**Deep Learning-Based Emotion Recognition via ResNet50 with Attention and Transformer Layers**

**Individual report: Hridi Prova Debnath**

*From Best Model to Image and Video Demo: A Comprehensive Exploration of*

*Real-Time Facial Emotion Recognition*

**1. Introduction**

Facial emotion recognition (FER) has become a pillar technology in contemporary human-computer interaction (HCI), mental health diagnosis, and affective computing. By interpreting facial expressions, machines are able to infer human emotions, making applications vary from empathetic AI assistants to clinical diagnostic tools for mood disorder detection. This report documents the process of porting a state-of-the-art Hybrid ResNet50 model—trained on the FER2013 dataset—into a real-time image and video demonstration system. The strategy was to transcend technical obstacles such as class imbalance, low-resolution inputs, and computational latency and maintain high accuracy across diverse emotional expressions.

The FER2013 dataset, although widely utilized, is particularly problematic: its 48x48 grayscale images contain less resolution than contemporary deep learning systems require, and its severe class imbalance (e.g., 547 "disgust" examples vs. 8,989 "happy" examples) skews model predictions. Our method is the combination of a Hybrid ResNet50 backbone with spatial-channel attention and transformer blocks, trained using a progressive training procedure. The technical considerations, implementation challenges, and performance results encountered in transitioning from a static model to a dynamic, real-world environment are discussed in this report.

**2. Background and Objectives**

**2.1 The Challenge of Real-World Emotion Recognition**

Emotion detection in unconstrained environments is much more different from lab environments. Varying illumination, face occlusions, and head orientations make feature extraction harder, and cultural differences in expressing emotions introduce labeling ambiguities. The FER2013 database, while commonly used, magnifies these challenges: its grayscale low-resolution images conceal subtle fine-grained features like micro-expressions (e.g., subtle lip twitches in "sadness"), and its class distribution mirrors real-world emotional frequency skew, with "happy" faces being most frequent.

**2.2 Objectives**

The primary aim was to miniaturize the Hybrid ResNet50 model—selected for high offline performance—into a real-time image and video processing framework. The principal goals were:

1.Input Adaptation: Design a preprocessing pipeline that can process FER2013's 48x48 grayscale images in such a way that it is backward compatible with the 224x224 RGB input size of ResNet50.

2.Class Imbalance Reduction: Employ methods to improve identification of minority classes like "disgust" and "fear."

3. Real-Time Optimization: Make the system run video streams at ≥24 FPS with ≤50 ms latency, suitable for interactive applications.

4. User-Centric Visualization: Develop a user-friendly interface displaying emotion labels, confidence values, and face bounding boxes.

**3. Methodology**

**3.1 Architectural Foundations: The Hybrid ResNet50 Model**

The Hybrid ResNet50 architecture integrates three innovations to counter FER2013's limitations:

**3.1.1 Residual Learning Backbone** ResNet50's residual blocks resolve the vanishing gradient issue through skip connections, allowing the network to learn identity functions when deeper layers offer no advantage. This structure allows for stable training of the 50-layer network, which is crucial for learning hierarchical features from edges to facial contours.

**3.1.2 Spatial-Channel Attention (SCA)** To compensate for information loss through low-resolution inputs, double attention mechanisms sharpen feature maps:

• Channel Attention: Squeeze-and-Excitation (SE) blocks adaptively reweight channel-wise responses. Channels capturing important features (e.g., mouth contours in "happy") are highlighted.

• Spatial Attention: A convolutional layer generates a spatial mask identifying regions like eyebrows (relevant to "anger") or lip corners (indicative of "sadness").

**3.1.3 Transformer-Augmented Context Modeling** A multi-head self-attention (MHSA) block learns long-range relations between facial landmarks. For instance, it associates furrowed brows with pursed lips in "anger," overcoming the constraint of convolutional layers, which are local-receptive-field sensitive. Positional encodings preserve spatial relations during processing of flattened feature maps.

**3.1.4 Class-Specific Branches** Separate dense clusters for emotion classes (e.g., "happy"/"surprise" vs. "disgust"/"fear") to reduce cross-class interference. This arrangement forces the model to learn distinct features for minority classes, improving their recall.

**3.2 Progressive Training Strategy**

**Training was conducted in three phases to balance stability and flexibility:**

1. Feature Extraction (Frozen Backbone): Initial training focused on attention and transformer blocks with frozen ResNet50 layers. This preserved pretrained ImageNet features (edges, textures) while fine-tuning the custom head to FER2013's emotion-specific patterns.

2. Gradual Fine-Tuning: Layers were unfrozen progressively—starting with the bottommost blocks (closest to the head)—to optimize high-level features like facial contours. Learning rates were reduced (1e-5 → 1e-6) to avoid overwriting valuable pretrained knowledge.

3. Head Optimization: The ResNet50 backbone was frozen again, and the classifier layers were fine-tuned with dropout (p=0.3) and L2 regularization (λ=1e-4) to prevent overfitting.

**3.3 Data Preprocessing and Augmentation**

To prepare the low-resolution, grayscale FER2013 dataset images for support, a series of preprocessing steps was conducted to convert the data into ResNet50 architecture-compatible data. The initial 48×48 grayscale images were upscaled to 224×224 resolution via bicubic interpolation. This strikes a balance between preserving essential facial features and maintaining computational efficiency.

Since ResNet50 takes three-channel RGB input, mere replication of the grayscale channel would be insufficient. Instead, a learnable 1×1 convolutional layer was utilized to project the single-channel images into the three-channel space. In contrast to simple replication, this layer learned to optimize the grayscale-to-RGB mapping such that it could be more accurately aligned with the pretrained ResNet50 filters derived from ImageNet.

Aside from input adaptation, large-scale data augmentation was applied to enhance variability and generalization. Geometric transformations such as rotations of up to ±35 degrees and shear distortions by 0.4 simulated varying head poses and perspective changes. Horizontal flipping was particularly effective for expressions like "happy" and "surprise," where facial symmetry plays a big role. Photometric augmentations were also included: brightness adjustments of 0.5 to 1.5 simulated changing lighting conditions, and zoom transformations of up to 40% helped the model cope with changing face-to-camera distances.

Class imbalance inherent in FER2013 was a major concern, especially for underclass classes like "disgust." Inverse-frequency weighting was included in the loss function to address this. Weights were calculated based on the total samples, class count, and samples per class. For instance, the disgust class, with a mere 547 instances, had a loss weight of 7.31 and therefore magnified its gradient contribution to training. Counterbalancing this, focused data augmentation was employed. Minority classes were oversampled and subjected to more aggressive transformations—e.g., ±25-degree rotations for "disgust"—to inflate their representation artificially and improve recall.

**4. Implementation: Bridging Model and Application**

**4.1 Image Demonstration Pipeline**

The image-based emotion recognition subsystem has a four-stage pipeline, which is specifically designed to be used with the FER2013 dataset. During the first stage, preprocessing, each gray input image is resized from the original 48×48 to the 224×224 input consumed by ResNet50. This single-channel image is then fed to a learnable 1×1 convolutional layer that transforms it into a three-channel pseudo-RGB representation for compatibility with pre-trained ResNet filters. The resulting tensor is normalized to the range [-1, 1] using ResNet's internal preprocessing function.

The second step is inference using the trained Hybrid ResNet50 model. The model outputs a probability distribution over the seven emotion classes for any given input image. Predictions for minority classes such as "disgust" and "fear" are boosted by class weights during training and are thus recalled better.

At the visualization phase, OpenCV operations mark the input image with bounding boxes around detected faces. The boxes are colored according to the predicted emotion (red for "angry," say) and overlaid with a label and confidence rating corresponding to the model's prediction.

Finally, during the analysis stage, per-class recall and confusion matrices are generated to see how well the model performs for each emotion. This is especially useful in seeing the performance improvement for classes that were historically underrepresented or were in danger of being misclassified.

**4.2 Real-Time Video Demonstration**

**4.2.1 System Architecture**

For real-time emotion recognition using video, a webcam stream is processed in real time. Face detection is accomplished through a Haar Cascade classifier processing grayscale frames. As a lightweight solution, the system can be executed at 24 frames per second (FPS) on mid-range GPUs and can be deployed in real time.

For enhanced robustness to illumination changes, adaptive histogram equalization is performed on each frame prior to face cropping. Rescaled cropped faces are rescaled to 224×224 and passed through the identical 1×1 convolutional layer for pseudo-RGB conversion. Preprocessing ensures standard input formatting to the emotion recognition model.

The emotion inference engine predicts one face per frame with the Hybrid ResNet50 model. TensorFlow graph optimizations enable the model to execute with approximately 50 milliseconds latency per frame. For smoothing prediction noise and output flicker minimization, a sliding window of the last 10 predicted labels is maintained. The most common label in this window is taken as the final output to be shown to the user.

A collage of people with different facial features

AI-generated content may be incorrect.

Figure: *Static image-based emotion classification using Hybrid ResNet50 on publicly available sample faces*

**4.2.2 User Interface and Visualization**

To interact with the user, the system provides bounding boxes around the detected faces in the video stream. Each box is colored based on the predicted emotion, e.g., yellow for "happy" or purple for "fear." Besides these boxes, text overlays indicate the predicted label and model confidence score, e.g., "Happy: 83%." These annotations provide immediate visual feedback and build user trust in the system's output. The interface also displays real-time performance metrics like frame rate and inference latency, which are essential in determining deployment readiness.

**4.2.3 Code Integration Challenges**

Adding tailored layers such as the ResNetLambda required the application of explicit serialization through TensorFlow's register\_keras\_serializable decorator. Otherwise, model loading would be unsuccessful due to unknown entities. Memory management was also high on the priority list. To load the model into consumer-grade GPUs' VRAM constraints, batch size during inference was reduced to the minimum of 1. Face detection and emotion classification were enabled concurrently with this optimization without resorting to oversubscription of accessible VRAM.

**5. Results and Analysis**

**5.1 Model Performance** Hybrid ResNet50 surpassed baseline CNNs on all the leading evaluation metrics. It achieved equalized accuracy of 58.87%, significantly superior to the baseline 45.20% benchmark for common CNNs. A macro F1 measure of 57.37% indicates the model maintaining good prediction stability across all emotion classes, not just the dominant ones. The log loss of 1.056 signifies good-calibrated confidence scores and high classification confidence.

Class-specific improvements were also seen. Retrieval for the "disgust" class was greatly improved—from 28% on the baseline CNN to 55%—due to strategic class weighting and hard augmentation. Blending up visually similar emotions, particularly "fear" and "surprise," was reduced from 38% to 26%, facilitated by the transformer module's global context modeling capability.

**5.2 Real-Time Demonstration Metrics** Real-time deployment operated at a frame rate of 24 FPS while running at a resolution of 1280x720. This was supported by an end-to-end latency of approximately 50 milliseconds per frame, enabling interactive and smooth user experiences. The model's inference engine required about 12 GB of VRAM, making it feasible to execute on consumer-grade GPUs such as the NVIDIA RTX 3060.

**A screenshot of a computer

AI-generated content may be incorrect.**

Figure: *Real-Time Emotion Recognition Demo Using Hybrid ResNet50 on Live Video Feed*

**5.3 Qualitative Outcomes** Qualitative assessments verified the stability of the model in practical environments. Seamless transitions between emotions and minimal label jitter were demonstrated by the video demo even when uneven lighting conditions were present. Image-based inference was highly accurate even when the faces were partially occluded, e.g., subjects wearing sunglasses, demonstrating the effectiveness of spatial attention modules to focus on prominent visible features.

**6. Future Directions**

**6.1 Model Improvements** In addition to this, future versions can look into replacing the ResNet50 backbone with a Vision Transformer (ViT-B/16) for better global feature representation. Further, the use of temporal modeling techniques such as 3D convolutional neural networks can allow for emotion recognition across sequences and hence detection of temporal dynamics in expressions.

**6.2 Deployment Optimization** For broader applicability, model quantization using TensorFlow Lite (e.g., INT8 conversion) can enable deployment on edge devices like smartphones and Raspberry Pi. Additionally, deploying the solution for web environments using TensorFlow.js would render browser-based inference backend-agnostic.

**8. Conclusion**

Transforming offline Hybrid ResNet50 into an online emotion detection system required overcoming multi-fold hurdles—ranging from computational to ethical. The final system achieves 58.87% balanced accuracy at 24 FPS, validating the efficacy of deep learning in real-world affective computing tasks. Democratization of access and resilience in the future through edge deployment and temporal and multimodal analysis.