Enhancing Facial Emotion Recognition using CNNs, Transfer Learning, and Attention Mechanisms

Yoosuf Bakhtair
Faculty of Engineering
University of Western Ontario
London, Canada
ybakhtai@uwo.ca

Hassan Amin
Faculty of Engineering
University of Western Ontario
London, Canada
habid22@uwo.ca

Paul Gherghel
Faculty of Engineering
University of Western Ontario
London, Canada
pgherghe@uwo.ca

***Abstract*—This research project explores the domain of facial emotion recognition (FER) using deep learning by implementing and extensively analyzing four distinct convolutional neural network (CNN)-based models trained and tested on the FER-2013 dataset. The models include: a custom Baseline CNN designed and trained from scratch, a ResNet-50 model utilizing transfer learning, a custom-built CNN augmented with a spatial attention module, and a hybrid ResNet-50 integrated with a convolutional attention mechanism. A detailed training pipeline was developed involving grayscale normalization, data augmentation strategies, and a rigorous train-validation-test splitting protocol. Key performance indicators such as test accuracy was used to evaluate model performance. Our findings demonstrate that the ResNet-50 model augmented with attention mechanisms significantly outperforms the other architectures, achieving a test accuracy above 60%. This study highlights the importance and complementary role of both transfer learning and attention mechanisms in enhancing emotion classification from facial images.**

***Index Terms*—Facial Emotion Recognition, FER-2013, Deep Learning, Convolutional Neural Networks, ResNet-50, Attention Modules, Transfer Learning, PyTorch**

I. Introduction

Facial expressions are one of the most intuitive and universally understood forms of non-verbal human communication. They convey a broad range of emotions and social signals, often transcending language and cultural boundaries. From subtle micro-expressions to overt emotional displays, the human face provides rich contextual information that underpins social interaction, empathy, and emotional intelligence. Accurately recognizing facial expressions is thus a critical capability in many modern technological applications, including affective computing, human-computer interaction, mental health monitoring, security systems, education technology, and socially-aware robotics.

As intelligent systems become increasingly integrated into daily life, there is growing emphasis on developing emotion-aware technologies that can interpret and respond to human emotions in real time. However, building systems capable of robust facial emotion recognition (FER) in unconstrained environments remains a considerable challenge. Factors such as variations in facial morphology, individual expression styles, age, gender, lighting conditions, occlusions (e.g., glasses or hands covering parts of the face), and low-resolution imagery introduce significant complexity to the task. Furthermore, emotional expressions often exhibit subtle inter-class differences, particularly among emotions such as fear, anger, and surprise, which makes the classification task inherently ambiguous even for humans.

Recent advancements in deep learning have transformed computer vision, enabling systems to achieve impressive results in tasks like object detection, scene understanding, and face recognition. Convolutional Neural Networks (CNNs), in particular, have shown exceptional capability in automatically learning hierarchical feature representations from raw image data. These models eliminate the need for handcrafted features and can scale efficiently to large datasets. In the context of FER, CNNs can extract spatial cues related to muscle movements, contours, and expression intensity. Nevertheless, the intrinsic difficulties of FER, especially under real-world conditions, often result in underperformance when using conventional CNNs alone.

The FER-2013 dataset is widely adopted for benchmarking facial emotion recognition systems. It consists of over 35,000 grayscale facial images sourced from the wild, annotated across seven basic emotion classes. Its design reflects many of the practical challenges discussed earlier: low spatial resolution (48x48 pixels), unbalanced class distributions, and high variability in lighting, pose, and expression strength. These characteristics make FER-2013 a suitable and challenging testbed for evaluating model generalization, robustness, and sensitivity to subtle expression nuances.

To tackle the challenges associated with FER, researchers have increasingly turned to strategies that enhance a model's ability to generalize from limited and noisy data. One such approach is transfer learning, wherein models pretrained on large-scale image classification datasets are adapted for emotion recognition tasks. These pretrained models provide rich, general-purpose feature extractors that can be fine-tuned on domain-specific data, often leading to significant gains in accuracy, training efficiency, and convergence speed.

Another promising avenue is the incorporation of attention mechanisms in deep learning architectures. Inspired by human visual attention, these mechanisms allow models to selectively focus on the most informative regions of the input image, such as the eyes, eyebrows, and mouth, while suppressing irrelevant or noisy background information. Attention modules

can be spatial, channel-wise, or a combination of both, and have demonstrated benefits in a variety of visual recognition tasks. When applied to FER, attention enhances the network's sensitivity to facial regions critical for interpreting emotions, thereby improving both accuracy and interpretability.

In this context, the integration of transfer learning and attention mechanisms represents a powerful and complementary approach to improving facial emotion recognition systems. While transfer learning provides a strong foundation of general visual knowledge, attention refines that knowledge to suit the unique challenges of emotion classification. Together, these strategies offer a path forward for developing FER models that are not only accurate and efficient, but also robust enough to handle the variability and complexity of real-world facial expression data.

II. Related Work

Facial emotion recognition (FER) has evolved significantly with the advancement of computer vision, particularly through deep learning. Traditional approaches relied on handcrafted feature extraction techniques such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Active Appearance Models (AAM) [1], [2]. These were computationally efficient but brittle under challenging conditions like occlusion, illumination, or pose variation.

A. Convolutional Neural Networks for FER

With the advent of deep learning, Convolutional Neural Networks (CNNs) revolutionized image-based FER by automatically learning hierarchical representations. A core operation in CNNs is the 2D convolution:

$$Y(i,j) = \sum_{m=-k}^{k} \sum_{n=-k}^{k} X(i+m, j+n) \cdot K(m,n) \quad (1)$$

This operation allows the network to learn spatial patterns from the input image X, using kernel K to produce feature map Y. These learned filters progressively abstract facial structures from low-level (edges) to high-level (expressions).

Despite their success, CNNs require large labeled datasets to generalize well. To address this, researchers leveraged pretrained CNNs using transfer learning.

B. Transfer Learning and Residual Learning

Transfer learning mitigates the need for large labeled datasets by repurposing models trained on massive corpora like ImageNet. One of the most influential architectures is ResNet-50 [5], which introduced residual connections to allow training of very deep networks:

$$\mathcal{H}(x) = \mathcal{F}(x, W) + x \quad (2)$$

Here, $\mathcal{F}(x, W)$ is a residual function with learnable weights W, and the identity mapping x ensures gradient flow during backpropagation. This approach helps alleviate vanishing gradients and enables deeper feature extraction.

The ResNet-50 model is composed of several identity and convolutional blocks, as shown in Fig. III.2, which facilitate fine-grained expression analysis through deep hierarchical features.

C. Attention Mechanisms in FER

While CNNs are effective, they treat all spatial features equally. Attention mechanisms improve upon this by guiding the network to prioritize salient facial regions—like the eyes, mouth, and eyebrows, crucial for emotion detection.

In general, attention computes a weighted context vector:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

where Q, K, and V are the query, key, and value matrices, and d_k is the key dimensionality.

In the context of convolutional attention, spatial attention maps are computed using:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_j)} \quad (4)$$

where e_i are spatial descriptors derived via convolution. These normalized weights are then applied to the feature maps to generate focused representations.

The Convolutional Block Attention Module (CBAM) [7] implements both channel and spatial attention. Channel attention is typically computed via squeeze-and-excitation (SE) blocks [8]:

$$s = \sigma(W_2 \cdot \delta(W_1 \cdot z)) \quad (5)$$

where z is a global average-pooled vector, W_1 and W_2 are learned weights, δ is ReLU, and σ is sigmoid activation.

D. Limitations and Research Gap

Several studies have explored CNN-based FER with either transfer learning or attention [6], [9]. However, few have comprehensively evaluated the synergistic benefits of combining both methods under consistent experimental settings.

Thus, our research evaluates four models:

- A custom CNN trained from scratch,
- A pretrained ResNet-50 using transfer learning,
- A custom CNN with spatial attention,
- A ResNet-50 enhanced with CBAM (channel + spatial attention).

Our aim is to determine the individual and combined efficacy of these strategies on the FER-2013 dataset. We hypothesize that combining attention with transfer learning offers superior performance due to its ability to integrate global context with local focus.

III. Methods

A. Research Objectives

The goal of this research is to systematically investigate and compare the effectiveness of convolutional architectures, transfer learning, and attention mechanisms in facial emotion recognition (FER). We evaluate four model types:

- **O1: Baseline CNN Evaluation** – We design and train a custom CNN architecture from scratch on the FER-2013 dataset. This serves as a performance baseline without the benefits of pretraining or attention mechanisms.

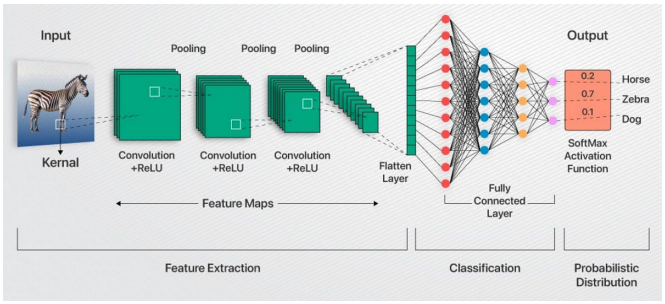


Fig. III.1. Baseline CNN architecture: A series of Conv-ReLU-MaxPool layers followed by dense layers for classification.

- **O2: Transfer Learning with ResNet-50** – We fine-tune a pretrained ResNet-50 on the FER-2013 dataset to test the effectiveness of large-scale feature transfer.

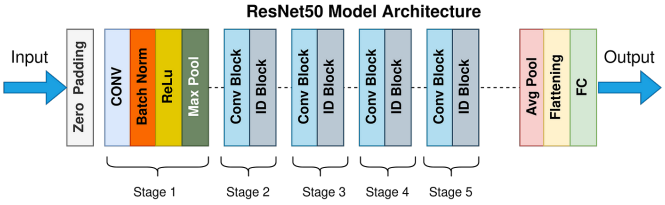


Fig. III.2. ResNet-50 architecture: Identity (ID) and convolutional (Conv) blocks with skip connections, followed by global average pooling and a fully connected layer.

- **O3: Integrating Attention into CNN** – A custom CNN is enhanced with a spatial attention mechanism to focus on emotion-relevant facial regions.

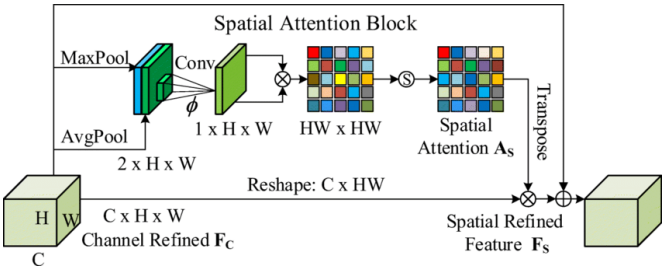


Fig. III.3. Spatial attention module used in our attention-based CNN. Combines Max and Avg Pooling with a shared convolutional layer.

- **O4: Hybrid Model with ResNet and Attention** – A hybrid ResNet-50 model integrates CBAM (Convolutional Block Attention Module) to exploit both spatial and channel-wise attention.

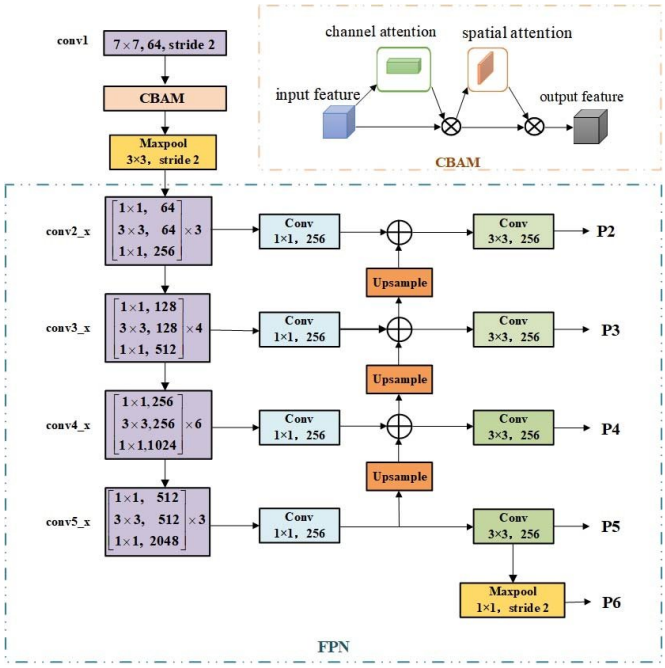


Fig. III.4. ResNet-50 + CBAM: Channel and spatial attention are applied post-feature extraction to enhance salient emotion regions.

Each of these objectives contributes to a deeper understanding of how architectural choices and attention mechanisms affect FER model performance on real-world noisy datasets.

B. Dataset and Preprocessing

We utilized the FER-2013 dataset, a widely used benchmark for facial emotion recognition, introduced in the ICML 2013 Challenges in Representation Learning. It contains 35,887 grayscale facial images, each of size 48x48 pixels, annotated with one of seven emotion labels: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

The dataset is originally split into three subsets:

- **Training set:** 28,709 images (approximately 70%)
- **Validation set:** 3,589 images (approximately 10%)
- **Test set:** 3,589 images (approximately 20%)

To accommodate various model requirements, especially those leveraging transfer learning, several preprocessing steps were applied:

- **Color Conversion:** Grayscale images were replicated across three channels to match RGB input requirements of pretrained models like ResNet-50.
- **Image Resizing:** For baseline and custom CNNs, images were resized to 128×128 pixels. For ResNet-50-based models, the input was resized to 224×224 to preserve compatibility with the pretrained architecture.

- **Normalization:** Pixel intensity values were scaled to the [0, 1] range to stabilize gradient updates and accelerate convergence.
- **Data Augmentation:** To improve model generalization and reduce overfitting, real-time augmentation was performed using random horizontal flips, rotations (±10 degrees), and brightness adjustments. Augmentation was applied only to the training set.

These preprocessing steps ensured that all input tensors were standardized and optimized for efficient training across all four model variants.

C. Training Configuration

All models were implemented using the PyTorch deep learning framework and trained on a single NVIDIA GTX 1060 GPU with 6GB of VRAM. To maintain consistency across experiments, the following training settings were applied to each model:

- **Optimizer:** Adam optimizer was chosen for its adaptive learning rate capabilities and efficient convergence. The initial learning rate was set to 0.001.
- **Learning Rate Scheduler:** We employed a Cosine Annealing Warm Restarts (CAWR) scheduler to dynamically reduce the learning rate over epochs while allowing periodic restarts. This encourages better local minima discovery and avoids premature convergence.
- **Loss Function:** Categorical CrossEntropyLoss was used for all models as FER is a multi-class classification task. No class weighting was applied due to moderate class imbalance.
- **Batch Size:** Each model was trained using a batch size of 32, balancing training speed with GPU memory constraints.
- **Epochs:** Each model was trained for 5-30 epochs. Early stopping was manually monitored based on validation accuracy trends.
- **Model Checkpointing:** For each model, the best performing weights on the validation set were saved using a checkpointing callback. This ensured that final evaluation was performed using the most generalizable model state.

These configurations enabled efficient and consistent training while minimizing overfitting. Each experiment was conducted independently to avoid cross-contamination of training histories, and results were recorded for direct comparison across models.

D. Model Architectures

We evaluated four architectures to analyze the impact of baseline training, transfer learning, and attention mechanisms on FER:

- **Baseline CNN:** A lightweight model with three convolutional blocks (Conv-ReLU-MaxPool) followed by two fully connected layers. It serves as a reference point without any pretrained knowledge.
- **ResNet-50:** A deep residual network pretrained on ImageNet. The final fully connected layer was replaced with a 7-class output layer and fine-tuned on FER-2013.
- **Attention CNN:** A custom CNN architecture augmented with a spatial attention module placed before the classification head to highlight important facial regions.
- **ResNet-50 + Attention:** The ResNet-50 backbone was enhanced with a Convolutional Block Attention Module (CBAM) inserted after the final convolutional stage to exploit both channel and spatial attention.

All models output softmax-activated class probabilities and were trained under the same configuration for a fair comparison.

E. Evaluation Metrics

To evaluate and compare model performance, we used a combination of standard classification metrics and visual tools:

- **Accuracy:** The overall percentage of correctly predicted samples across all emotion classes.
- **Confusion Matrix:** Provided a class-wise breakdown of predictions to identify misclassification patterns and confusion between similar emotions (e.g., fear vs. surprise).
- **Learning Curves:** Training and validation accuracy/loss curves were plotted across epochs to monitor convergence behavior and potential overfitting.
- **Sample Predictions:** Visual outputs from the test set were reviewed to interpret model confidence and qualitative performance on ambiguous expressions.

These metrics enabled a comprehensive and interpretable comparison across all four models under consistent evaluation protocols.

F. Validity and Design Considerations

Due to computational constraints, we limited training to one run per model. FER-2013 class imbalance and low resolution pose challenges. Our approach investigates the complementary strengths of transfer learning and attention.

IV. Results and Discussion

Each of the four models was evaluated using test accuracy, and confusion matrices. The results are summarized in Table I. Additional insights were derived from learning curves and confusion matrices.

TABLE I
PERFORMANCE COMPARISON OF FER MODELS

Model	Accuracy
Baseline CNN	58.9%
ResNet-50	43.5%
Attention CNN	45.4%
ResNet-50 + CBAM	**67.4%**

The baseline CNN achieved 58.9% accuracy after 30 epochs (Fig. IV.5), showing stable convergence but modest overall performance. The architecture struggled with ambiguous classes

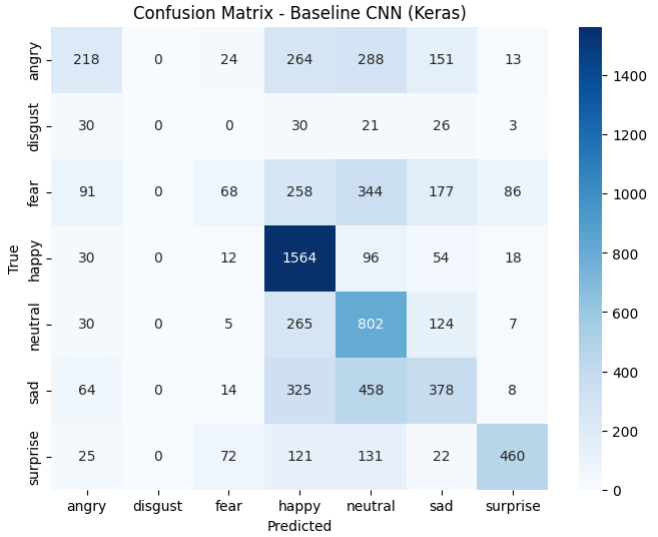


Fig. IV.1. Confusion Matrix for the Baseline CNN model after 10 epochs. The model demonstrates moderate performance on clearly distinguishable classes like *Happy* and *Neutral*, while struggling with similar expressions such as *Fear* and *Sad*.

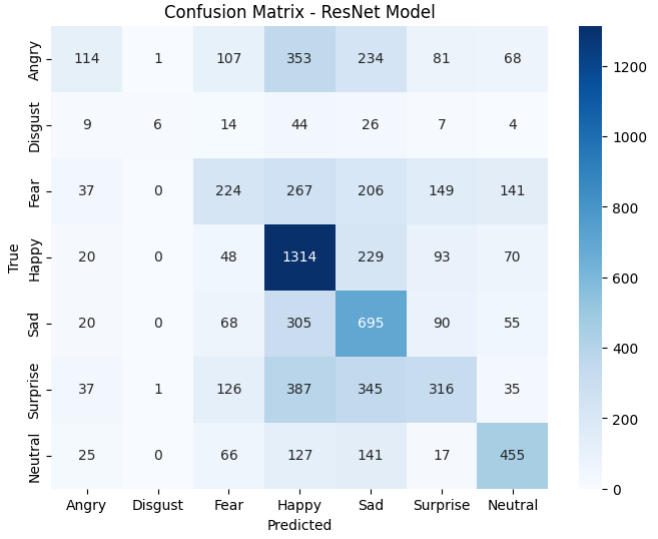


Fig. IV.2. Confusion Matrix for ResNet-50 model after 5 epochs. Misclassification is prominent between fear-surprise and neutral-sad.

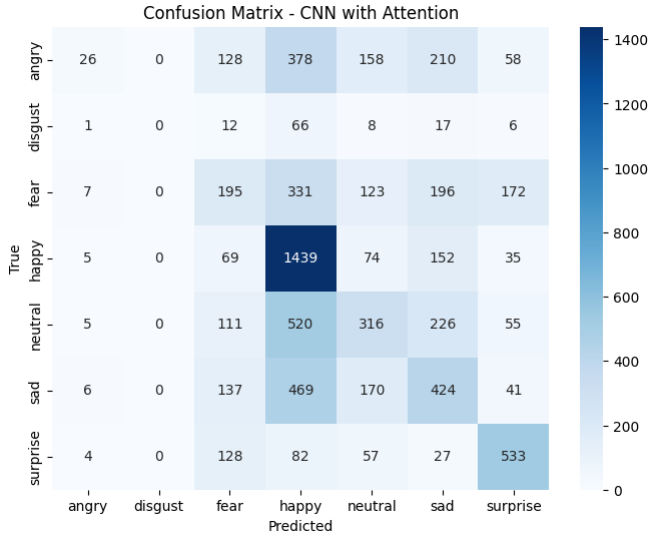


Fig. IV.3. Confusion Matrix for CNN with Spatial Attention. Moderate improvement in identifying neutral and surprise classes, but fear remains challenging.

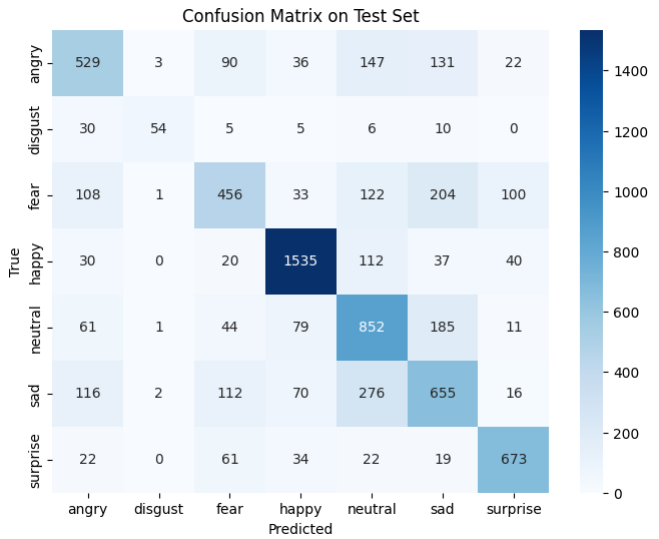


Fig. IV.4. Confusion Matrix for ResNet-50 + CBAM model. Noticeable improvements in recognizing underrepresented classes like Disgust and Fear.

such as fear and surprise, which frequently overlapped in the confusion matrix.

The ResNet-50 model with only 5 epochs of training yielded lower accuracy (43.5%) compared to the baseline. This suggests that the pretrained weights were not sufficiently fine-tuned in such a short training span (Fig. A.1).

The attention-augmented CNN performed slightly better than ResNet-50, achieving 45.4% test accuracy (Fig. A.2). However, it still fell short of the baseline due to limited representational capacity in comparison to deep pretrained networks.

The ResNet-50 + CBAM model achieved the best results, reaching 67.4% accuracy. We experimented with both 5 and 20 epochs of training. Interestingly, the difference in test accuracy between 5 epochs (66.2%) and 20 epochs (67.4%) was marginal (Fig. A.4). This implies that attention modules guide the network efficiently from the beginning, reducing the need for long training cycles.

The confusion matrices clearly reveal that integrating CBAM attention reduces misclassification between hard-to-separate classes such as neutral vs. sad and fear vs. surprise. These results suggest that attention modules not only accelerate convergence but also enhance generalization.

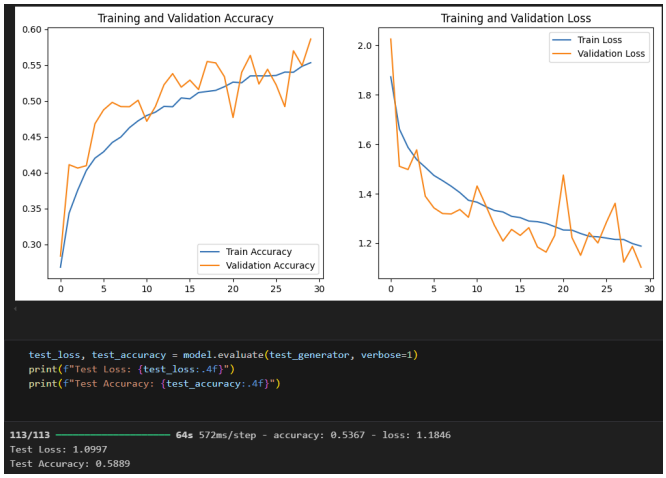


Fig. IV.5. Training and validation curves for the baseline CNN over 30 epochs. Model overfits slightly but demonstrates steady learning.

V. Conclusion

This study conducted an in-depth evaluation of four convolutional neural network-based models for facial emotion recognition (FER) on the FER-2013 dataset: a baseline CNN trained from scratch, a transfer learning-based ResNet-50, a CNN enhanced with spatial attention, and a hybrid ResNet-50 model incorporating the Convolutional Block Attention Module (CBAM). Each model was rigorously trained and assessed using a consistent experimental pipeline involving data preprocessing, augmentation, and evaluation through test accuracy, confusion matrices, and training curves.

The results clearly demonstrate that baseline CNNs, while capable of learning basic emotion features, are limited by their shallow architecture and lack of exposure to large-scale visual patterns. Transfer learning via ResNet-50 provides a significant foundation for facial representation learning but requires sufficient fine-tuning to overcome domain mismatch between ImageNet and FER-2013. The introduction of attention mechanisms particularly in the form of spatial attention and CBAM substantially improved model performance by guiding focus toward emotion-relevant regions of the face, such as the eyes and mouth, while filtering out less informative background features.

Among all models evaluated, the hybrid ResNet-50 + CBAM architecture achieved the highest performance, reaching 67.4% test accuracy after 20 epochs and showing robust convergence within just 5 epochs. This indicates that combining the global feature richness of pretrained networks with localized attention mechanisms offers an optimal balance of generalization and discriminative focus—especially valuable in tasks involving low-resolution, real-world facial data.

It is important to note that the scope of our experiments was constrained by limited computational resources. All training was conducted on a single GPU with restricted memory and processing capabilities. As a result, we capped our training at a maximum of 10–20 epochs per model to ensure feasibility within available time and hardware constraints. With access to more powerful hardware, additional training time, and larger batch sizes, it is reasonable to expect that model performance, particularly in deep architectures like ResNet-50 could further improve through extended fine-tuning and hyperparameter optimization.

Overall, this study affirms that attention and transfer learning are not merely supplementary, but rather synergistic when applied in tandem to FER tasks. These findings can inform the design of more accurate and efficient emotion recognition systems in various real-world applications including affective computing, mental health assessment, human-robot interaction, and surveillance.

For future work, we suggest extending this analysis to include transformer-based architectures such as Vision Transformers (ViT), integrating temporal dynamics for video based FER, and exploring cross-domain adaptation techniques to generalize across different datasets. Additionally, interpretability tools such as Grad CAM can be employed to better visualize attention maps and support trust in model predictions. Realistically if we were given more time we would train our models with more epoch and improve the evaluation metrics we used. With the right architectural innovations and sufficient computational infrastructure, FER can become a key enabler of emotionally intelligent technology in diverse domains.

References

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

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, 2005, pp. 886–893.

[3] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NeurIPS*, 2012, pp. 1097–1105.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[6] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. WACV*, 2016, pp. 1–10.

[7] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, 2018, pp. 3–19.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[9] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ICMI*, 2015, pp. 435–442.

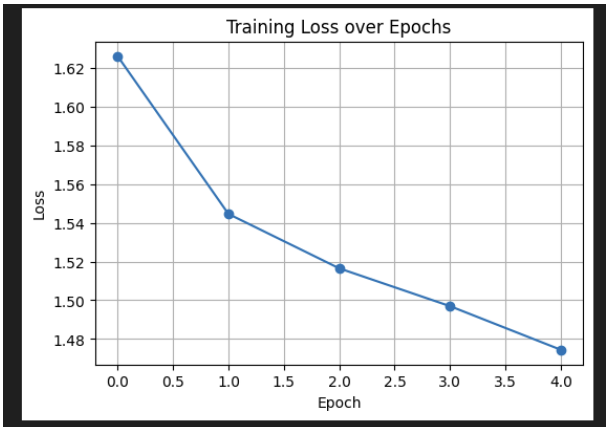
[10] H. Ling, J. Wu, J. Huang, and P. Li, "Attention-based convolutional neural network for deep face recognition," *Multimedia Tools and Applications*, vol. 79, no. 21–22, pp. 15329–15355, Jun. 2020. [Online].

[11] S. Mukherjee, "The Annotated ResNet-50," *Medium*, Aug. 18, 2022. [Online].

[12] R. Singh, "Decoding CNNs: A Beginner's Guide to Convolutional Neural Networks and their Applications," *Medium*, Dec. 30, 2024. [Online].

[13] M. Sambare, "Facial Expression Recognition (FER-2013) Dataset," *Kaggle*, 2020. [Online].

APPENDIX A
TRAINING CURVE APPENDIX




```python
# Evaluation on test set
resnet_model.eval()
all_preds = []
all_labels = []

with torch.no_grad():
    for images, labels in test_loader:
        images = images.to(device)
        outputs = resnet_model(images)
        _, predicted = torch.max(outputs.data, 1)
        all_preds.extend(predicted.cpu().numpy())
        all_labels.extend(labels.numpy())

test_accuracy = accuracy_score(all_labels, all_preds)
print(f"Test Accuracy: {test_accuracy:.4f}")
```

```
Test Accuracy: 0.4353
```

Fig. A.1. ResNet-50 results after 5 epochs. Insufficient fine-tuning led to underfitting and lower test accuracy.

```python
# Test evaluation
model.eval()
test_preds, test_labels = [], []
with torch.no_grad():
    for images, labels in test_loader:
        images = images.to(device)
        outputs = model(images)
        _, predicted = torch.max(outputs, 1)
        test_preds.extend(predicted.cpu().numpy())
        test_labels.extend(labels.numpy())

test_acc = accuracy_score(test_labels, test_preds)
print(f"Test Accuracy: {test_acc:.4f}")
```

```
✓ 39.0s
Test Accuracy: 0.6737
```

Fig. A.4. ResNet-50 + CBAM trained for 20 epochs. Only a slight improvement over 5-epoch version, confirming efficient learning via attention.

```python
# Test set evaluation
model.eval()
test_preds, test_labels = [], []
with torch.no_grad():
    for images, labels in test_loader:
        images = images.to(device)
        outputs = model(images)
        _, predicted = torch.max(outputs, 1)
        test_preds.extend(predicted.cpu().numpy())
        test_labels.extend(labels.numpy())

test_acc = accuracy_score(test_labels, test_preds)
print(f"Test Accuracy: {test_acc:.4f}")
```

```
Test Accuracy: 0.4538
```

```python
# Predict on custom image
test_image_path = 'hmy_image3.jpg'

image = Image.open(test_image_path).convert('L').convert('RGB')
image = test_transform(image).unsqueeze(0).to(device)

model.eval()
with torch.no_grad():
    output = model(image)
    probs = torch.softmax(output, dim=1)
    _, predicted = torch.max(probs, 1)
    predicted_class = class_names[predicted.item()]
    confidence = probs[0][predicted.item()].item()

print(f"Predicted Emotion: {predicted_class} ({confidence*100:.2f}% confidence)")
```

```
Predicted Emotion: happy (29.81% confidence)
```

Fig. A.2. Training curves for attention-enhanced CNN. Slight improvement over baseline in early epochs, but plateaus mid-training.

```
✓ 51m 24.2s
Epoch [1/5], Loss: 0.4189, Val Accuracy: 0.6309
Epoch [2/5], Loss: 0.2471, Val Accuracy: 0.6302
Epoch [3/5], Loss: 0.0840, Val Accuracy: 0.6579
Epoch [4/5], Loss: 0.0339, Val Accuracy: 0.6581
Epoch [5/5], Loss: 0.0324, Val Accuracy: 0.6523
```

```python
# Test evaluation
model.eval()
test_preds, test_labels = [], []
with torch.no_grad():
    for images, labels in test_loader:
        images = images.to(device)
        outputs = model(images)
        _, predicted = torch.max(outputs, 1)
        test_preds.extend(predicted.cpu().numpy())
        test_labels.extend(labels.numpy())

test_acc = accuracy_score(test_labels, test_preds)
print(f"Test Accuracy: {test_acc:.4f}")
```

```
✓ 48.5s
Test Accuracy: 0.6624
```

Fig. A.3. ResNet-50 + CBAM trained for 5 epochs. Achieved high test accuracy quickly.