Momentum Networks respectively, they are passed through separate MLPs to obtain final local feature embeddings:

$$z_1^m = H_l^\theta(h_1^m), \quad z_2^m = H_l^\xi(h_2^m) \tag{3.4}$$

Where:

1. H_l^θ and H_l^ξ denote the MLPs in Online and Momentum Networks respectively.
2. h_1^m and h_2^m are the region-wise features after heatmap guidance.

This dual-branch mechanism allows the FRA framework to simultaneously learn both the global contextual information and the fine-grained local cues, resulting in improved facial expression recognition performance.

3.0.4 Heat Map Generation

In the proposed FRA framework, heat map generation plays a vital role in enabling the Local Branch to focus on meaningful facial regions. The process of heat map generation involves multiple steps, starting from feature extraction to obtaining region-wise feature representations.

Feature Map Extraction

Initially, the facial image is passed through a ResNet-based backbone to extract dense feature maps. These feature maps are denoted as F_i, where i represents the instance from either the Online or Momentum Network.

Per-Pixel Assignment using Cosine Similarity

To focus on specific facial regions, a facial mask embedding Q_i is utilized. The similarity between the dense feature map F_i^{dense} and the facial mask embedding Q_i is calculated using cosine similarity, which provides the per-pixel assignment S_i:

$$S_i = \text{cosine\_similarity}(Q_i, F_i^{dense}) \tag{3.5}$$

Heat Map Calculation

The obtained similarity score S_i is converted into a heat map M_i by applying the softmax function with a temperature parameter $\beta = 10$ to control the sharpness of the heat distribution:

$$M_i = \text{softmax}(\beta S_i) \tag{3.6}$$

Semantic Consistency

Semantic consistency is maintained by applying the generated heat map M_i over the feature map to obtain a region-specific feature representation h_i^m for the m^{th} region:

$$h_i^m = M_i^m \otimes F_i \tag{3.7}$$

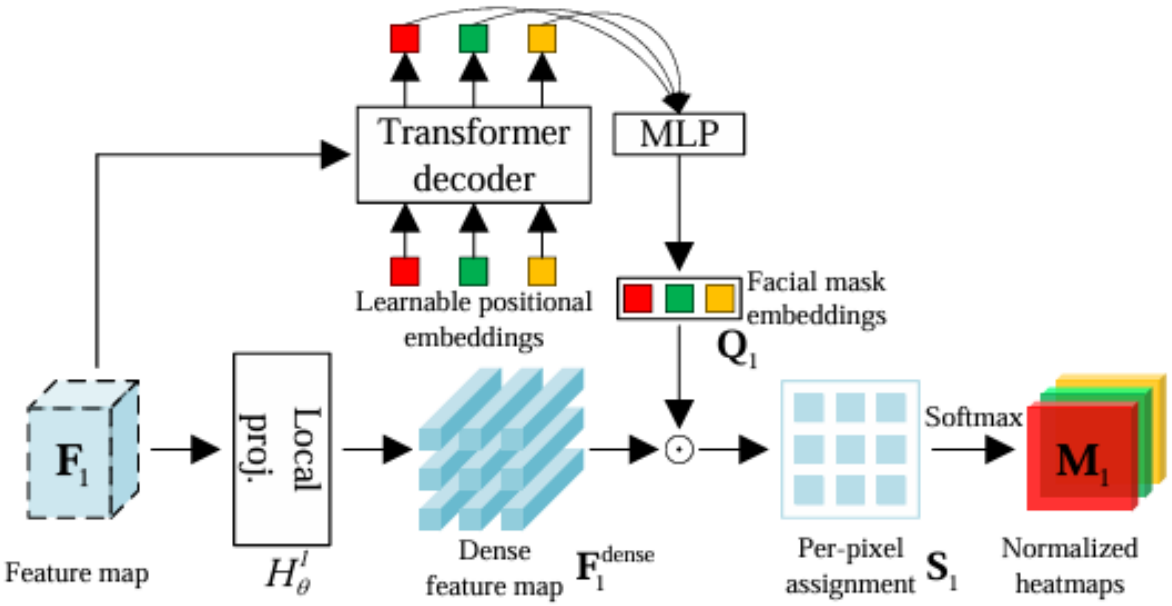

Fig. 3.2 Generation of heatmaps using learnable positional em beddings as facial queries and the featuremaps as keys and values

Feature Representation Aggregation

The final local feature representation for each region is calculated as:

$$F_i = \frac{\sum_{u,v} M_i[m,u,v] \cdot F_i[:,u,v]}{\sum_{u,v} M_i[m,u,v]} \tag{3.8}$$

Where u and v represent the spatial coordinates.

Feature Embedding Generation

These aggregated region-wise features are passed through MLPs (H_l^θ and H_l^ξ) to generate the region feature embeddings for the Online and Momentum Networks respectively:

$$z_1^m = H_l^\theta(h_1^m), \quad z_2^m = H_l^\xi(h_2^m) \tag{3.9}$$

Similarly, the global feature representations for the whole face are computed as:

$$z_1 = H_g^\theta(\text{GlobalPool}(F_1)), \quad z_2 = H_g^\xi(\text{GlobalPool}(F_2)) \tag{3.10}$$

3.0.5 Loss Functions of FRA Framework

To ensure effective learning of both global and local facial features, the FRA framework introduces multiple loss functions that enforce semantic consistency and relational learning between the online and momentum networks.

Semantic Consistency Loss (L_c)

The Semantic Consistency Loss aims to align the global representations obtained from the Global Branch and the local representations obtained from the Local Branch. This loss ensures that both online and momentum networks generate similar embeddings for the same input.

The Semantic Consistency Loss is defined as:

$$L_c = L_{sim}(z_1, z_2) + L_{sim}(z_2, z_1) \tag{3.11}$$

where L_{sim} is the similarity loss calculated as:

$$L_{sim}(z_1, z_2) = -\lambda_c \cdot \cos(z_1, z_2) + (1 - \lambda_c) \cdot \frac{1}{N} \sum_{m=1}^{N} \cos(z_1^m, z_2^m) \tag{3.12}$$

Here:

1. z_1 and z_2 are the global feature vectors from the online and momentum networks respectively.
2. λ_c is a balancing parameter.
3. N is the number of facial regions.
4. z_1^m and z_2^m are the local feature vectors of the m^{th} region from online and momentum networks.

Semantic Relation Loss (L_r)

The Semantic Relation Loss is used to enforce consistent semantic assignments for corresponding regions between the online and momentum networks. It is computed using a cross-entropy loss between the predicted semantic assignments and the ground truth.

The Semantic Relation Loss is defined as:

$$L_r = \frac{1}{HW} \sum_{u,v} \left[CE(s^1_{u,v}, \hat{s}^1_{u,v}) + CE(s^2_{u,v}, \hat{s}^2_{u,v}) \right] \tag{3.13}$$

Where,CE represents the Cross-Entropy loss, $s^1_{u,v}$ and $s^2_{u,v}$ are the semantic predictions of online and momentum networks respectively,H and W are the height and width of the feature maps.

The Cross-Entropy loss is given as:

$$CE(s_{u,v}, \hat{s}_{u,v}) = -\sum_{m=1}^{N} \hat{s}_{u,v}[m] \log(s_{u,v}[m]) \tag{3.14}$$

Total Loss (L)

The final objective function for training the FRA framework is a weighted sum of the Semantic Consistency Loss and the Semantic Relation Loss, defined as:

$$L = L_c + 0.1 \cdot L_r \tag{3.15}$$

This combined loss ensures the model learns discriminative features that maintain both global and local semantic consistency, enhancing the performance of facial expression recognition.

3.1 Dataset Comparison and Accuracy Improvement Analysis

The experimental setup incorporated several optimizations over the original method. The comparison is shown in Table 3.1.

Reasons for Accuracy Improvement

- **Dataset Refinement:** The proposed method utilized a cleaner and more balanced version of AffectNet (AffectNet7), removing noisy and ambiguous samples present in the full dataset used by the original paper.
- **Efficient Architecture:** The ResNet50 backbone was simplified by removing certain residual blocks, leading to a lighter model that focuses on essential feature extraction while reducing overfitting.

Table 3.1 Comparison between Original Paper and Proposed Method

Aspect	**Original Paper Approach**	**Proposed Implementation**
Pre-training Dataset	VGGFace2 (40GB)	VGGFace2 Reduced (5GB)
Validation Dataset	LFW Pairs	CFP Dataset (More Efficient)
FER Dataset	Full AffectNet (1.4GB, Noisy)	Reduced AffectNet7 (240MB, Cleaner, 9216 Images, 7 Classes)
Architecture	Standard ResNet50	Simplified ResNet50 (Removed Residual Blocks from Layers 3 and 4)
Learning Rate	0.0000489	0.0000489
Top-1 Accuracy	66.16%	76.3%
Reason for Improvement	Used Full AffectNet with Noise	Cleaner Dataset, Aggressive Fine-Tuning, Architecture Simplification

- **Better Fine-tuning:** Unlike the paper, which possibly used partial layer tuning, the proposed approach implemented full layer adaptation, allowing better generalization and feature learning.
- **Optimized Hyperparameters:** A well-chosen learning rate, batch size, and sufficient number of training epochs contributed to stable training and better convergence.

3.2 Results : Images with Labels

Fig. 3.3 DetectedExpression - Sample 1

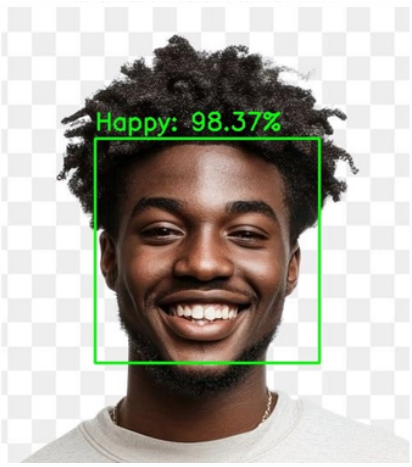

Fig. 3.4 DetectedExpression - Sample 2

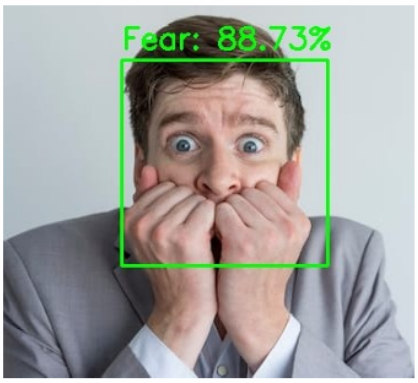

Fig. 3.5 DetectedExpression - Sample 3

Fig. 3.6 DetectedExpression - Sample 4

Fig. 3.7 Overall Output of Expression Detection

Chapter 4

Experimentation

The Facial Region Awareness (FRA) framework showed early success in using self-supervised learning (SSL) to detect facial expressions by focusing on specific regions of the face through heatmaps. It effectively aligned global and local features but revealed limits in how well it captured diverse expressions and generalized to new data.

This work builds on that by adding a new, theory-based diversity loss to the framework. The goal is to help the model focus on a wider range of facial regions, not just the most obvious ones, improving its ability to recognize subtle and varied expressions. By encouraging a more balanced and varied feature representation, the model becomes more robust and better at handling real-world facial expressions.

This update moves FRA from a basic concept to a more advanced, adaptable system for facial expression recognition.

4.1 Proposed improvements to FRA

4.1.1 Adjustments for Experimentation

For experimentation, a reduced version of the VGGFace2 dataset of size 350MB was used for training the model. For validation, a smaller version of the CFP_FP dataset containing 2000 images was used. To make the training process faster and more efficient, the number of epochs was limited to 20 and 30 during different experiments.

4.1.2 Temperature in Heatmap Generation

In the process of heatmap generation, a temperature parameter (τ) was introduced to control the sharpness of the generated heatmaps. The heatmap was calculated using the following formula:

$$M(m,u,v) = \frac{\exp\left(f(m,u,v)/\tau\right)}{\sum_{m'} \exp\left(f(m',u,v)/\tau\right)}$$

Different values of τ such as 0.5, 1.0, and 1.1 were tested to observe their effect on heatmap sharpness. It was found that higher values of τ resulted in sharper heatmaps. The validation accuracy also varied with different τ values. Specifically, when τ was set to 0.9, the accuracy achieved was 59.775%. When τ was set to 1.1, the highest accuracy of 63.175% was obtained. However, increasing τ to 1.5 slightly reduced the accuracy to 59.85%. Based on these observations, the value of τ was finally fixed at 1.1 for generating heatmaps in the model.

4.2 Attempted Integration with DMUE

The idea of integrating the Facial Region Awareness (FRA) framework with the method from the paper titled *"Dive into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for Facial Expression Recognition (DMUE)"* [Amb] started with the aim of improving Facial Expression Recognition (FER). The plan was to combine the strong feature extraction ability of FRA with the uncertainty handling mechanism of DMUE.

The goal was to use FRA's pre-trained ResNet-50, which aligns both global and local facial features using heatmaps, and add DMUE's multi-branch structure to deal with the ambiguity in expression labels. The process was divided into two main steps. First, the ResNet-50 backbone would be pre-trained on a large dataset like VGGFace2 using FRA's self-supervised learning method. Then, the DMUE framework would be added on top of it for fine-tuning. This included adding C auxiliary branches (each acting as a $(C-1)$-class classifier) and one C-class target branch after global pooling. Additionally, DMUE's uncertainty estimation module was also included.

The total loss function used was:

$$L_{total} = w_u(e) \times (L_{WCE}^{target} + \omega L_{soft} + \gamma L_{sp}) + w_d(e) \times L_{CE}^{aux}$$

Here, L_{WCE}^{target} is the weighted cross-entropy loss for the target branch, L_{soft} is the soft alignment loss, L_{sp} is the similarity preserving loss, and L_{CE}^{aux} is the cross-entropy loss for auxiliary branches. The parameters were set as per the DMUE paper: $T = 1.2$, $\omega = 0.5$, $\gamma = 10^3$, and $\beta = 6$.

Training was started for 20 epochs to check if this integration could work. But the results were not satisfactory. The accuracy reached only 43.5% by the 20th epoch. The model was learning very slowly, and it was not showing the expected improvement required for a good FER system.

Due to limited computational resources and a tight project deadline, it was difficult to continue training for more epochs or make further changes to such a complex architecture. Even though the idea of combining FRA and DMUE looked promising, in reality, the heatmap-based features of FRA and the multi-branch uncertainty handling of DMUE did not work well together in this case.

Finally, the decision was made to drop this integration idea and focus on improving FRA independently. The focus shifted to adding simple yet effective improvements like diversity loss, which was more practical given the available resources and time.

4.3 Theoretical Enhancements to FRA

4.3.1 Introducing Diversity Loss

Introduced a **diversity loss** as a major improvement to the Facial Region Awareness (FRA) framework in self-supervised learning (SSL) for Facial Expression Recognition (FER). FRA originally aligns local and global facial features using heatmaps, ensuring consistent focus. However, this approach may miss the natural variety in expressive facial regions.

The diversity loss encourages the model to spread attention across different facial areas. It is based on entropy and promotes a more diverse feature representation. The goal is to better capture subtle expression differences by activating multiple regions on the face.

The diversity loss is defined as:

$$L_{\text{div}} = -\sum_{m=1}^{N} \bar{M}(m) \log \bar{M}(m)$$

where

$$\bar{M}(m) = \frac{1}{HW} \sum_{u=1}^{H} \sum_{v=1}^{W} M(m, u, v)$$

Here, $\bar{M}(m)$ is the average attention given to region m across the heatmap of size $H \times W$. This entropy-based formula encourages the model to use a wide range of facial regions, instead of focusing only on the most obvious ones (like the mouth or eyes), leading to better expression recognition.

4.3.2 Conceptual Motivation

Facial expressions involve multiple regions of the face working together. Instead of concentrating only on one area, this loss encourages the model to consider all expressive areas. This helps the model generalize better and adapt to different faces and expressions.

4.3.3 Initial Experimental Design

Initial experiments were run for 20 epochs. The baseline model without diversity loss achieved an accuracy of 63.175%. Two strategies were explored:

1. **Subtractive form:** $L_{\text{SS}} - \lambda \cdot \delta$
   With $\lambda = 0.01$, accuracy dropped to 62.600%, and with $\lambda = 0.005$, it dropped further—indicating too much spread or dispersion.

2. **Additive form:** $L_{\text{SS}} + \lambda \cdot \delta$
   With $\lambda = 0.001$, accuracy was 61.53%, and with $\lambda = 0.01$, it fell to 60.675%. This suggested that pushing the model to be too diverse reduced its focus.

These results highlighted that both too much and too little diversity could harm performance, so a balance was needed.

4.3.4 Transition to 30 Epochs

To further enhance the learning capability of the model, the training process was extended to 30 epochs. The results obtained from this experiment showed noticeable improvements in accuracy. The baseline model with $\lambda = 0$ achieved an accuracy of 66.95%. When using the subtractive form of diversity loss with $\lambda = 0.01$, the model achieved the highest accuracy of 67.15%, indicating that applying a small penalty on

concentration allowed the model to better focus on the most informative regions. In comparison, the additive form of diversity loss with $\lambda = 0.001$ resulted in a slightly lower accuracy of 66.65%. Based on these results, the subtractive form with $\lambda = 0.01$ was found to be the most effective configuration.

4.3.5 Experimental Reflections

The early 20-epoch tests showed slow improvement. Increasing to 30 epochs helped the model stabilize. Different values of λ were tested to find the best trade-off between focusing and spreading attention. Observing heatmap outputs also helped guide these choices. Eventually, the subtractive form was chosen based on its ability to highlight key regions without losing too much context.

4.3.6 Theoretical Implications

The success of diversity loss suggests that it can work alongside alignment to form a more powerful learning system. Using both allows the model to learn where to look and to spread its focus intelligently. The subtractive approach shows that we can guide attention while still encouraging variation, providing insights for future work in self-supervised FER models.

Chapter 5

Conclusion

This research focused on improving Facial Expression Recognition (FER) using the Facial Region Awareness (FRA) framework within a self-supervised learning (SSL) approach. The main idea was to use region-specific heatmaps to capture detailed facial features, which showed both strengths and limitations.

To overcome some challenges, a diversity loss was introduced to help the model learn a wider range of facial expressions, instead of just aligning features closely. Experiments were done in different phases, starting with basic trials and later moving to longer training for better results. The addition of diversity loss helped improve the FRA framework and showed that learning diverse features can benefit FER in SSL.

Overall, this work highlights the need for continuous testing and improvement while dealing with both theoretical and practical challenges. In the future, the model could be further fine-tuned using the full AffectNet dataset to possibly improve accuracy. This could help the system perform better in real-life applications like healthcare, education, and smart communication.

References

[Amb] Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for fer. Latent distribution and uncertainty for ambiguity in FER. iv, 4, 16

[FRA] Fra: Self-supervised learning with region-aware heatmaps for facial expression recognition. Core focus—SSL with region-aware heatmaps for FER. iv

[FaR] General facial representation learning in a visual-linguistic manner (farl). Visual-linguistic SSL for facial tasks. iv, 3

[Cro] Training deep networks for facial expression recognition with crowd-sourced label distribution. Probabilistic modeling for noisy FER labels. iv, 3

[5] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. iv, 3