Final Report:
Facial Emotion Recognition using a Convolutional Neural Network

Antonia Härle Clara Sophie Negwer Andre Ngo Jonas Saathoff

Abstract

Automatic facial expression recognition (FER) is an extensively studied task in deep learning. Accurate emotion recognition can provide valuable insights in psychological or consumer behaviour research, or even in the detection of early signs of mental health disorders. While numerous CNN-based approaches exist, they mostly rely on pretrained, resource-intensive architectures. Our work explores the potential of training a lightweight CNN from scratch. The proposed CNN is based on an initial architecture of four convolutional layers, average pooling, and a softmax output layer. By progressively integrating residual and squeeze-and-excitation (SE) blocks, dropout, and batch normalization, we improved validation accuracy from 58.50% over 71.86% on FER-2013 to 86.32% on FER+, while reducing model size from 34M to 10M parameters. The resulting model operates in real-time with only 794 MFLOPs per input and achieves competitive performance without pretraining. Grad-CAM visualizations confirm that the model focuses on semantically meaningful facial regions. Our results show that architectural precision and careful training strategies can outperform larger, pretrained models in accuracy. Additionally, comparing FER-2013 and FER+ underscores the importance of annotation quality, with model performance improving by nearly 15% when trained on the revised FER+ labels. Finally, we present an interface for real-time, video- and image-based FER based on our trained CNN, showing that practical applications can be built with relatively moderate computational effort.

1. Introduction

Recognizing other people's emotions is essential in everyday life and plays an important role in social interaction. One perspective argues that facial expressions convey information about the emotional state of a person [10], whereas from a more evolutionary standpoint, they are understood as signals to regulate social interactions [12]. Regardless of the theoretical perspective, the ability to identify and interpret facial expressions is a fundamental task that supports and facilitates human interaction. Research suggests that only 7% of communication is verbal, 38% is vocal and 55% is mediated by visual cues such as facial expression and gestures [26].

While expression recognition often occurs unconsciously, the task itself is somehow complex. This is underlined by the fact that humans typically do not achieve accuracies higher than 90% and often only 60 - 80% in recognizing emotions in other persons [13, 31]. Moreover, individuals with certain mental health conditions, such as autism spectrum disorder or personality disorders, often experience difficulties in emotion recognition and imitation, which can significantly impair social communication [9, 45].

Facial expression recognition (FER) is a technique to detect human emotions in images or videos to identify information about other persons' facial expressions and emotional state [21]. With advances in machine learning and computer vision, various automated FER approaches were developed. Early systems relied on simple classifiers, but nowadays modern machine learning models, especially convolutional neural networks (CNN), are able to achieve accuracies exceeding 70 - 80% in classifying emotions from facial images, suggesting that their performance in the FER task is at least comparable to that of humans [20, 21, 28, 29].

These improvements open up a wide range of practical fields that can benefit from improving machine learning models. For example, Drimalla et al. [9] used an open-source tool to extract facial action units from participants and analyze differences in emotion recognition and imitation between healthy controls and individuals with autistic spectrum conditions. Research in automatic emotion encoding has further potential in areas such as early detection of mental health disorders, consumer science or lie detection [14, 15]. Developing effective and accessible deep learning models for FER is therefore of growing importance.

This work contributes to that goal by developing a lightweight, custom CNN trained from scratch, created as part of a university practical to deepen the understanding of CNN models without relying on pretrained architectures. The model is integrated into tools that are capable of classifying emotions in real-time using webcam input, static

images, or video recordings. As these tools are intended as a prototype, future work could expand and refine them, opening up possibilities for applications in psychological research and diagnostics or consumer science.

2. Related Work

A coherent understanding of emotional theory is fundamental to the reliable application of machine learning in emotion recognition. In one of the best studied works of emotion recognition by Paul Ekman, the basic emotions are split into six principal emotions: happiness, sadness, anger, surprise, fear and disgust [10]. Later, this was extended by a seventh category, *neutral* and/or *contempt*, depending on the specific labeling strategy or research focus.

Translating these emotion theories into machine learning applications requires large, annotated datasets that reflect the defined emotion categories. Over the years, numerous databases have been collected to support this goal. Since our work focuses on building a model that can classify emotions in real-world conditions, datasets that contain images with variability in lighting, background and facial appearance are especially relevant. The most prominent datasets, that contain real-world-scenario images, include FER-2013 [34], FER+ [3], AffectNet [2], AFEW [11] and RAF-DB [38]. Among these, AffectNet is the largest, containing approximately 1,000,000 high-resolution images of eight emotions, which are represented in a highly diverse manner. FER-2013 is the second largest, with nearly 36,000 grayscale images labeled according to Ekman's seven basic emotions [10]. Despite its slightly lower resolution, it remains one of the most studied datasets in CNN-based FER research. FER+ is an enhanced version of FER-2013, introduced to address issues with inaccurate labeling in the original dataset. RAF-DB includes around 30,000 RGB images, so roughly 6,000 fewer than FER-2013, but offers a higher resolution. AFEW, in contrast, is based on short video clips, allowing for temporal modeling of emotions.

In the beginning, machine learning approaches on facial emotion recognition (FER) based on extracting features from the images as a first step and then using a classifier (e.g. support vector machine, random forest or neural network [4, 7, 36]) as a second step. With computational resources improving and image classification tasks becoming more complex, several research groups developed deep learning based approaches for FER, as these have achieved state-of-the-art performance across a variety of applications [8].

Various deep learning architectures have been explored for FER - for example, deep belief networks (DBNs), recurrent neural networks (RNNs), or general adversarial networks (GANs). However, CNNs are currently part of the most prominent models for working with FER datasets. A typical CNN consists of three main layer types: convolutional, pooling, and fully connected layers. Convolutional layers apply learnable filters on the input images to extract features such as facial components, edges or shapes. Pooling layers reduce the spatial dimensions of feature maps, improving computational efficiency and helping prevent overfitting [48]. Finally, fully connected layers use the extracted features to perform classification [51]. Hyperparameters like depth, stride, and padding influence the network's complexity and output size [42].

One of the first CNNs that was applied to FER was proposed by Matsugu et al. [25]. Over the past decades, numerous techniques have been introduced to enhance the performance of CNNs. These include advanced pooling strategies, regularization methods, extensive data augmentation, batch normalization, as well as improvements in optimization algorithms and learning rate scheduling. Another important aspect influencing CNN performance is the network depth, i.e. the number of layers use to capture abstract features. This led to deeper architectures like AlexNet [22], VGG [39] and ResNet [41] that further improved the performance, particularly when pretrained on large-scale datasets. For the FER-2013 dataset, as this is often used as a benchmark in comparing model performance, pretrained models like VGG, Inception and ResNet achieved accuracies between 65.20% and 73.28% in two reasearch papers [20, 33].

In addition to approaches based on pretrained models, some research has demonstrated that well-designed, compact CNNs trained from scratch can also yield competitive results. For example, Minaee et al. [28] proposed a lightweight attentional CNN consisting of four convolutional layers, followed by two fully connected layers and a spatial transformer network that directs attention to the most salient facial regions [19]. Despite its simplicity and shallow depth, the model achieved 70.02% accuracy on FER-2013, outperforming several deeper architectures and demonstrating the effectiveness of focused architectural design. Mohan et al. [29] employed a similar base architecture and enhanced it by applying batch normalization to the outputs of both convolutional and fully connected layers, achieving 79% accuracy on FER-2013 and 82% on RAF-DB.

Additionally, many targeted modifications have been proposed to further enhance the performance of CNNs. These include residual connections to address vanishing gradient problems in deeper networks [16, 52], and squeeze-and-excitation (SE) blocks to adaptively adjust channel-wise feature responses [18]. For instance, Ma et al. [24] incorporated a SE module into a lightweight attention-based CNN to emphasize informative feature channels and suppress less relevant ones based on global context, achieving an accuracy of 87% on the RAF-DB dataset.

The most recent research has increasingly focused on Vi-

sion Transformer (ViT) based approaches for FER, as these architectures often outperform conventional CNN models in terms of classification accuracy, particularly on large and complex datasets (e.g. [1, 30]). However, these approaches and other high-performing models often rely on deep and computationally intensive architectures. To keep the model lightweight and due to our educational focus on building a CNN from scratch, we drew inspiration from compact CNN designs proposed by Minaee et al. [28] and Mohan et al. [29], which demonstrate that shallow networks can still perform effectively. Building on this foundation, we enhanced our model by integrating residual SE blocks, applying batch normalization, and experimenting with a spatial transformer [19, 28]. In addition, we explored various data augmentation strategies to improve generalization and reduce overfitting.

3. Methodology

3.1. Model Architecture

For our emotion classification model, we used the CNN architecture proposed by Minaee et al. [28] as a basis, though with some major modifications. They demonstrated that a compact CNN with fewer layers can achieve good results on benchmark FER datasets. While their full model integrates a spatial transformer to dynamically focus on salient facial regions, we omitted this module to maintain a lower implementation complexity and to focus on core CNN performance. We then expanded this reduced model by also including batch normalization, drop out layers, residual blocks [16], as well as Squeeze-And-Excitation blocks [18].

The structure of our model is illustrated in Fig. 1, Fig. 2, Fig. 3 and Fig. 4. After preprocessing of the data, features are extracted by a convolutional layer and a batch normalization layer. These are then followed by a ReLU activation function, four residual SE blocks. Lastly, adaptive average pooling is applied for downsamling Fig. 1.

These features are then passed to the classifier, which consists of a layer flattening the data and a dropout layer with a dropout rate of 0.4. This is then passed to a fully connected linear layer followed by a ReLU activation function, a dropout of rate 0.3 and another fully connected linear layer Fig. 2.

Residual blocks create direct paths for gradients during backpropagation through skip connections. This is to tackle the so-called vanishing gradients issue, which is a common problem especially in deeper networks. During backpropagation, the gradients can become ever so small due to repeated multiplication of small values, causing the first layers to update their gradients slowly or not at all.

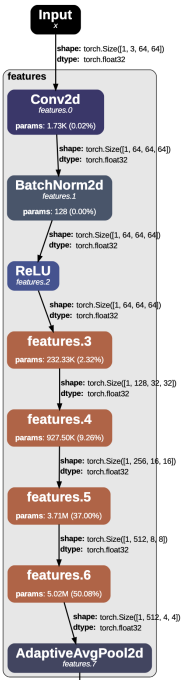

Figure 1. Feature extraction. Conv Layer → BatchNorm → ReLU → 4 Custom Blocks → MaxPooling. The custom feature blocks consist of a Residual Block (Fig. 3) and a Squeeze-And-Excitation Block (Fig. 4).

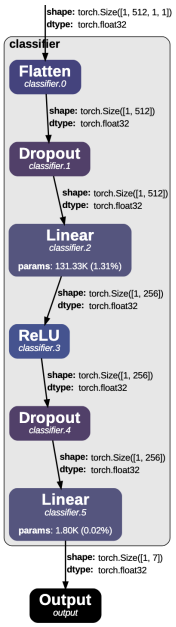

Figure 2. Classification Head. Flatten → Dropout → Linear (fully connected) → ReLU → Dropout → Linear (fully connected).

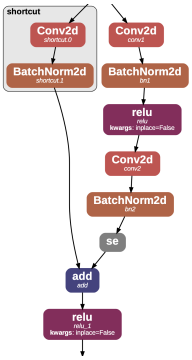

Figure 3. Residual Block. Includes a skip connection (shortcut), providing a direct path for gradients to flow backwards during backpropagation.

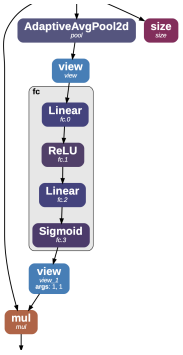

Figure 4. Squeeze-And-Excitation Block. Contained in the Residual Block (named "se" in Fig. 3), it employs channel attention to focus on important and put less weight on less important channels.

To resolve this, the input x (also called identity) of a residual block is added to the block's learned transformation $F(x)$, resulting in the output formula

$$y = F(x) + x \tag{1}$$

When $F(x)$ is small, the x summand "dominates" the term, since its derivative during backpropagation is 1, which is essentially an identity mapping. This prevents gradients from completely vanishing and leads to at least the identity component to flow through. [16]. This way, residual blocks enable CNNs to learn faster and more efficiently. The equivalent implemention used in our model is visualized in Fig. 3.

The squeeze-and excitation block is shown in Fig. 4. With it, the model is able to focus on more important channels for classification by generating attention weights for each channel. During the squeeze operation, Global Average Pooling is used to get a single value representation of each channel. Then, in the excitation phase, a small neural network consisting of two FC-Layers, ReLU, and a Sigmoid function generate "attention weights" from 0 to 1. In the last step, the original channels are multiplied by these weights, weighting the respective channels according to their importance. For little computational effort, this often leads to small improvements in accuracy. [18]

3.2. Training Setup

For training our CNN models, we used the PyTorch framework with the AdamW optimizer and a cross-entropy loss function. The dataset was organized into predefined folders for training and validation. Approximately 80 % of the images were used for training and 20 % for validation. This static split was applied consistently throughout all experiments.

The learning rate was initialized at 1×10^{-3} and adjusted using a ReduceLROnPlateau scheduler, which reduces the learning rate by a factor of 0.5 if the validation loss does not improve for 3 consecutive epochs. Early stopping with a patience of 25 epochs was employed to prevent overfitting. A minimum learning rate of 1×10^{-6} was enforced to prevent the model from stagnating due to excessively small updates.

We trained all models for a maximum of 200 epochs using a batch size of 32. Each training epoch was followed by evaluation on the validation set, and the model with the best validation loss was saved.

The training pipeline includes logging of loss and accuracy for both training and validation phases, allowing for detailed performance tracking. Additionally, we visualized the confusion matrix and per-class metrics to better understand model behavior.

All experiments were conducted using Google Colab with GPU acceleration. We ensured consistent random seeds and hardware settings to allow reproducibility and fair comparison between configurations.

3.3. Grad-CAM Architecture and Integration

Grad-CAM was chosen due to its strong class-specific localization, architectural flexibility, and ease of integration with PyTorch. Compared to alternative methods such as occlusion, CAM, or generic activation maps, Grad-CAM provides high interpretability while being applicable to a wide range of CNN architectures [35, 50, 53]. It generates heatmaps by backpropagating the gradients of the target class into the last convolutional layer, allowing for intuitive visualizations of the model's decision focus.

Our PyTorch-based implementation registers forward and backward hooks on a target convolutional layer to access its activations and the gradients during backpropagation. Given an input image and a predicted class c, we compute the importance weight α_k^c of the k-th feature map A^k by taking the global average of the gradients:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2}$$

These weights are then used to compute a class-specific localization map L^c as a weighted combination of the feature maps:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{3}$$

We normalize this heatmap and overlay it on the input image for visualization.

Target Layer Selection. We evaluated several layers within the feature extractor to determine which one yields the most informative heatmaps. A summary of our selection criteria is shown below:

- `features[0--2]`: Too shallow; mostly low-level patterns.
- `features[3--4]`: Optional; better spatial maps, but less semantic meaning.
- `features[5]`: Deeper and still spatial; semantically meaningful.
- `features[6]`: Best trade-off; deep and spatially preserved.
- `features[7]` and FC layers: No spatial output; unsuitable for Grad-CAM.

In our final setup, we selected `features[6]` as the target layer for Grad-CAM, as it retains both semantic richness and spatial structure. The resulting activation patterns across layers are illustrated in Figure A2.

3.4. Preprocessing and Augmentation

Since AffectNet is the largest FER dataset containing facial expressions captured in real-world scenarios, it would be well-suited to our research objectives. However, due to its restricted access, we initially selected FER-2013 as our primary dataset. Although FER-2013 is considerably smaller and consists of lower-resolution images (48×48 pixels), it offers a diverse range of emotional expressions under real-world conditions and remains a widely used benchmark in FER research. Given that the dataset has been criticized for noisy labels, which may partly explain why even human accuracy on FER-2013 only reaches around 65% [13], we further evaluated our model by training it on FER+ [3].

For preprocessing, all images were resized to 64×64 pixels using bilinear interpolation to provide more spatial detail to the network. They were then normalized using the dataset-specific mean and standard deviation (FER-2013: $\mu = 0.5073$, $\sigma = 0.2061$; FER+: $\mu = 0.5072$, $\sigma = 0.2062$). Augmentations were implemented using the Albumentations library [5], and the final image tensors were converted to PyTorch format.

To improve generalization and reduce overfitting, the training images were augmented using different transformation techniques [43]. The final augmentation pipeline included horizontal flipping with a 50% probability ($p = 0.5$), random brightness and contrast adjustments ($p = 0.3$), and combined affine transformations ($p = 0.7$) – namely, random shifts(±10%), scaling (±15%), and rotations (±15°). To further increase robustness, Gaussian blur ($p = 0.1$) was applied to simulate low-resolution input, and Coarse Dropout ($p = 0.25$) was used to randomly mask out rectangular regions of the image to simulate occlusions.

To focus on expressive emotions, the neutral (and contempt for FER+) class was excluded from training, resulting in six categories: angry, disgust, fear, happy, sad, and surprise. The dataset was split into training and validation sets, and data was loaded in batches of size 32.

3.5. Video Demo

To demonstrate the practical application of our model in real-time settings, we implemented a live webcam demo using OpenCV. The system captures video frames from the webcam, performs face detection using Haar cascades, and classifies emotions on the detected faces using our trained CNN model.

Each frame is first converted to grayscale and scanned for frontal faces using a pretrained Haar feature-based cascade classifier provided by OpenCV. Once a face is detected, the region of interest is extracted, converted to RGB, and passed through the same preprocessing pipeline as used during model training, including resizing to 64×64 pixels, normalization, and conversion to PyTorch tensors via Albumentations [5]. The preprocessed image is then forwarded through the CNN, and emotion probabilities are computed using the softmax function. The predicted label, along with class probabilities, is overlaid directly onto the live video stream. To maintain responsiveness, inference is performed on a single frame at a time using the best-performing model weights (‘best-weights.pt‘).

Compared to batch-based image classification, real-time video inference introduces constraints in terms of latency and model throughput. To balance performance and usability, the system processes every frame sequentially and omits Grad-CAM visualization due to runtime constraints. Nevertheless, this interactive prototype illustrates the practical deployment potential of our emotion recognition model in real-world scenarios.

4. Experiments

4.1. Datasets and Augmentation

To evaluate the impact of different data augmentation strategies on generalization and overfitting, we conducted a series of experiments on FER-2013 and FER+ [27, 37]. The re-

sulting training and validation accuracies are summarized in Tab. 1.

Without augmentation, the model clearly overfits on FER-2013, achieving high training accuracy but relatively low validation performance. Introducing horizontal flipping notably improved validation accuracy but could not fix overfitting. Vertical flipping was even less effective, yielding a lower validation accuracy than horizontal flipping. Moreover, since vertical flips are less realistic for facial expressions we chose to exclude them from further strategies. The *moderate* augmentation strategy, which included horizontal flipping, brightness shifts, and affine transformations, effectively reduced overfitting and led to strong performance across both datasets. It achieved a good balance between training and validation accuracy, especially on FER+. The *aggressive* strategy, which added Gaussian blur and coarse dropout, produced slight additional improvements. Although the gain was modest, we chose the aggressive setup due to its consistent validation advances and the assumption that coarse dropout may enhance robustness in real-world scenarios like live emotion recognition.

	FER-2013		**FER+**	
Augmentation	Train acc	Val acc	Train acc	Val acc
None	99.71	59.92	–	–
Horizontal Flip	99.23	62.20	–	–
Vertical Flip	97.74	56.97	–	–
Moderate	77.62	69.86	89.54	85.58
Aggressive	77.71	71.86	89.80	86.32

Table 1. Training and validation accuracies in percent under different data augmentation strategies for FER-2013 and FER+. Moderate: horizontal flip, brightness shift, and affine transformation. Aggressive: additionally Gaussian blur and coarse dropout. Training accuracy of last epoch. Mean validation accuracy over epochs.

4.2. Evaluation Metrics

To evaluate the model beyond overall accuracy, we report precision, recall, and F1-score per class, as well as macro-averaged values across all six emotion categories. This is especially relevant for FER+ due to class imbalance. While the model achieves a strong validation accuracy of 86.32%, this metric alone may be misleading in multi-class settings.

Precision quantifies the proportion of correctly predicted instances among all predicted instances for a class. Recall captures the proportion of correctly predicted samples among all actual samples of that class. The F1-score combines both via the harmonic mean and is considered the most robust metric in imbalanced classification tasks.

On the FER+ validation set, our model achieved:

- **Accuracy**: 86.32%
- **Macro-Precision**: 78.96%
- **Macro-Recall**: 78.09%
- **Macro-F1-score**: 78.35%
- **Validation loss**: 0.444

A detailed breakdown of performance per class is shown in Table 2. The model performs best on the majority classes *happiness* and *surprise*, with F1-scores of 0.94 and 0.87, respectively. Minority classes such as *disgust* and *fear* show lower performance.

Class	**Precision**	**Recall**	**F1-Score**
Anger	0.77	0.85	0.81
Disgust	0.70	0.56	0.62
Fear	0.60	0.64	0.62
Happiness	0.95	0.92	0.94
Sadness	0.86	0.84	0.85
Surprise	0.87	0.88	0.87

Table 2. Per-class evaluation metrics on the FER+ validation set.

Figure 5 further illustrates model performance via the confusion matrix. The most frequent misclassifications occur between *fear* and *surprise*, and between *disgust* and *anger*.

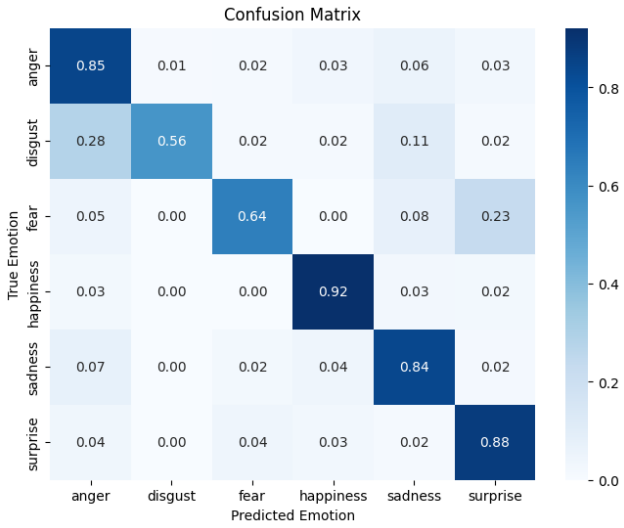

Figure 5. Confusion matrix on the FER+ validation set.

As seen in Table 2, the emotions *happiness* and *sadness* show balanced and high scores in both precision and recall. In contrast, *disgust* exhibits the weakest recall, indicating that many true instances of this class are missed. *Fear* also shows moderate performance, with more false positives likely arising from confusion with other expressions. The classes *anger* and *surprise* demonstrate particularly high recall, suggesting that they are reliably detected even under uncertain conditions.

4.3. Results

We conducted extensive training experiments on the FER-2013 and FER+ datasets using multiple data augmentation strategies. As summarized in Table 1, augmentations had a substantial impact on validation performance, especially for FER+. Horizontal flipping and affine transformations already yielded noticeable gains, while aggressive strategies further improved generalization.

In parallel, we evaluated several architectural variants to identify the most effective configuration. Our baseline CNN—composed of four convolutional layers with ReLU activations, two max-pooling layers, and a large fully connected classifier—contained approximately 34 million parameters and achieved a validation accuracy of 79.98% on FER+. Despite its size, the model showed limitations in generalization and efficiency, suggesting that model capacity alone is not sufficient for strong performance.

We also evaluated a variant with a Spatial Transformer Network (STN) at the input, inspired by Jaderberg et al. [19]. However, performance degraded to 19.23%, and training was unstable—likely due to the low input resolution (64×64), which limits the STN's ability to learn meaningful transformations.

We next introduced residual blocks to the architecture. In doing so, we restructured the network to include four custom blocks with skip connections and strided convolutions for spatial downsampling. This significantly reduced the number of parameters to 4.83 million and increased validation accuracy to 85.19%. This improvement highlights not only the effectiveness of residual learning in preventing vanishing gradients, but also the architectural benefit of gradually reducing spatial dimensions while increasing feature richness.

Building on this design, we further integrated Squeeze-and-Excitation (SE) blocks into each residual unit to enable channel-wise attention. The resulting model reached a validation accuracy of 86.32% on FER+ with only 10.02 million parameters. These results underscore that with careful architectural refinement, smaller models can outperform larger baseline variants both in accuracy and efficiency.

The training trajectory of the final model is shown in Figure 6. Accuracy increased steadily throughout training, while the loss curve exhibited minor oscillations caused by batch-level variance and augmentation effects. Thanks to batch normalization, dropout, and early stopping, the model converged stably without signs of overfitting.

Despite its compact design, the model performs on par with more complex architectures, such as those proposed by Minaee et al. [28], while maintaining real-time inference capabilities. Specifically, the model requires approximately 794 million floating point operations (MFLOPs) per 64×64 input image. Around 99.7% of the computational

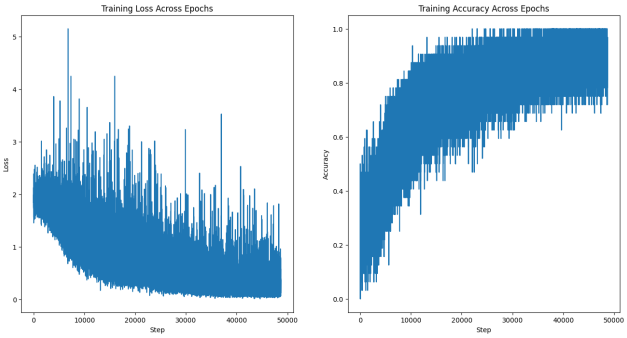

Figure 6. Training loss and accuracy across training steps.

load stems from convolutional layers, while contributions from normalization, pooling, and fully connected layers are minimal. Operations within the SE blocks (e.g., `sigmoid`, `mul`, `add_`) are not included in the FLOP count due to limitations of the analysis tool, though their impact is negligible.

4.4. Explainable AI and Demonstration

To improve the interpretability of the trained model, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize which facial regions contributed most to the model's predictions. For correctly classified samples, the Grad-CAM heatmaps consistently highlighted semantically meaningful areas such as the mouth region for *happy*, the eyebrows and forehead for *angry*, or the eyes and wrinkles for *disgust* and *fear*. Figure A1 shows representative examples of accurate predictions along with their corresponding attention maps.

To better understand the internal representation across different network depths, we also visualized intermediate activations at various convolutional stages. As illustrated in Figure A2, early layers primarily capture low-level textures and facial contours, while deeper layers become more specialized in detecting class-relevant semantic features such as muscle tension, eye shape, or mouth curvature.

In the case of misclassifications, Grad-CAM provides useful insights into the model's failure modes. Figure A3 displays an example where a *surprise* expression was incorrectly predicted as *sad*. The heatmap reveals that the model's attention is largely focused on the subject's arms and, to a lesser extent, the mouth region. This indicates that the model may be distracted by non-discriminative visual elements and fails to capture essential facial cues such as widened eyes or raised eyebrows, which are characteristic of *Surprise* expressions.

These visualizations demonstrate that, while the model achieves high classification accuracy overall, it can still attend to irrelevant or misleading regions when key facial areas are occluded or ambiguous. This highlights the importance of robustness to partial occlusions and the need for

better integration of holistic facial features—especially in real-world scenarios requiring interpretability and reliability.

Although both FER+ and RAF-DB were employed during model training and evaluation, the FER+ model proved significantly more robust and responsive in live webcam inference. Compared to RAF-DB, the FER+ dataset contains a larger variety of spontaneous expressions and greater intra-class variability, which better reflects the diversity and dynamics of real-world facial behavior [3]. In contrast, RAF-DB features more posed and controlled expressions, limiting its generalizability to natural video input. In practice, the RAF-trained model often produced unstable or erratic predictions during continuous inference. Accordingly, all live demonstrations and the webcam-based interface were based on the FER+ model.

5. Discussion

5.1. Conclusion

Our study demonstrates that high classification performance in FER can be achieved with a compact, custom CNN architecture trained entirely from scratch. Through systematic architectural refinements, including the addition of residual SE blocks, we increased validation accuracy from 58.50% over 71.86% on FER-2013 to 86.32% on FER+, while maintaining low computational costs suitable for real-time applications. These results highlight that careful network design can outperform larger, pretrained models in FER.

One of our key findings is the significant performance gap between FER-2013 and FER+, despite identical model architectures and image inputs. This underscores the important role of label quality: FER+'s revised annotations reduce ambiguity, enabling more effective learning [3]. Inaccurate annotations not only degrade performance but also risk leading the model to learn spurious or non-generalizable features [46, 47].

Notably, Grad-CAM visualizations showed that the model consistently focused on facial regions relevant to the displayed emotions, indicating that it learned semantically meaningful and human-interpretable representations. Additionally, real-time webcam testing displayed that models trained on FER+ were substantially more robust than those trained on FER-2013, further validating their practical utility in real-world scenarios.

5.2. Outlook

Nonetheless, certain challenges persist - most notably the frequent misclassification of minority emotion classes such as *disgust* and *fear*. While class weighting and augmentation mitigated some imbalance, further improvement could be achieved via targeted oversampling, alternative loss functions such as focal loss, or synthetic data generation.

One possibility for future work lies in the incorporation of synthetic data augmentation using generative adversarial networks (GANs). While results vary, Porcu et al. [32] report performance gains on the CK+ dataset using GAN-augmented samples, whereas Singh et al. [40] found no significant improvement when applying GANs to FER-2013. This suggests that the effectiveness of GAN-based augmentation may depend on dataset-specific factors such as noise level and label consistency.

Another promising direction is the combination of multiple datasets to improve the diversity and quantity of training data, especially for underrepresented emotion classes such as *disgust*. Previous studies by Zavarez et al. [49] and Kirbiz et al. [23] demonstrate that merging datasets can lead to more robust and generalizable models, further underlining the possible advantage of using FER+ instead of FER-2013.

When selecting training data, it is also essential to account for inherent dataset biases. Real-world applications of emotion recognition require training sets that reflect variation across age, gender, and ethnicity within all emotion categories. As noted by Hosseini et al. [17], mitigating bias in the training distribution may directly reduce bias in model predictions. Combining diverse datasets could be a practical approach to achieving more equitable and representative emotion classification.

To train models directly on video sequences may offer significant advantages for real-world emotion recognition. Temporal models can exploit contextual information such as gesture dynamics, facial motion trajectories, and temporal coherence of expressions. For example, Vielzeuf et al. [44] propose a multimodal architecture that processes both still frames and sequences of consecutive frames through a dedicated CNN, and augment their pipeline with audio features. This allows their system to capture emotional cues not only from static visual information but also from prosodic features such as tone and volume. Similarly, Caschera et al. [6] explore a linguistically grounded, multimodal approach that integrates speech, gestures, and textual input in addition to facial expressions. They find that facial and gestural input is especially effective for distinguishing *anger* and *happiness*, while vocal features contribute more to the classification of *sadness* and *fear*. These findings support the intuition that humans rely on a rich set of multimodal signals when interpreting emotions and that future FER systems could benefit from adopting similar strategies. Combining multimodal information could lead to more accurate, context-aware FER models suitable for deployment in real-world environments.

Overall, our findings suggest that with high-quality labels, compact models, and accurate training strategies, FER systems can achieve both high performance and real-world viability - without relying on heavy pretraining and too complex architecture.

References

[1] Mouath Aouayeb, Wassim Hamidouche, Catherine Soladie, Kidiyo Kpalma, and Renaud Seguier. Learning vision transformer with squeeze and excitation for facial expression recognition. *arXiv preprint arXiv:2107.03107*, 2021. 3

[2] M. S. T. J. Azida. Affectnet dataset [data set]. https://www.kaggle.com/datasets/mstjebashazida/affectnet. 2

[3] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 279–283, 2016. 2, 5, 8

[4] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 568–573, 2005. 2

[5] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. 5

[6] Maria Chiara Caschera, Patrizia Grifoni, and Fernando Ferri. Emotion classification from speech and text in videos using a multimodal approach. 6, 2022. 8

[7] Junkai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. Facial expression recognition based on facial components detection and hog features. In *Proceedings of the International Workshops on Electrical and Computer Engineering Subfields*, pages 884–888, Istanbul, Turkey, 2014. 2

[8] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and trends® in signal processing*, 7 (3–4):197–387, 2014. 2

[9] Hanna Drimalla, Irina Baskow, Behnoush Behnia, Stefan Roepke, and Isabel Dziobek. Imitation and recognition of facial emotions in autism: a computer vision approach. *Molecular autism*, 12(1):27, 2021. 1

[10] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971. 1, 2

[11] Australian National University Few Group. Acted facial expressions in the wild (afew) [data set]. https://users.cecs.anu.edu.au/. 2

[12] Alan J Fridlund. *Human facial expression: An evolutionary view*. Academic press, 2014. 1

[13] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. *arXiv preprint arXiv:1307.0414*, 2013. 1, 5

[14] Karol Grabowski, Agnieszka Rynkiewicz, Amandine Lassalle, Simon Baron-Cohen, Björn Schuller, Nicholas Cummins, Alice Baird, Justyna Podgórska-Bednarz, Agata Pieniażek, and Izabela Łucka. Emotional expression in psychiatric conditions: New technology for clinicians. *Psychiatry and Clinical Neurosciences*, 73(2):50–62, 2019. 1

[15] Rajana Harika, T. Uday, M. Lalitha Sirisha, M. Sri Lakshmi Sahitya, K. Drugaanjali, and M. Satya Srinivas. A review of advancements in facial emotion recognition and detection using deep learning. In *Proceedings of the 2024 International Conference on Social and Sustainable Innovations in Technology and Engineering (SASI-ITE)*, pages 290–295, 2024. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2, 3, 4

[17] Mohammad Mehdi Hosseini, Ali Pourramezan Fard, and Mohammad H. Mahoor. Faces of fairness: Examining bias in facial expression recognition datasets and models, 2025. 8

[18] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019. 2, 3, 4

[19] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer Networks, 2016. arXiv:1506.02025 [cs]. 2, 3, 7

[20] Yousif Khaireddin and Zhuofa Chen. Facial emotion recognition: State of the art performance on fer2013. *arXiv preprint arXiv:2105.03588*, (arXiv:2105.03588), 2021. 1, 2

[21] Thomas Kopalidis, Vassilios Solachidis, Nicholas Vretos, and Petros Daras. Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets. *Information*, 15(33):135, 2024. 1

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 2

[23] Serap Kırbız. Improving facial emotion recognition through dataset merging and balanced training strategies. *Journal of the Franklin Institute*, 362(7):107659, 2025. 8

[24] Hui Ma, Turgay Celik, and Heng-Chao Li. Lightweight attention convolutional neural network through network slimming for robust facial expression recognition. *Signal, Image and Video Processing*, 15(7):1507–1515, 2021. 2

[25] Masakazu Matsugu, Katsuhiko Mori, Yusuke Mitari, and Yuji Kaneda. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks*, 16(5):555–559, 2003. 2

[26] Albert Mehrabian and Susan R Ferris. Inference of attitudes from nonverbal communication in two channels. *Journal of consulting psychology*, 31(3):248, 1967. 1

[27] Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pages 117–122. IEEE, 2018. 5

[28] Shervin Minaee, Mehdi Minaei, and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(99):3046, 2021. 1, 2, 3, 7

[29] Karnati Mohan, Ayan Seal, Ondrej Krejcar, and Anis Yazidi. Fer-net: facial expression recognition using deep neural net. *Neural Computing and Applications*, 33(15):9125–9136, 2021. 1, 2, 3

[30] Uzma Nawaz, Zubair Saeed, and Kamran Atif. A novel transformer-based approach for adult's facial emotion recognition. *IEEE Access*, 13:56485–56508, 2025. 3

[31] Nicole L Nelson and James A Russell. Universality revisited. *Emotion Review*, 5(1):8–15, 2013. 1

[32] Simone Porcu, Alessandro Floris, and Luigi Atzori. Evaluation of data augmentation techniques for facial expression recognition systems. *Electronics*, 9(11):1892, 2020. 8

[33] Christopher Pramerdorfer and Martin Kampel. Facial expression recognition using convolutional neural networks: state of the art. *arXiv preprint arXiv:1612.02903*, 2016. 2

[34] M. Sambare. Fer2013 [data set]. https://www.kaggle.com/datasets/msambare/fer2013. Accessed: Jul. 03, 2025. 2

[35] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2020. 4

[36] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009. 2

[37] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 5

[38] Shuvoalok. Raf-db dataset [data set]. https://www.kaggle.com/datasets/shuvoalok/raf-dbdataset. 2

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2015. 2

[40] Shekhar Singh and Fatma Nasoz. Facial expression recognition using cnns and gans: A study on classification and generation of synthetic images. *International Journal of Information Technology*, 2025. 8

[41] Sasha Targ, Diogo Almeida, and Kevin Lyman. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*, 2016. 2

[42] Eva Tuba, Nebojša Bačanin, Ivana Strumberger, and Milan Tuba. Convolutional neural networks hyperparameters tuning. In *Artificial Intelligence: Theory and Applications*, pages 65–84. Springer, Cham, 2021. 2

[43] Saiyed Umer, Ranjeet Kumar Rout, Chiara Pero, and Michele Nappi. Facial expression recognition with trade-offs between data augmentation and deep learning features. *Journal of Ambient Intelligence and Humanized Computing*, 13(2):721–735, 2022. 5

[44] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 569–576, 2017. 8

[45] WHO. International classification of diseases 11th revision. *The Global Standard for Diagnostic Health Information.*, 2018. 1

[46] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015. 8

[47] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024. 8

[48] Afia Zafar, Muhammad Aamir, Nazri Mohd Nawi, Ali Arshad, Saman Riaz, Abdulrahman Alruban, Ashit Kumar Dutta, and Sultan Almotairi. A comparison of pooling methods for convolutional neural networks. *Applied Sciences*, 12 (17):8643, 2022. 2

[49] Marcus Vinicius Zavarez, Rodrigo F. Berriel, and Thiago Oliveira-Santos. Cross-database facial expression recognition based on fine-tuned deep convolutional network. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 405–412, 2017. 8

[50] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2014. 4

[51] Chen-Lin Zhang, Jian-Hao Luo, Xiu-Shen Wei, and Jianxin Wu. In defense of fully connected layers in visual representation transfer. In *Advances in Multimedia Information Processing – PCM 2017*, pages 807–817. Springer, Cham, 2018. 2

[52] Weiguang Zhang, Xuguang Zhang, and Yinggan Tang. Facial expression recognition based on improved residual network. *IET image processing*, 17(7):2005–2014, 2023. 2

[53] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *arXiv preprint arXiv:1512.04150*, 2015. 4

Appendix A: Additional Visualizations

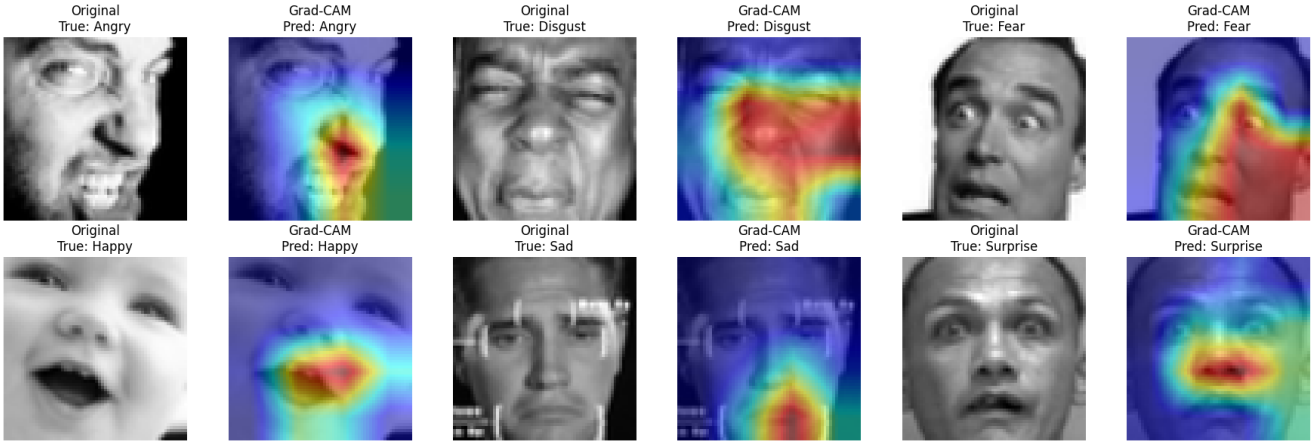

Figure A1. Grad-CAM results for correctly classified samples from various emotion classes.

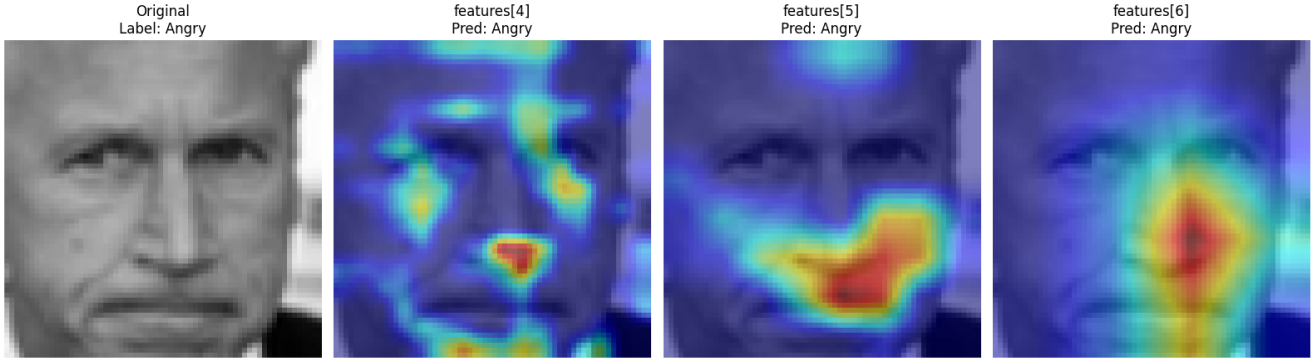

Figure A2. Grad-CAM activation maps for different layers (`features[4]`, `features[5]`, and `features[6]`) for a sample classified as *Angry*.

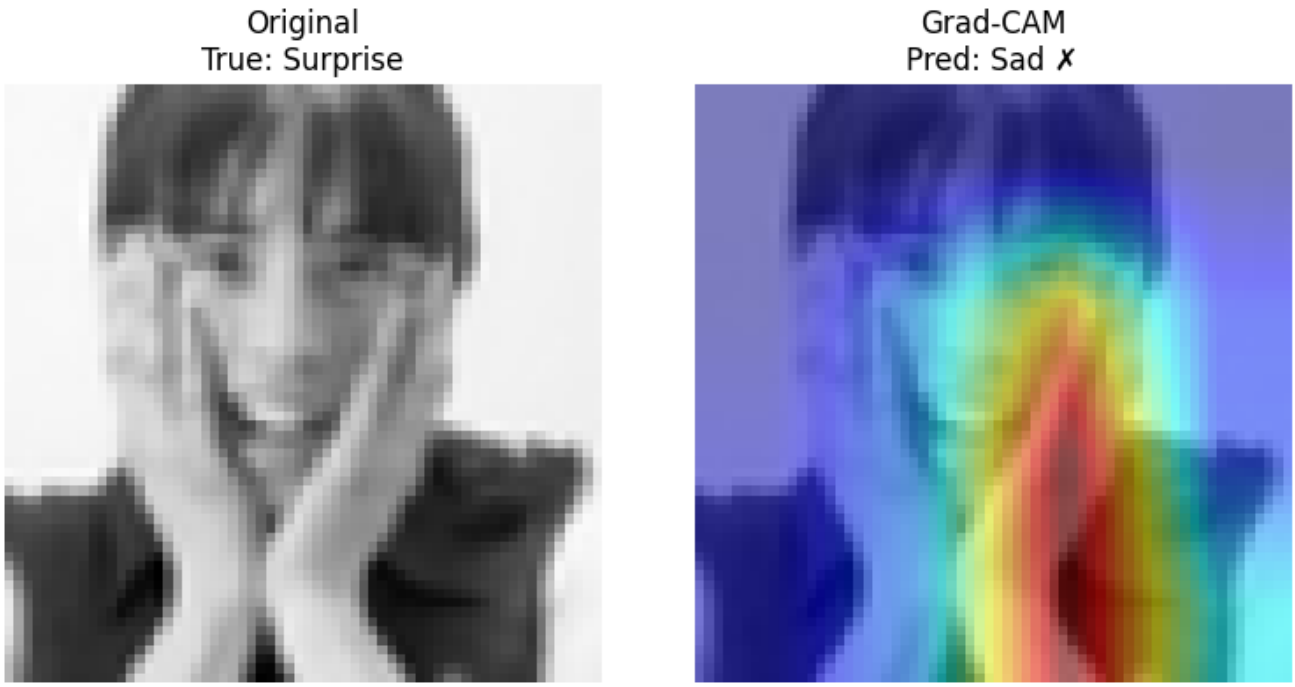

Figure A3. Grad-CAM result for a misclassified sample (*Surprise* misclassified as *Sad*). The model primarily attends to the subject's arms and mouth.

Appendix B: Contributions

Antonia Härle

- Conducted training experiments on FER-2013, systematically comparing different augmentation strategies and pooling methods (max vs. average)
- Adapted the training pipeline for FER+, performed experiments, and analyzed differences in performance across datasets
- Implemented the CSV generation pipeline for batch inference on image folders
- Finalized and corrected code, cleaned up files
- Wrote Sections: Abstract, 1 *Introduction*, 2 *Related Work*, 3.4 *Preprocessing and Augmentation*, 4.1 *Datasets and Augmentation*, 5.1 *Conclusion*
- Maintained the LaTeX document

Clara Sophie Negwer

- Did literature research
- Wrote 5.2 *Outlook*

Andre Ngo

- Designed and decided the core CNN architecture, including residual block integration
- Optimized training routine: introduced learning rate scheduling (ReduceLROnPlateau) and regularization strategies
- Performed extensive training and model tuning on FER-2013 dataset
- Developed the live webcam emotion recognition demo using OpenCV and PyTorch
- Created architectural figures (Fig. 1–4) to visualize feature extractor and classifier pipeline
- Wrote Section 3.1: *Model Architecture* and contributed to Section 3.2: *Training Setup*

Jonas Saathoff

- Designed and implemented the baseline CNN architecture and extended it with Squeeze-and-Excitation (SE) blocks
- Conducted extensive training experiments on FER-2013 and RAF-DB; performed label filtering to remove *neutral* class and align datasets
- Developed and evaluated a Spatial Transformer Network (STN) variant and analyzed failure cases
- Implemented Grad-CAM for explainability; visualized attention maps for correctly and incorrectly classified samples and across feature extractor layers
- Built the full video-based classification pipeline including preprocessing, face detection, and live inference
- Performed FLOP analysis and optimized model for low-latency real-time inference
- Wrote Sections 3.2 *Training Setup*, 3.3 *Grad-CAM Architecture and Integration*, 3.5 *Video Demo*, 4.2 *Evaluation Metrics*, 4.3 *Results*, 4.4 *Explainable AI and Demonstration*