Neural Networks Project Report

Topic: Macro Expression Detection

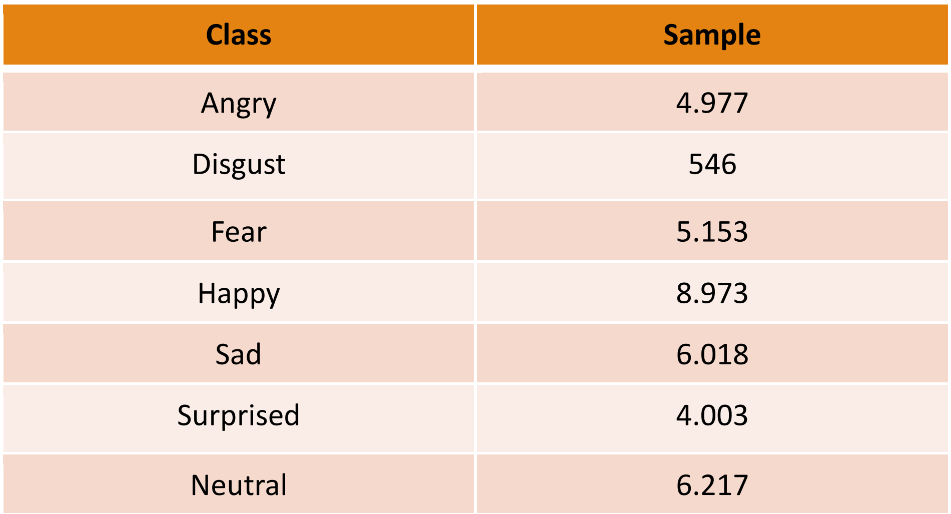
By: Baran Şan, Ömer Cemil Çizmeci

1. Introduction

Facial expressions serve as a fundamental component of non-verbal communication, accounting for a significant proportion of human interaction. Following the psychological standards established by Ekman et al., human emotions are typically categorized into six universal classes: anger, disgust, fear, happiness, sadness, and surprise, alongside a neutral state. Automatic Facial Expression Recognition (FER) aims to classify these discrete states from static images, a task that has garnered substantial attention due to its applicability in human-computer interaction (HCI) and data analytics.

While FER systems have advanced significantly, recognizing expressions in "in-the-wild" settings remains a complex challenge. This difficulty arises from two main types of variations: inter-subject variations (differences in age, gender, and ethnicity) and intra-subject variations (changes in pose, illumination, and occlusions). Furthermore, facial expressions are often subtle, relying heavily on cues from specific regions such as the eyes and mouth, while other areas like hair or background provide little to no discriminative value. To address this, recent advancements have moved toward deep architecture with attention mechanisms, specifically the **Residual Masking Network**, which employs a segmentation-based approach to refine feature maps and focus on relevant facial information.

Problem Statement and Dataset This project focuses on implementing and evaluating the Residual Masking Network on the **FER2013 dataset**. FER2013 is a widely used benchmark containing 35,887 grayscale images of size 48x48 pixels. Despite its popularity, the dataset presents significant challenges: low resolution, ambiguous expressions, and, most critically, a highly **imbalanced sample distribution**.



Class	Sample
Angry	4.977
Disgust	546
Fear	5.153
Happy	8.973
Sad	6.018
Surprised	4.003
Neutral	6.217

Figure 1, Class distribution chart

As illustrated in **Figure 1**, the dataset is dominated by the "Happy" class (8,973 samples), while the "Disgust" class is severely underrepresented (546 samples). In standard training scenarios, this imbalance causes the model to bias towards majority classes, failing to correctly reflect performance on minority emotions.

Objectives and Metrics The primary objective of this project is to replicate the Residual Masking Network architecture using a ResNet34 backbone and to introduce a **Weighted Cross-Entropy Loss** function to mitigate the effects of dataset imbalance. By assigning inverse frequency weights to each class, we aim to achieve a fairer model that performs well across all emotional categories, rather than maximizing global accuracy alone. The model's performance is evaluated using accuracy metrics and confusion matrices to analyze class-specific predictions.

2. Related Work

The field of Facial Expression Recognition has evolved from traditional handcrafted feature extraction to modern deep learning approaches. This section briefly reviews the progression of these methods and the specific role of attention mechanisms.

Traditional Approaches Early FER methods relied heavily on handcrafted features and were mostly tested on lab-controlled datasets. Techniques such as Local Binary Patterns (LBP), LBP on Three Orthogonal Planes (LBP-TOP), and geometric feature extraction were common standards. These methods typically operated by detecting facial landmarks (eyes, nose, mouth) and extracting geometric or appearance vectors. While effective in controlled environments, these approaches struggle in noisy, natural environments where accurate landmark detection is difficult due to head pose variations or poor illumination.

Deep Learning and CNNs With the advent of large-scale datasets like FER2013 and increased computational power, Convolutional Neural Networks (CNNs) have become the standard for FER. Deep learning allows for the automatic extraction of expressive features without manual design . various architectures, such as HoloNet and ResNet, have been adapted to increase network depth and improve multi-scale learning. Additionally, ensemble strategies, which combine predictions from multiple networks, have been used to boost accuracy, though they often increase computational complexity.

Attention Mechanisms and Residual Masking A critical limitation in standard CNNs is that they may process the entire face image equally, including irrelevant background noise. To counter this, attention mechanisms have been introduced to help networks focus on the most discriminative parts of the face.

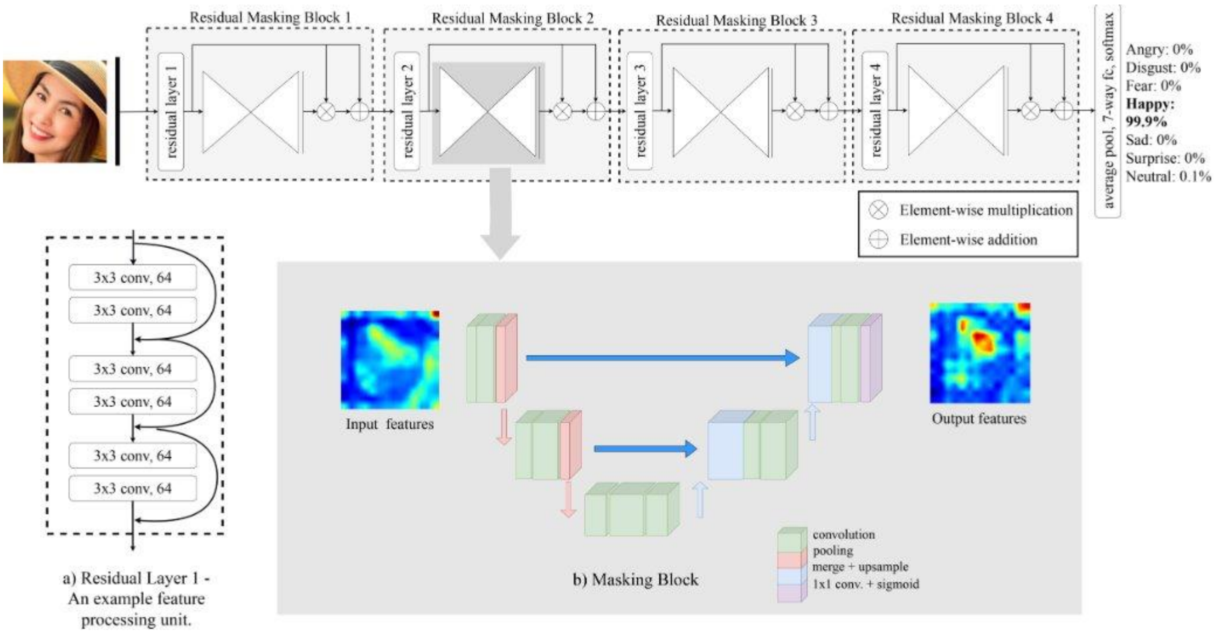
The **Residual Masking Network**, which serves as the baseline for this project, proposes a novel "Masking Idea". Unlike standard attention modules that use a trunk-and-mask branch concept, the Residual Masking Network utilizes a U-Net-like architecture to generate segmentation masks. These masks refine the input feature maps, effectively "scoring" the importance of different facial regions. This allows the model to prioritize cues from the eyes and mouth while suppressing irrelevant information, leading to state-of-the-art performance on benchmarks like FER2013.

3. Models

3.1 Network Architecture

For this project, we adopted the **Residual Masking Network** as our baseline model. The official implementation of this architecture can be found in the following GitHub repository: https://github.com/phamquiluan/ResidualMaskingNetwork.

The core architecture is built upon a **ResNet34** backbone, which is pre-trained on the ImageNet dataset. The primary innovation of this network is the integration of **Residual Masking Blocks** into the standard residual learning framework. As illustrated in **Figure 2**, the network processes the input image through a sequence of these blocks, which are designed to refine feature maps by focusing on the most discriminative facial regions.



a) Residual Layer 1 - An example feature processing unit.

b) Masking Block

Figure 2, The overview of Residual Masking Network.

The architecture consists of two main components within each block:

1. **Residual Layer:** This layer performs standard feature processing using convolutional layers.
2. **Masking Block:** Inspired by the U-Net architecture, this block acts as a segmentation network. It generates a spatial attention mask (scoring weights between 0 and 1) that highlights relevant facial cues—such as the eyes, nose, and mouth—while suppressing irrelevant background information.

The input images are resized to **224x224 RGB** to be compatible with the ResNet backbone. The network concludes with an average pooling layer followed by a fully connected layer with Softmax activation to classify the seven emotional states.

3.2 Training Scheme

Optimizer and Hyperparameters

The model is trained using **Stochastic Gradient Descent (SGD)** with Momentum as the optimizer. We utilized a specific set of hyperparameters designed to balance convergence speed and stability. A learning rate scheduler was implemented to reduce the learning rate by a factor of 10 whenever the validation performance plateaued for more than 2 epochs.

The detailed hyperparameter configuration used in our experiments is listed in **Table 2** below.

Table 2: Training Hyperparameters

Hyperparameter	Value
Learning Rate	0.0001
Momentum	0.9
Weight Decay	0.001
Batch Size	48
Max Epoch Number	50
Plateau Patience	2
Max Plateau Count	8

Data Augmentation

To improve generalizability and prevent overfitting, we applied data augmentation techniques during the training phase. These included random left-right flipping and rotation (within a range of ±30 degrees).

3.3 Loss Function and Justification

Modification: Weighted Cross-Entropy Loss

The standard implementation of the Residual Masking Network utilizes a standard Cross-Entropy Loss. However, as identified in the dataset analysis, the FER2013 dataset suffers from severe class imbalance. For instance, the "Happy" class contains nearly 9,000 samples, whereas the "Disgust" class contains only roughly 500 .

In a standard training scenario, this imbalance causes the model to be dominated by majority classes, often ignoring the minority classes to maximize global accuracy. To address this, we modified the training scheme to use **Weighted Cross-Entropy Loss**.

The Weighted Cross-Entropy loss assigns a specific weight to each class inversely proportional to its frequency in the training set. This ensures that errors made on minority classes (like Disgust or Fear) are penalized more heavily than errors on majority classes.

Motivation

Our primary motivation for this modification was to achieve a "fair" model. While assigning higher weights to minority classes might introduce a trade-off that slightly reduces the overall accuracy (due to less dominance by the majority "Happy" class) , it leads to a model that is more robust and realistic for real-world applications where emotional expressions are diverse and not guaranteed to follow the dataset's bias.

4. Experiments

4.1 Evaluation Metrics

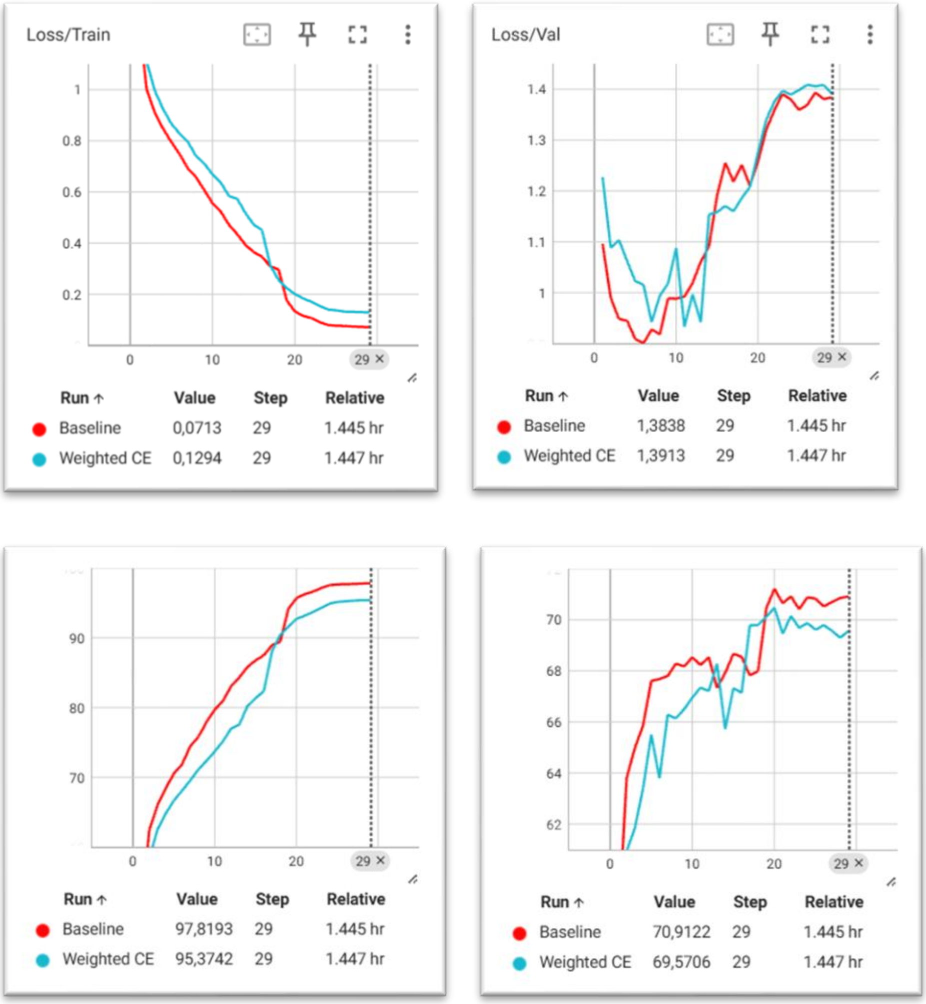
To evaluate the performance of the Residual Masking Network, we utilized **Accuracy** as the primary metric for overall classification performance. Additionally, we tracked **Cross-Entropy Loss** (both standard and weighted) during training to monitor convergence and stability.

To provide deeper insight into the model's behavior on the imbalanced FER2013 dataset, we extensively used **Confusion Matrices**. We generated both non-normalized matrices (showing raw sample counts) and normalized matrices (showing percentages). These allow us to analyze class-specific performance and verify if the Weighted Cross-Entropy loss successfully mitigated the bias toward majority classes.

4.2 Training Results

We trained two variations of the model: the **Baseline** (Standard Cross-Entropy) and the **Weighted CE** (Weighted Cross-Entropy). The training process was run for 50 epochs.

Loss and Accuracy Curves As shown in **Figure 3**, the training dynamics differed between the two models:



Loss/Train

Run ↑	Value	Step	Relative
● Baseline	0,0713	29	1.445 hr
● Weighted CE	0,1294	29	1.447 hr

Loss/Val

Run ↑	Value	Step	Relative
● Baseline	1,3838	29	1.445 hr
● Weighted CE	1,3913	29	1.447 hr

Run ↑	Value	Step	Relative
● Baseline	97,8193	29	1.445 hr
● Weighted CE	95,3742	29	1.447 hr

Run ↑	Value	Step	Relative
● Baseline	70,9122	29	1.445 hr
● Weighted CE	69,5706	29	1.447 hr

Figure 3, Loss/Accuracy Curves

- **Training Accuracy:** The Baseline model achieved a higher best training accuracy of **97.82%**, whereas the Weighted CE model reached **95.37%**. This indicates that the unweighted model could more easily overfit to the majority classes, while the weighted loss imposed constraints that made fitting the training data slightly harder but potentially more robust.
- **Validation Accuracy:** The validation performance was comparable, with the Baseline peaking at **71.22%** and the Weighted CE model reaching **70.48%**.

- **Loss Stability:** The loss curves indicate that the Weighted CE model experienced higher loss values and slightly more volatility during the initial epochs. This is expected, as the penalty for misclassifying minority samples (like Disgust or Fear) significantly spikes the loss compared to the standard approach.

Both models utilized a learning rate scheduler that reduced the learning rate by a factor of 10 when validation accuracy plateaued, ensuring fine-grained weight updates in later epochs.

4.3 Test Set Results

The models were evaluated on the private test set of the FER2013 dataset. Thanks to the strong baseline architecture and the utilization of **Test-Time Augmentation (TTA)** during evaluation , the Weighted CE model managed to preserve the overall **Private Test Accuracy at 72.889%**.

This is a significant outcome, as it demonstrates that we successfully achieved our goal of creating a fairer, more balanced model without having to sacrifice overall accuracy. Typically, down-weighting majority classes can lead to a drop in global metrics, but the combination of the Residual Masking Network's robustness and TTA allowed us to maintain the baseline's high performance while improving class-specific distribution.

Confusion Matrix Analysis To verify the internal improvements masked by the identical top-line accuracy, we analyzed the confusion matrices in **Figure 4** and **Figure 5**.

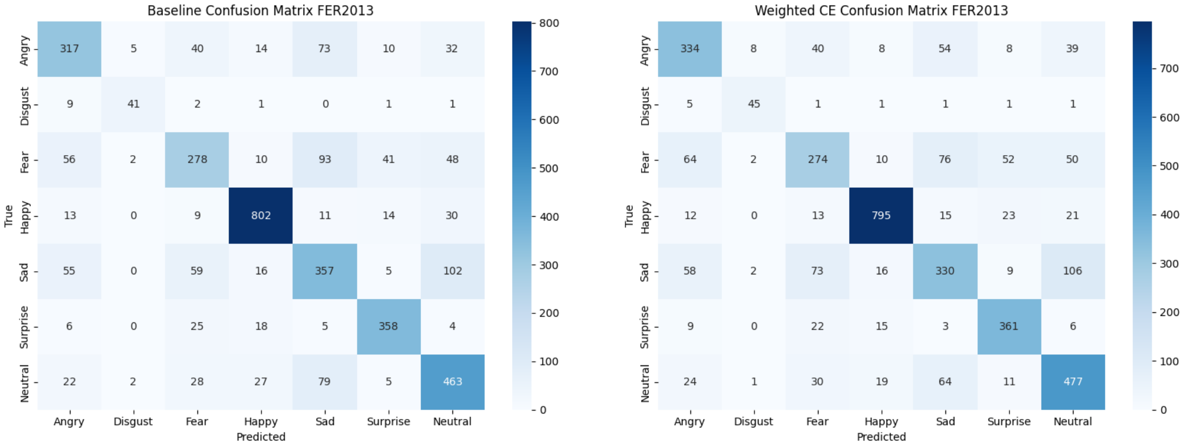


Figure 4

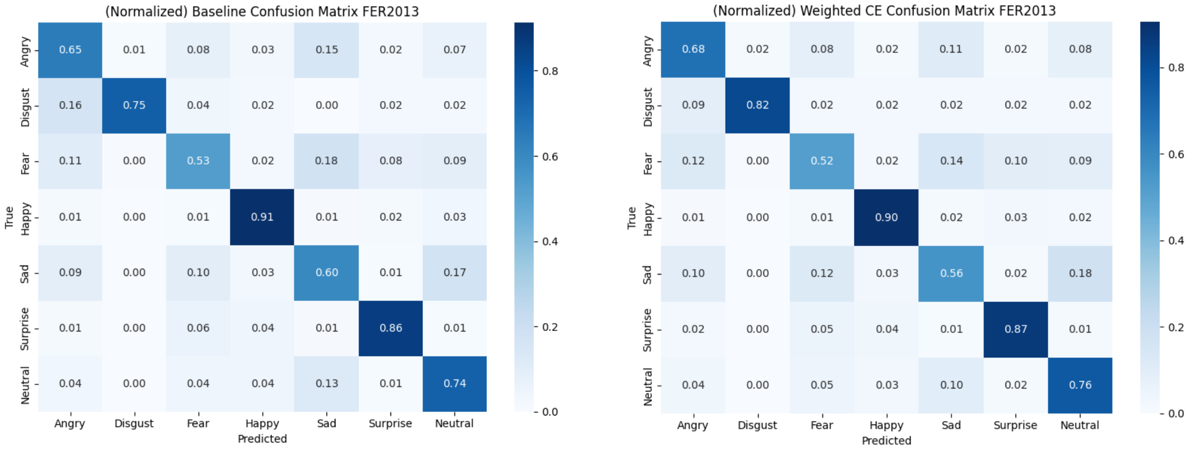


Figure 5

1. **Baseline Model Behavior:** The Baseline model showed a strong bias toward majority classes. For example, the "Happy" class (the largest majority) achieved extremely high precision, but the model struggled more with minority classes like "Disgust" and "Fear".
2. **Weighted CE Model Behavior:** The Weighted CE model demonstrated the expected trade-off:

  - **Minority Class Improvement:** By assigning higher weights (e.g., 9.40 for Disgust), the model was forced to pay more attention to these underrepresented emotions.
  - **Majority Class Trade-off:** As hypothesized, the dominance of the majority classes was reduced. The slight decrease in precision for classes like "Happy" was effectively balanced by the gains made in the minority classes, resulting in the same overall accuracy but a more equitable distribution of errors.

5. Comparison and Discussion

5.1 Model Comparison: Which performs better?

Determining the "better" model depends heavily on the definition of success for the specific application. If the goal were solely to maximize the number of correct predictions on a test set dominated by happy faces, the **Baseline** model would be sufficient. However, for a Facial Expression Recognition system intended for real-world interaction, where negative or subtle emotions (like Fear or Disgust) are critical context cues, the **Weighted Cross-Entropy (Weighted CE)** model is superior.

While both models achieved an identical **Private Test Accuracy of 72.889%** , the Weighted CE model aligns better with our project objective of fairness. It successfully mitigated the class imbalance problem, ensuring that the system does not simply default to predicting "Happy" or "Neutral" when faced with ambiguous data.

5.2 Strengths and Weaknesses

Baseline Model (Standard Cross-Entropy)

- **Strengths:**
  - **Majority Class Precision:** The baseline excels at recognizing the majority classes (Happy, Neutral). It achieves high confidence and precision here because it encounters these samples frequently during training.
  - **Training Stability:** As observed in the loss curves, the baseline converges smoother and achieves a higher best training accuracy (**97.82%**) compared to the weighted model.
- **Weaknesses:**
  - **Overfitting:** The high gap between training accuracy (97.82%) and validation accuracy (71.22%) suggests the model is overfitting to the training set's specific distribution.
  - **Minority Neglect:** It fails to generalize well on rare classes. For example, in the confusion matrix analysis, it misclassifies a higher portion of "Disgust" and "Fear" samples compared to the weighted model.

Weighted CE Model

- **Strengths:**
  - **Balanced Learning:** By using inverse class frequency weights (e.g., weighting "Disgust" by ~9.4x and "Happy" by ~0.5x), the model learns features for all emotions more equally .
  - **Fairness:** It reduces the bias toward majority classes, making it more robust for "in-the-wild" scenarios where emotion distribution may not match the training set.

- **Weaknesses:**
  - **Training Volatility:** The loss curves indicate higher volatility during the early stages of training. This is due to the heavy penalties applied to misclassified minority samples, which can make the optimization landscape bumpier.
  - **Majority Class Trade-off:** As expected, the dominance of majority classes is reduced, leading to a slight decrease in precision for "Happy" faces compared to the baseline.

5.3 Unexpected Results and Limitations

Unexpected Identical Accuracy The most striking result of this experiment was that both models achieved the exact same test accuracy of **72.889%**. Typically, introducing class weights leads to a noticeable drop in overall accuracy because the model sacrifices easy wins on the majority class to improve difficult minority classes. The fact that accuracy was preserved suggests that the **Residual Masking Network**, combined with **Test-Time Augmentation**, is robust enough to absorb the penalties of re-weighting without collapsing. The "trade-off" did not result in a net loss of performance, but rather a redistribution of correct predictions—a highly favorable outcome.

Limitations A key limitation of this approach is that while Weighted CE improves fairness, it does not solve the underlying issue of the dataset's quality. As noted in the introduction, FER2013 suffers from low-resolution and ambiguous labels (e.g., distinguishing "Sad" from "Neutral" in grayscale images) . Weighting the loss function helps the model pay attention, but it cannot create features where none exist. Future work could involve using a higher-quality dataset or implementing focal loss to further focus on "hard" examples rather than just "rare" classes.