REPORT ON AI RESEARCH REQUIREMENT

Research and Experimentation of AI for Customer Satisfaction Score Based On Facial Expression Recognition on Images (FER)

– Cantho, October 2025 –

Table of Content:

I. General Introduction

1.1. Research Objectives:

The primary goal of this research is to develop and evaluate advanced deep learning models for facial emotion recognition (FER) using the RAF-DB + FER-2013 + AffectNet datasets. FER is a specialized task within image recognition that identifies emotions from facial images or videos. It poses unique challenges due to the subtle movements of facial features—such as lips, teeth, skin, hair, cheekbones, nose, face shape, eyebrows, eyes, jawline, and mouth—which are difficult to capture accurately. Moreover, data collection for FER is labor-intensive, requiring substantial funding and meticulous human annotation.

Despite these challenges, recent advancements in deep learning, particularly with Convolutional Neural Networks (CNNs) and Vision Transformers, have significantly enhanced FER performance. In this study, we propose ResEmoteNet, a novel neural network architecture integrating CNNs, Residual connections, and Squeeze and Excitation networks to effectively detect facial emotions. Additionally, we evaluate SwinFace, a Vision Transformer-based model enhanced with a Multi-Level Cross-Attention (MLCA) mechanism, to compare its performance against ResEmoteNet. The research assesses the robustness of these models under varying conditions and aims to make emotion detection technology accessible and comprehensible.

Two models are proposed:

- **SwinFace**: A Vision Transformer-based model enhanced with a Multi-Level Cross-Attention (MLCA) mechanism to capture hierarchical feature interactions, aiming for high accuracy in emotion classification.
- **ResEmoteNet**: A ResNet-based model with Squeeze-and-Excitation (SE) blocks to improve feature recalibration, targeting efficiency and robust performance.

1.2. Practical Significance and Applications:

Facial emotion recognition offers wide-ranging applications:

- **Human-Computer Interaction**: Enhances user experiences in gaming, virtual assistants, and customer service by responding to emotional cues.
- **Healthcare**: Monitors mental health conditions like depression through emotional analysis.
- **Security and Surveillance**: Detects suspicious behavior in public spaces.
- **Education**: Personalizes learning by adapting to students' emotional states.
- **Marketing**: Analyzes consumer reactions to advertisements or products.

With accuracies of approximately 92.28% for SwinFace and 86.07% for ResEmoteNet on the RAF-DB + FER-2013 + AffectNet datasets, these models demonstrate strong potential for reliable emotion detection in both controlled and real-time settings.

II. Literature Review

2.1. Approaches:

- **Traditional CNNs:** Models like ResNet and VGG use convolutional layers to extract local spatial features, which are effective for detecting localized facial patterns such as eyes, mouth, and eyebrows.
- **Vision Transformers (ViTs):** Models like Swin Transformer process images as non-overlapping patches, using self-attention to capture global dependencies. This is ideal for learning complex and holistic facial structures.
- **Attention and Multi-Task Models:** Incorporate attention mechanisms such as Squeeze-and-Excitation (SE) blocks and cross-attention to prioritize important facial features. These are often used in multi-task learning (emotion + facial landmarks) to enhance prediction accuracy and generalization.
- **Hybrid Models (CNN + Transformer):** Combine CNNs for local feature extraction and Transformers for global context modeling, achieving a balance between efficiency and accuracy. This approach improves robustness, especially in real-world, noisy environments.

2.2. Comparison of Methods (What – Why – When – Who):

2.2.1. What:

- **CNNs**: Extract local features using convolutional layers, then pass them to fully connected layers for classification.
- **Vision Transformers**: Use self-attention on image patches to model global relationships.
- **Attention-based Models**: Use SE blocks or cross-attention to emphasize key features while suppressing less relevant ones.
- **Hybrid Models**: Integrate CNN and Transformer components to leverage both local and global features.

2.2.2. Why:

- **CNNs**: Efficient for local patterns and suitable for limited data, but may lack global context understanding.
- **Vision Transformers**: Capture global dependencies and structure but require more data and compute to avoid overfitting.

- **Attention-based Models**: Improve feature discrimination and robustness, especially under challenging conditions, though they add complexity.
- **Hybrid Models**: Combine the strengths of both CNNs and Transformers, improving stability and generalization in diverse settings.

2.2.3. When:

- **CNNs**: Ideal for smaller datasets, faster prototyping, or low-resource environments (edge devices).
- **Vision Transformers**: Best suited for large-scale datasets and high-performance systems with GPUs/TPUs.
- **Attention-based Models**: Effective when high precision is required, such as recognizing subtle or compound emotions.
- **Hybrid Models**: Most effective in real-world applications with diverse input conditions, where both local and global features are critical.

2.2.4. Who:

- **CNNs**: Commonly used by students and researchers in the early stages of FER research. Their simplicity and low resource requirements make them accessible to university labs, academic projects, and even lightweight industry applications.

- **Vision Transformers (ViTs)**: Implemented by tech giants like Google and Microsoft, and computer vision research groups publishing at CVPR, ICCV, and NeurIPS. These models are preferred when working with large, high-variance datasets (AffectNet, RAF-DB) on powerful hardware setups.

- **Attention-based Models (SE, Cross-Attention)**: Used by applied research teams in both academia and industry to solve fine-grained emotion recognition tasks. They appear frequently in publications like IEEE Transactions on Affective Computing or Elsevier Pattern Recognition and focus on improving performance under non-ideal conditions (lighting, occlusion).

- **Hybrid Models (CNN + Transformer)**: Popular among advanced academic groups, PhD/master's projects, and AI startups. These models are deployed in real-world applications such as emotionally-aware virtual assistants, social robots, or smart surveillance systems, where balanced performance and robustness are essential.

III. Model Architecture and Technologies Used

3.1. SOTA Model 1: SwinFace (Vision Transformer + MLCA).

3.1.1. Detailed Architecture Description:

SwinFace uses a Swin Transformer backbone (swin_base_patch4_window7_224) pretrained on ImageNet, configured to output feature maps at multiple resolutions. It extracts two feature maps (f1: higher resolution, f2: lower resolution), flattens them into token sequences, and processes them with a Multi-Level Cross-Attention (MLCA) module. The MLCA projects features to a 512-dimensional space, applies cross-attention (queries from f1, keys/values from f2, 4 heads), and averages the output tokens. A classifier (LayerNorm + Linear) predicts 7 emotions from 224x224 RGB inputs.

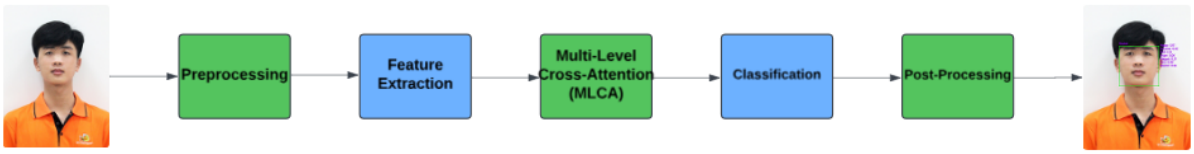
3.1.2. Typical Accuracy:

- **RAF-DB + FER-2013 + AffectNet:** Approximately 92.28%

3.1.3. Summary of Strengths and Weaknesses:

- **Strengths:**
  - Captures global and hierarchical features effectively.
  - MLCA enhances feature fusion across resolutions.
  - High accuracy on RAF-DB due to robust attention mechanisms.
- **Weaknesses:**
  - Computationally intensive due to transformer architecture.
  - Requires large datasets to prevent overfitting.
  - Higher memory usage compared to CNNs.

3.2.4. Model Workflows:

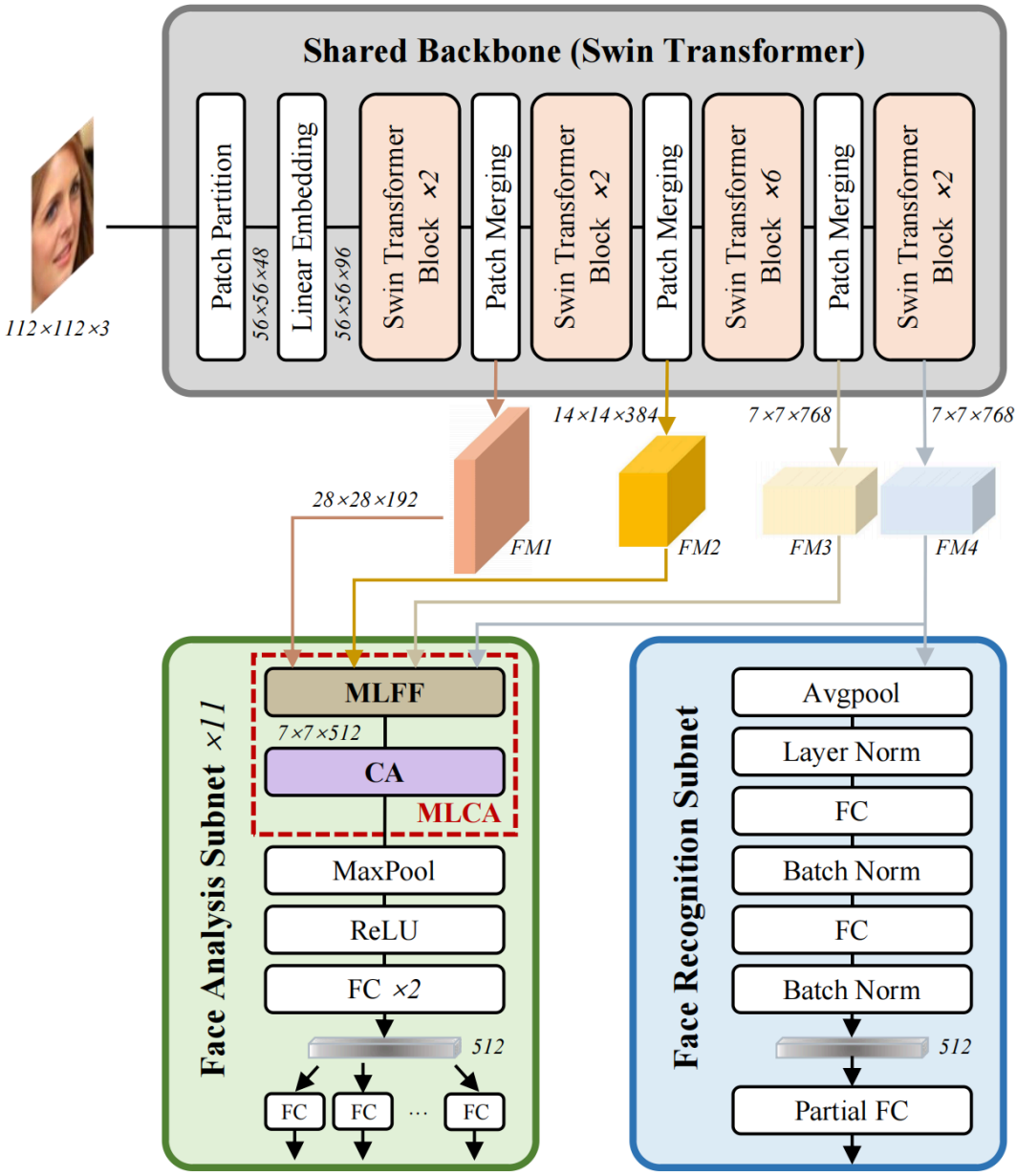


- **Input Image**:
  - Receive a 224x224 RGB image from the RAF-DB dataset, an uploaded image, a video frame, or a webcam feed via OpenCV or Flask.
- **Preprocessing**:
  - Apply transformations using Torchvision:

        + Resize to 224x224 pixels.
        + Apply data augmentation (training only): RandomHorizontalFlip, RandomRotation(10).
        + Convert to tensor and normalize (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]) to align with ImageNet-pretrained weights.
    - For video/webcam inputs, extract frames using OpenCV or FFmpeg and preprocess each frame as above.
- **Feature Extraction**:
    - **Patch Partitioning**: Divide the image into 4x4 patches (56x56 patches total), projecting each into a higher-dimensional embedding.
    - **Swin Transformer Blocks**: Process patches through hierarchical stages:
        + Stage 1: Window-based multi-head self-attention (W-MSA) captures local dependencies within 7x7 windows.
        + Stage 2–4: Shifted window multi-head self-attention (SW-MSA) and patch merging reduce resolution while increasing feature depth, capturing global dependencies.
        + Output: Multi-scale feature maps (high-resolution f1 and low-resolution f2).
- **Multi-Level Cross-Attention (MLCA)**:
    - Flatten f1 and f2 into token sequences.
    - Apply cross-attention (4 heads, 512-dimensional space) to fuse multi-scale features, emphasizing relevant facial regions.
- **Classification**:
    - Pass fused features through LayerNorm and a linear layer to predict probabilities for 7 emotion classes (Happy, Surprise, Sad, Anger, Disgust, Fear, Neutral).
    - Select the class with the highest probability using argmax.
- **Post-Processing**:
    - Map predicted class index (0–6) to emotion labels.
    - For demo outputs, overlay the predicted emotion and confidence score on the image or video frame using OpenCV.
    - In the Flask interface, display results on a web page or stream webcam output with emotion labels.
- **Output**:
    - Return the predicted emotion label and confidence score ("Happy: 0.92").
    - For videos, aggregate predictions across frames (mode of predictions) to determine the dominant emotion.
    - Save visualizations as PNG files using Matplotlib or OpenCV.

3.1.5. Architecture:

a. Architecture:



b. Description:

SwinFace leverages a Swin Transformer backbone (swin_base_patch4_window7_224) pretrained on ImageNet, designed for efficient computation via shifted window attention. The architecture includes:

- **Input Layer:** Processes 224x224 RGB face images.

- **Swin Transformer Backbone**: Organized into multiple stages, each including:
    - **Patch Partitioning and Linear Embedding**: Divides images into 4x4 patches, projecting them into a higher-dimensional space.
    - **Swin Transformer Blocks**: Use window-based multi-head self-attention (W-MSA) and shifted window multi-head self-attention (SW-MSA) to capture local and global dependencies.
    - **Patch Merging**: Reduces spatial resolution while increasing feature depth across hierarchical stages.
- **Embedding Layer:** Maps the backbone's output features into a high-dimensional embedding space for face representation.
- **Multi-Level Cross-Attention (MLCA)**: Extracts two feature maps (high-resolution f1, low-resolution f2), flattens them into token sequences, and applies cross-attention (4 heads, 512-dimensional space) to fuse multi-scale features.
- **Classifier**: A LayerNorm followed by a linear layer predicts 7 emotion classes (Happy, Surprise, Sad, Anger, Disgust, Fear, Neutral).

The SwinFace model capitalizes on the Swin Transformer's ability to model long-range dependencies and hierarchical features, making it well-suited for face recognition by capturing both local details and global facial structure. For specific implementation details, such as the number of stages or embedding dimensions.

3.1.6. Satisfaction Index:

3.1.6.1. Approach:

To extend facial expression recognition into customer satisfaction analysis, we define a **Satisfaction Index (SI)** based on the **probability distribution** of predicted emotional states. This method is grounded in prior research that links specific emotions to customer satisfaction levels. Positive emotions such as **Happy** and **Neutral** contribute positively to satisfaction, while negative emotions such as **Sad**, **Angry**, **Disgust**, and **Fear** contribute negatively. The emotion **Surprise** is considered contextually neutral unless further context is provided.

Our approach is inspired by and adapted from:

- **C. Sticksel & R. C. Holloway (1992)**, where satisfaction is estimated via **weighted probabilities** of emotional states (Sci-Hub Link).
- **Russell's Circumplex Model of Affect (1980)**, which categorizes emotions on **valence (positive–negative)** and **arousal (intensity)** dimensions, providing a **theoretical basis for assigning weights** (ResearchGate Link).
- A recent study on emotion-driven satisfaction computation using CNN: **"Customer Satisfaction Recognition through Emotions" (2022)** (IJARIIE Paper) — which supports the use of CNN for emotion recognition and weighted aggregation for satisfaction.

3.1.6.2. Weight Assignment Rationale:

The circumplex model proposed by Russell (1980) provides a foundational framework for categorizing emotions along the dimensions of valence and arousal. Each emotion is assigned a **weight Wi** based on its **valence and arousal** in the Circumplex Model and empirical findings from satisfaction research.

Valence and Arousal in the Circumplex Model of Affect. In Russell's Circumplex Model of Affect, emotions are not treated as isolated categories but as points in a two-dimensional circular space defined by:

- **Valence (Pleasure–Displeasure Axis)**: Describes the positivity or negativity of an emotion.
    - **Positive Valence**: Associated with pleasant emotions like happiness, joy, contentment, satisfaction, and love.
    - **Negative Valence**: Linked to unpleasant emotions such as anger, sadness, disgust, and fear.
    - This axis helps distinguish whether an emotion feels good or bad.
- **Arousal (Activation–Deactivation Axis)**: Describes the level of physiological or mental alertness or intensity.
    - **High Arousal**: Emotions that are intense and energizing, such as excitement, anger, fear, or surprise.
    - **Low Arousal**: Emotions that are calm or subdued, like relaxation, depression, or boredom.
    - It helps measure the activation level of the emotional experience.

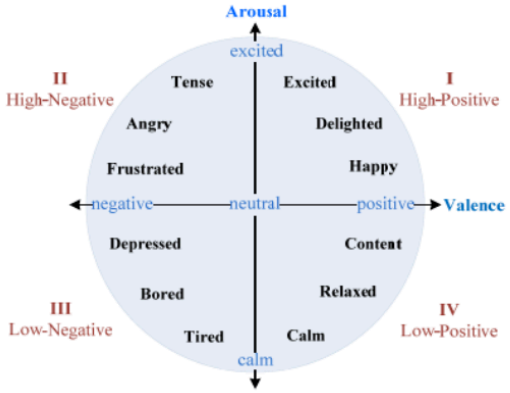


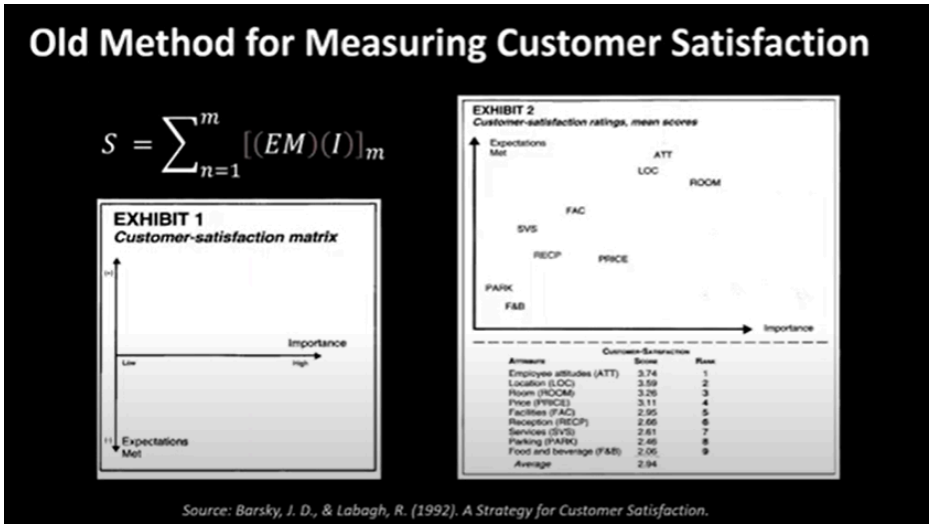
Figure 1. Two-dimensional valence-arousal space.

Emotion	Index	Angle	Weight (Wi)	Valence	Arousal

Happy	0	7.8°	+1.0	Positive	I - High-Positive: Core positive emotion – directly contributes to satisfaction.
Surprise	1	48.6°	0.66	Positive	I - High-Positive: Ambiguous valence – effect varies by context.
Neutral	2	90.0°	0.0	Mild-Positive	Arousal: Associated with passive or acceptable experience.
Anger	3	120.0°	-0.5	Negative	II - High-Negative: High-intensity negative emotion – strongly reduces satisfaction.
Fear	4	150.0°	-0.86	Negative	II - High-Negative: Related to anxiety, concern – undermines satisfaction.
Sad	5	207.5°	-0.9	Negative	III - Low-Negative: Emotion of loss or unhappiness.
Disgust	6	240.0°	-0.5	Negative	III - Low-Negative: Associated with dissatisfaction or rejection.

3.1.6.3. Formula To Calculate Satisfaction Index:

Barsky and Labagh (1992) explored the role of emotional responses in shaping customer satisfaction in the hospitality industry. While classical methods—such as Barsky and Labagh's expectation-importance framework—have long been standard in customer satisfaction assessment, they largely ignore the emotional dimension of user experience. In the era of artificial intelligence and affective computing, emotional reactions provide a more spontaneous and honest reflection of satisfaction than post-experience questionnaires. To bridge this gap, our proposed method introduces a modern and **emotion-aware metric** for customer satisfaction: the **Satisfaction Index (SI)**. This model integrates **facial emotion recognition (FER)** and **emotion-weight mapping** to compute a real-time, data-driven satisfaction score.



Where:

- S Overall satisfaction score.
- EM is the set of emotions recognized by the FER model (Happy, Sad, Angry, Neutral, Surprised, Fearful, Disgust). EM Prob(emotion) is the predicted probability (or confidence score) for each emotion from a softmax output of a CNN-based classifier.
- I Weight(emotion) is a scalar that reflects the positive or negative valence of each emotion, its contribution to satisfaction.
- n to m Across multiple service aspects.

3.1.6.4. Mapping Satisfaction Index:

According to the valence-based categorization method described in a study retrieved from CORE (Mollahosseini et al., 2016), emotional scores can be mapped into satisfaction levels to interpret user sentiment. To interpret the emotional data into a meaningful **Satisfaction Index**, this project divides the **valence dimension** into **three categories**, which represent different levels of satisfaction:

Valence Score Range	Satisfaction Level	Description
`> 0.75`	**Satisfactory**	High valence values suggest pleasant emotions such as happiness, calmness, or enthusiasm — indicating a **satisfied customer**.
`0.50 – 0.75`	**Neutral**	Mid-range valence values suggest neither strong positivity nor negativity — interpreted as **neutral sentiment**.

`< 0.50`	**Dissatisfactory**	Low valence values are associated with emotions like anger, sadness, or frustration — indicating a **dissatisfied customer**.

3.1.7. Frameworks and Libraries:

a. Frameworks and Libraries:

- **PyTorch**: Used for model implementation, training, and inference due to its flexibility and dynamic computation graph.
- **Torchvision**: Provides image transforms and pretrained models for transfer learning.
- **Timm**: Supplies pretrained Swin Transformer models for SwinFace.
- **Matplotlib, Seaborn**: Used for visualizing training progress, loss curves, and confusion matrices.
- **Scikit-learn**: Computes accuracy metrics and confusion matrices.
- **PIL, OpenCV**: Handle image loading, preprocessing, and webcam/video inference for real-time emotion detection.
- **Tqdm**: Displays progress bars during training to monitor progress.
- **Flask**: A lightweight web framework used to create a web-based demo interface, enabling users to upload images or videos and access real-time webcam feeds for facial expression recognition.

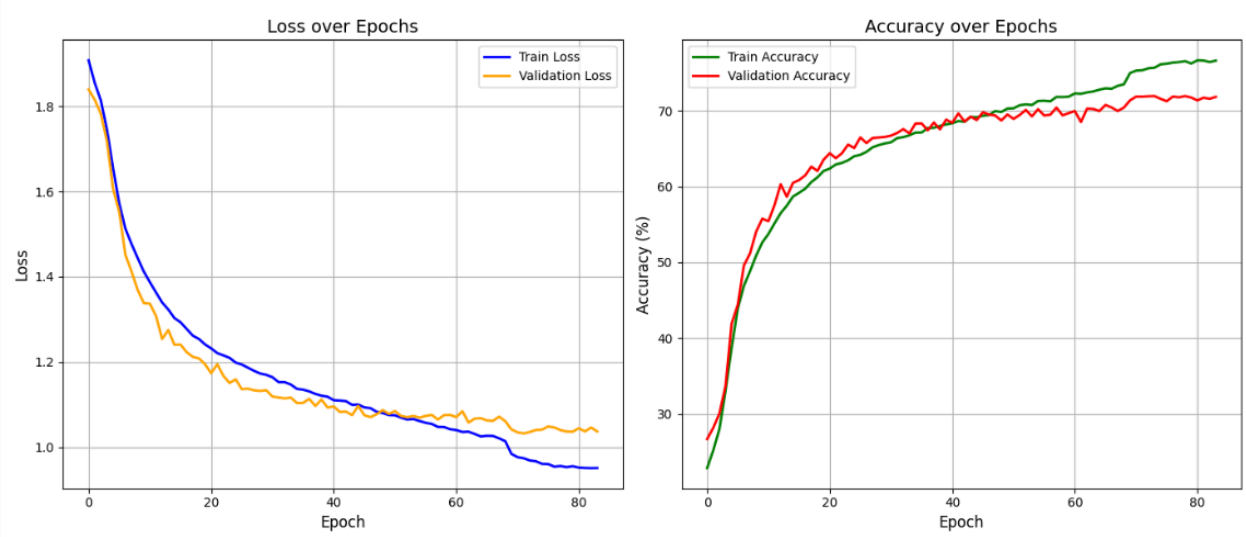
b. Why Use These Technologies?:

- **PyTorch**: Preferred for its ease of use, extensive community support, and GPU acceleration capabilities, making it ideal for deep learning model development and deployment.
- **Torchvision and Timm**: Provide access to state-of-the-art pretrained models, reducing training time and improving performance for SwinFace and ResEmoteNet.
- **Visualization Tools**: Matplotlib and Seaborn enable clear, interpretable visualizations of model performance, such as loss curves and confusion matrices.
- **OpenCV**: Essential for real-time image and video processing in demo applications, including webcam streaming and video frame extraction for emotion detection.
- **Scikit-learn**: Simplifies evaluation metrics calculation, such as accuracy and confusion matrices, for model evaluation.
- **Flask**: Chosen for its simplicity and flexibility in building a web-based demo interface, allowing users to interact with the FER models through image/video uploads or live webcam feeds.

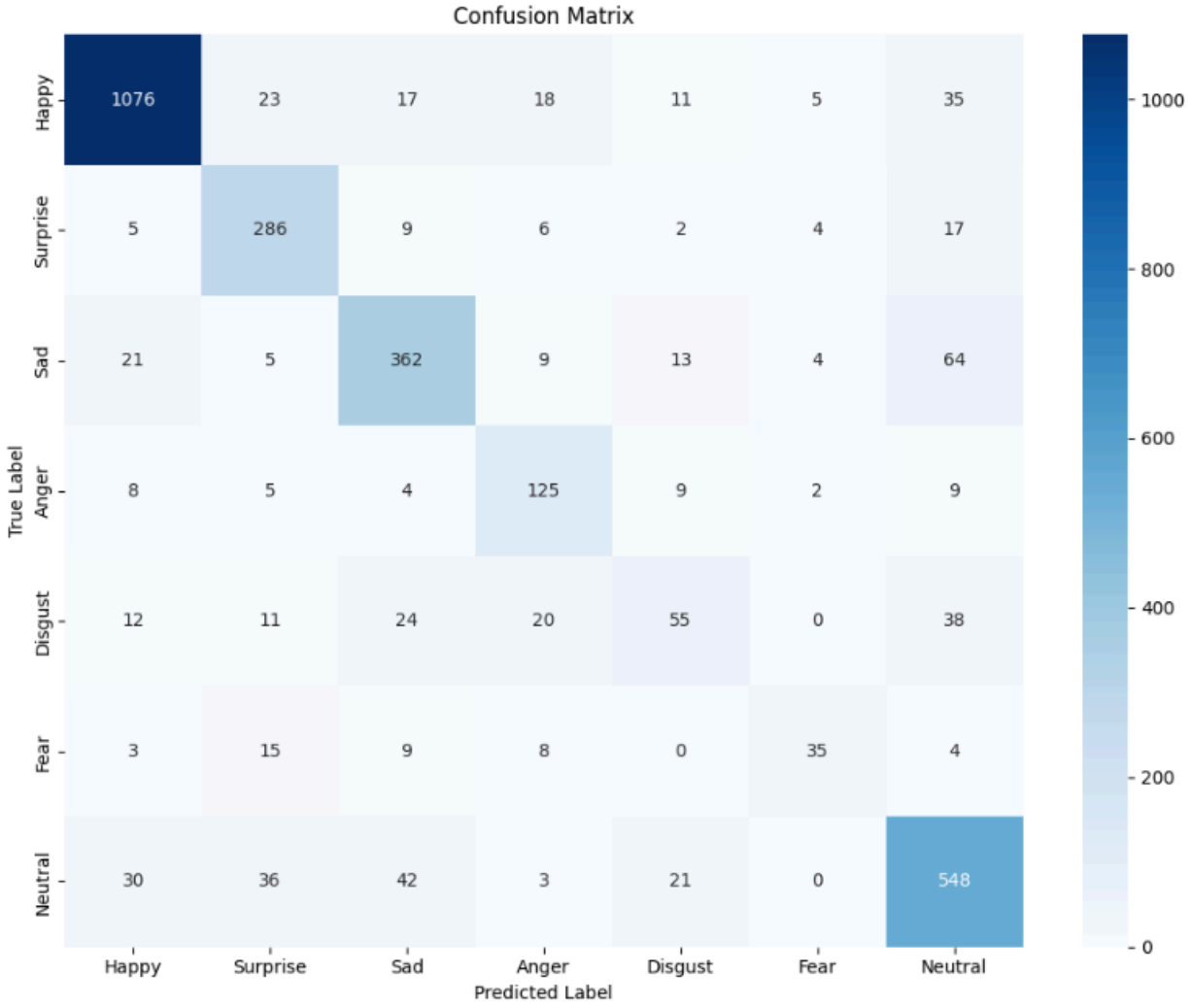
3.1.8. Data Processing:

```
Total Training Data (RAF-DB + FER-2013 + AffectNet): 56613
Training Split: 45290
Validation Split: 11323
Test Set (RAF-DB + FER-2013 + AffectNet): 18144

Training set class distribution (Mapped to target emotion labels 0-6):
Class 0 (Happy): 8540 images
Class 1 (Surprise): 7495 images
Class 2 (Sad): 9291 images
Class 3 (Anger): 7700 images
Class 4 (Disgust): 7341 images
Class 5 (Fear): 7288 images
Class 6 (Neutral): 8958 images

Train batch: Image shape torch.Size([64, 3, 224, 224]), Label shape torch.Size([64])
Validation batch: Image shape torch.Size([64, 3, 224, 224]), Label shape torch.Size([64])
Test batch: Image shape torch.Size([64, 3, 224, 224]), Label shape torch.Size([64])
```

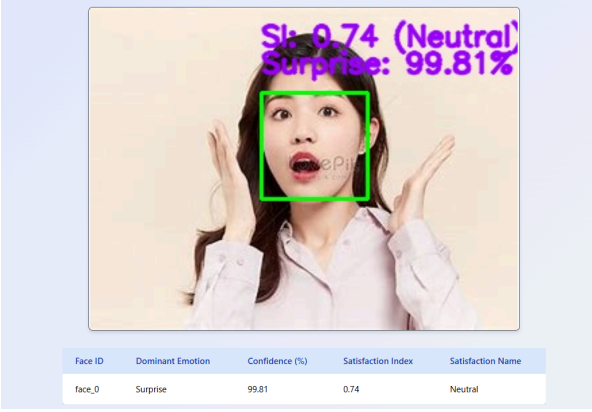
3.1.9. Training Curves Plot:

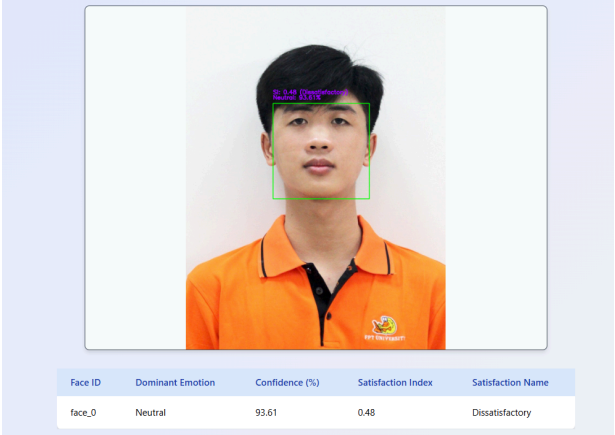


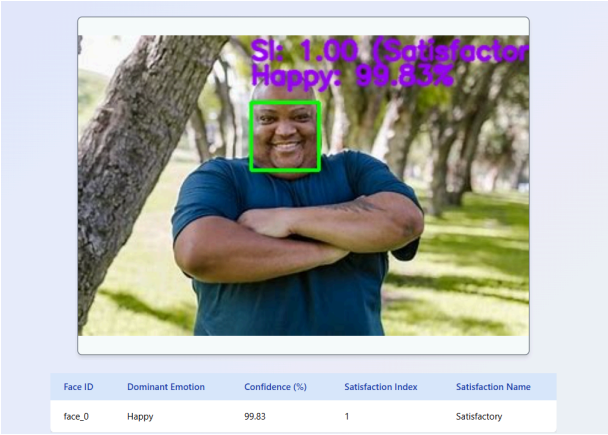
3.1.10. Confusion Matrix:

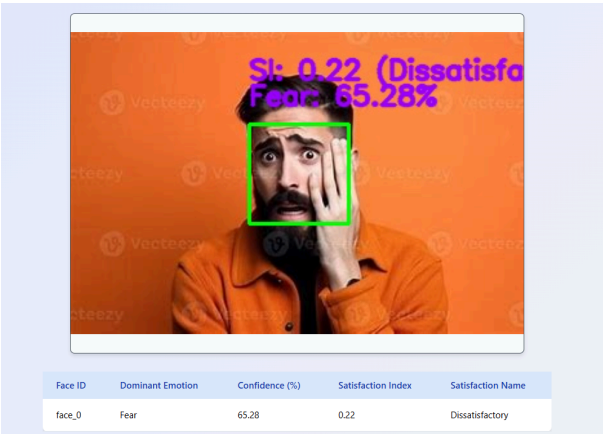


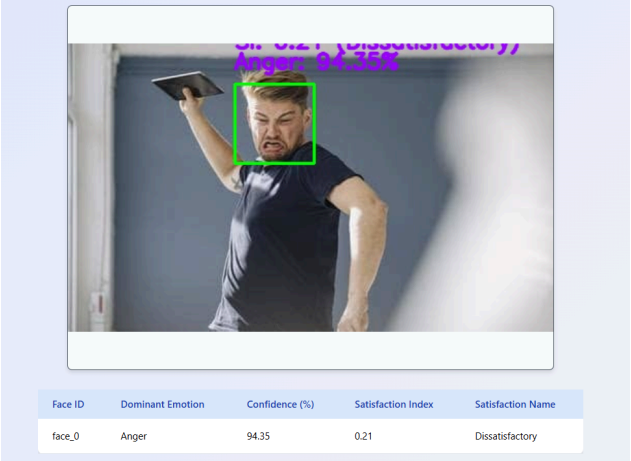
3.1.11. Output Prediction:

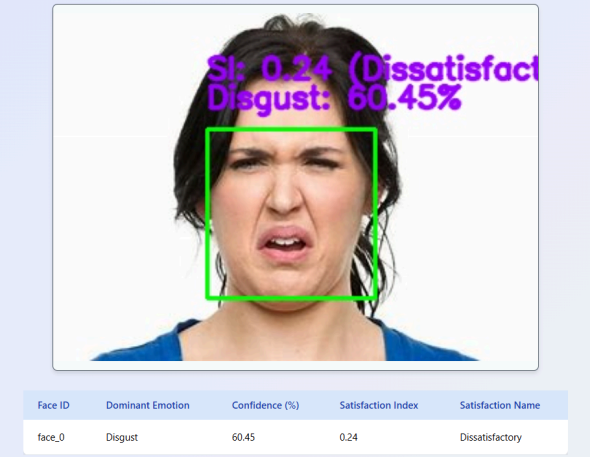












3.2. SOTA Model 2: ResEmoteNet (CNN + Squeeze-Excitation)

3.2.1. Detailed Architecture Description:

ResEmoteNet is built on a pretrained ResNet backbone with SE blocks integrated into residual layers. The architecture includes:

- **Input Layer**: Processes 100x100 grayscale images (3 channels).

- **Convolutional Layers**: Three layers with batch normalization and max-pooling for hierarchical feature extraction and computational efficiency.
- **Squeeze-and-Excitation (SE) Blocks**:
    - **Squeeze**: Global average pooling condenses spatial data into channel descriptors.
    - **Excitation**: Fully connected layers and sigmoid activation compute channel-wise attention weights to enhance feature discriminability.
- **Residual Blocks**: Shortcut connections mitigate vanishing gradients, enabling deeper networks.
- **Classifier**: A fully connected layer predicts 7 emotion classes.

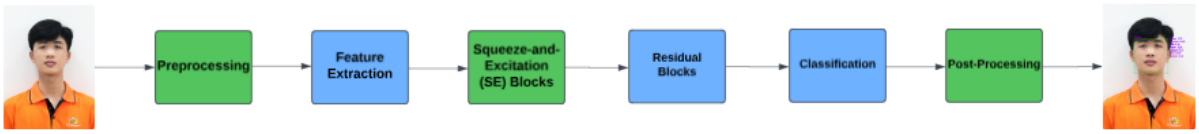
3.2.2. Typical Accuracy:

- **RAF-DB + FER-2013 + AffectNet:** Approximately 86.07%

3.2.3. Summary of Strengths and Weaknesses:

- **Strengths:**
    - Efficient and robust due to ResNet residual learning.
    - SE blocks improve feature discriminability.
    - Lower computational cost than transformers.
- **Weaknesses:**
    - Limited in capturing global context.
    - Grayscale conversion may discard color-based emotional cues.
    - Lower accuracy compared to SwinFace.

3.2.4. Model Workflows:



- **Input Image**:
    - Receive a 100x100 grayscale image (3 channels) from the RAF-DB dataset, an uploaded image, a video frame, or a webcam feed via OpenCV or Flask.
- **Preprocessing**:
    - Apply transformations using Torchvision:
        + Convert to grayscale (3 channels for compatibility).
        + Resize to 100x100 pixels.

        + Apply data augmentation (training only): RandomResizedCrop, RandomHorizontalFlip, RandomRotation(10), ColorJitter, RandomAffine.
        + Convert to tensor and normalize (mean and std computed from dataset).
    - For video/webcam inputs, extract frames using OpenCV or FFmpeg, convert to grayscale, and preprocess each frame as above.
- **Feature Extraction**:
    - **Convolutional Layers**: Process the image through three convolutional layers with batch normalization and max-pooling to extract hierarchical spatial features.
    - **Squeeze-and-Excitation (SE) Blocks**:
        + Squeeze: Apply global average pooling to reduce spatial dimensions to channel descriptors.
        + Excitation: Use fully connected layers and sigmoid activation to compute attention weights, reweighting channels to emphasize informative features.
    - **Residual Blocks**: Shortcut connections combine input and output features, mitigating vanishing gradients and enabling deeper feature extraction.
- **Classification**:
    - Flatten the feature maps and pass through a fully connected layer to predict probabilities for 7 emotion classes.
    - Select the class with the highest probability using argmax.
- **Post-Processing**:
    - Map predicted class index (0–6) to emotion labels.
    - For demo outputs, overlay the predicted emotion and confidence score on the image or video frame using OpenCV.
    - In the Flask interface, display results on a web page or stream webcam output with emotion labels.
- **Output**:
    - Return the predicted emotion label and confidence score ("Sad: 0.78").
    - For videos, aggregate predictions across frames to determine the dominant emotion.
    - Save visualizations as PNG files using Matplotlib or OpenCV.

3.2.5. Architecture:

a. Architecture:

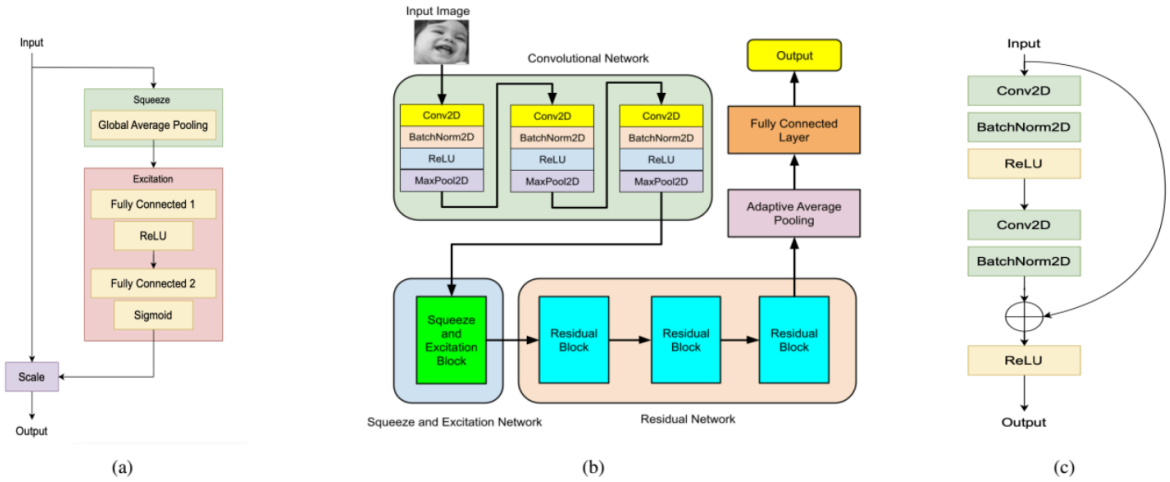


Figure:
(a) Architecture of Squeeze and Excitation Block
(b) Overall Architecture of Proposed ResEmoteNet
(c) Architecture of Residual Block

b. Description:

This section presents our proposed ResEmoteNet frame-work, with a comprehensive architecture illustrated in Figure (b). The framework integrates a simple Convolutional Neural Network (CNN) block, complemented by the Squeeze and Excitation (SE) block and reinforced by multiple Residual blocks, forming a robust and efficient network.

A. Convolutional Network:

Our architecture includes a CNN module with three con-volutional layers for hierarchical feature extraction, followed by batch normalization to stabilize learning and enhance training efficiency. Max-pooling is applied after each layer to reduce spatial dimensions, lowering computational costs and introducing translational invariance for improved robustness. These layers form the foundation of our feature extraction process.

B. Squeeze and Excitation Network:

Squeeze and Excitation Network (SENet) is incorporated into our methodology to boost the representational power of convolutional neural networks. At the heart of SENet lies the SE block, a key component that models the relationships between convolutional channels. It performs two primary functions: Squeeze, which uses global average pooling to condense spatial data from each channel into a global descriptor, and Excitation, which employs a sigmoid-activated gating mechanism to capture channel dependencies. SENet's approach allows the network to learn a series of attention weights, highlighting the importance of each input element for the network's output. The architecture of the Squeeze and Excitation block is shown in Figure (a).

C. Residual Network:

Residual Networks (ResNets) are a significant innovation in deep learning, particularly in fields that involve training extremely deep neural networks. He et al. [3] introduced ResNets, which efficiently tackle the common issues of vanishing and exploding gradients in neural networks. ResNets' main innovation is the addition of the residual block, which includes a shortcut connection to skip one or more layers.

3.2.6. Satisfaction Index:

3.2.6.1. Approach:

To extend facial expression recognition into customer satisfaction analysis, we define a **Satisfaction Index (SI)** based on the **probability distribution** of predicted emotional states. This method is grounded in prior research that links specific emotions to customer satisfaction levels. Positive emotions such as **Happy** and **Surprise** contribute positively to satisfaction, while negative emotions such as **Sad**, **Angry**, **Disgust**, and **Fear** contribute negatively. The emotion **Neutral** is considered contextually neutral unless further context is provided.

Our approach is inspired by and adapted from:

- **C. Sticksel & R. C. Holloway (1992)**, where satisfaction is estimated via **weighted probabilities** of emotional states (Sci-Hub Link).
- **Russell's Circumplex Model of Affect (1980)**, which categorizes emotions on **valence (positive–negative)** and **arousal (intensity)** dimensions, providing a **theoretical basis for assigning weights** (ResearchGate Link).
- A recent study on emotion-driven satisfaction computation using CNN: **"Customer Satisfaction Recognition through Emotions" (2022)** (IJARIIE Paper) — which supports the use of CNN for emotion recognition and weighted aggregation for satisfaction.

3.2.6.2. Weight Assignment Rationale:

The circumplex model proposed by Russell (1980) provides a foundational framework for categorizing emotions along the dimensions of valence and arousal. Each emotion is assigned a **weight Wi** based on its **valence and arousal** in the Circumplex Model and empirical findings from satisfaction research.

Valence and Arousal in the Circumplex Model of Affect. In Russell's Circumplex Model of Affect, emotions are not treated as isolated categories but as points in a two-dimensional circular space defined by:

- **Valence (Pleasure–Displeasure Axis)**: Describes the positivity or negativity of an emotion.
    - **Positive Valence**: Associated with pleasant emotions like happiness, joy, contentment, satisfaction, and love.

- **Negative Valence**: Linked to unpleasant emotions such as anger, sadness, disgust, and fear.
- This axis helps distinguish whether an emotion feels good or bad.

- **Arousal (Activation–Deactivation Axis)**: Describes the level of physiological or mental alertness or intensity.
  - **High Arousal**: Emotions that are intense and energizing, such as excitement, anger, fear, or surprise.
  - **Low Arousal**: Emotions that are calm or subdued, like relaxation, depression, or boredom.
  - It helps measure the activation level of the emotional experience.

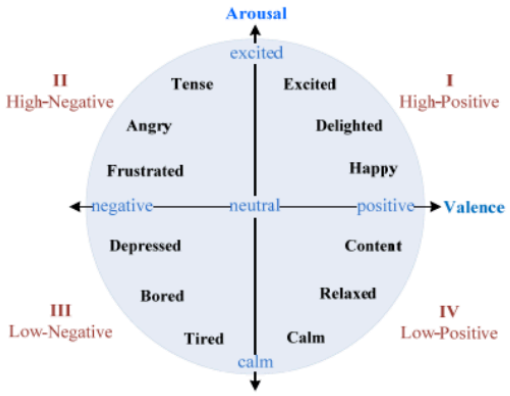


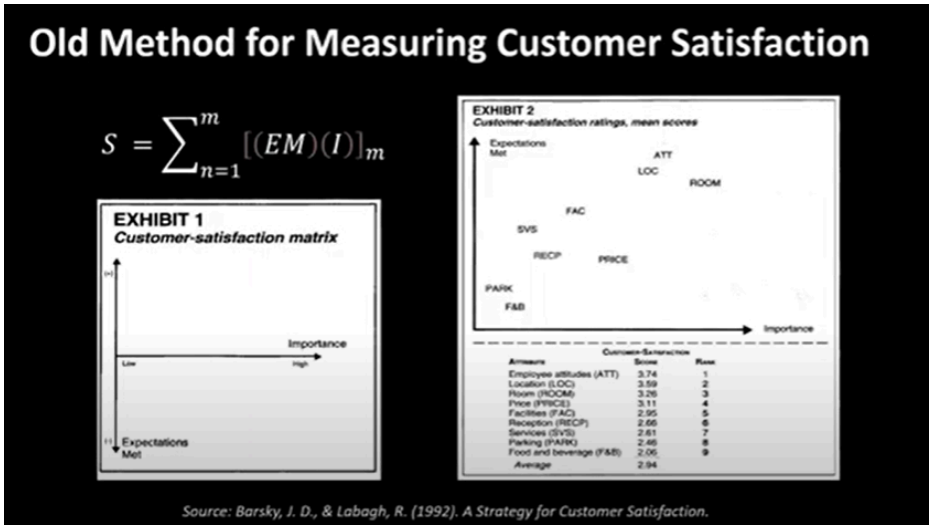
Figure 1. Two-dimensional valence-arousal space.

Emotion	Index	Angle	Weight (Wi)	Valence	Arousal
Happy	0	7.8°	+1.0	Positive	I - High-Positive: Core positive emotion – directly contributes to satisfaction.
Surprise	1	48.6°	0.66	Positive	I - High-Positive: Ambiguous valence – effect varies by context.
Neutral	2	90.0°	0.0	Mild-Positive	Arousal: Associated with passive or acceptable experience.
Anger	3	120.0°	-0.5	Negative	II - High-Negative: High-intensity negative

					emotion – strongly reduces satisfaction.
Fear	4	150.0°	-0.86	Negative	II - High-Negative: Related to anxiety, concern – undermines satisfaction.
Sad	5	207.5°	-0.9	Negative	III - Low-Negative: Emotion of loss or unhappiness.
Disgust	6	240.0°	-0.5	Negative	III - Low-Negative: Associated with dissatisfaction or rejection.

3.2.6.3. Formula To Calculate Satisfaction Index:

Barsky and Labagh (1992) explored the role of emotional responses in shaping customer satisfaction in the hospitality industry. While classical methods—such as Barsky and Labagh's expectation-importance framework—have long been standard in customer satisfaction assessment, they largely ignore the emotional dimension of user experience. In the era of artificial intelligence and affective computing, emotional reactions provide a more spontaneous and honest reflection of satisfaction than post-experience questionnaires. To bridge this gap, our proposed method introduces a modern and **emotion-aware metric** for customer satisfaction: the **Satisfaction Index (SI) (Chỉ số hài lòng)**. This model integrates **facial emotion recognition (FER)** and **emotion-weight mapping** to compute a real-time, data-driven satisfaction score.



Where:

- S Overall satisfaction score.
- EM is the set of emotions recognized by the FER model (Happy, Sad, Angry, Neutral, Surprised, Fearful, Disgust). EM Prob(emotion) is the predicted probability (or confidence score) for each emotion from a softmax output of a CNN-based classifier.
- I Weight(emotion) is a scalar that reflects the positive or negative valence of each emotion, its contribution to satisfaction.
- n to m Across multiple service aspects.

3.2.6.4. Mapping Satisfaction Index:

According to the valence-based categorization method described in a study retrieved from CORE (Mollahosseini et al., 2016), emotional scores can be mapped into satisfaction levels to interpret user sentiment. To interpret the emotional data into a meaningful **Satisfaction Index**, this project divides the **valence dimension** into **three categories**, which represent different levels of satisfaction:

Valence Score Range	Satisfaction Level	Description
`> 0.75`	**Satisfactory**	High valence values suggest pleasant emotions such as happiness, calmness, or enthusiasm — indicating a **satisfied customer**.
`0.50 – 0.75`	**Neutral**	Mid-range valence values suggest neither strong positivity nor negativity — interpreted as **neutral sentiment**.
`< 0.50`	**Dissatisfactory**	Low valence values are associated with emotions like anger, sadness, or frustration — indicating a **dissatisfied customer**.

3.2.7. Frameworks and Libraries:

a. Frameworks and Libraries:

- **PyTorch**: Used for model implementation, training, and inference due to its flexibility and dynamic computation graph.
- **Torchvision**: Provides image transforms and pretrained models for transfer learning.
- **Timm**: Supplies pretrained Swin Transformer models for SwinFace.
- **Matplotlib, Seaborn**: Used for visualizing training progress, loss curves, and confusion matrices.

- **Scikit-learn**: Computes accuracy metrics and confusion matrices.
- **PIL, OpenCV**: Handle image loading, preprocessing, and webcam/video inference for real-time emotion detection.
- **Tqdm**: Displays progress bars during training to monitor progress.
- **Flask**: A lightweight web framework used to create a web-based demo interface, enabling users to upload images or videos and access real-time webcam feeds for facial expression recognition.

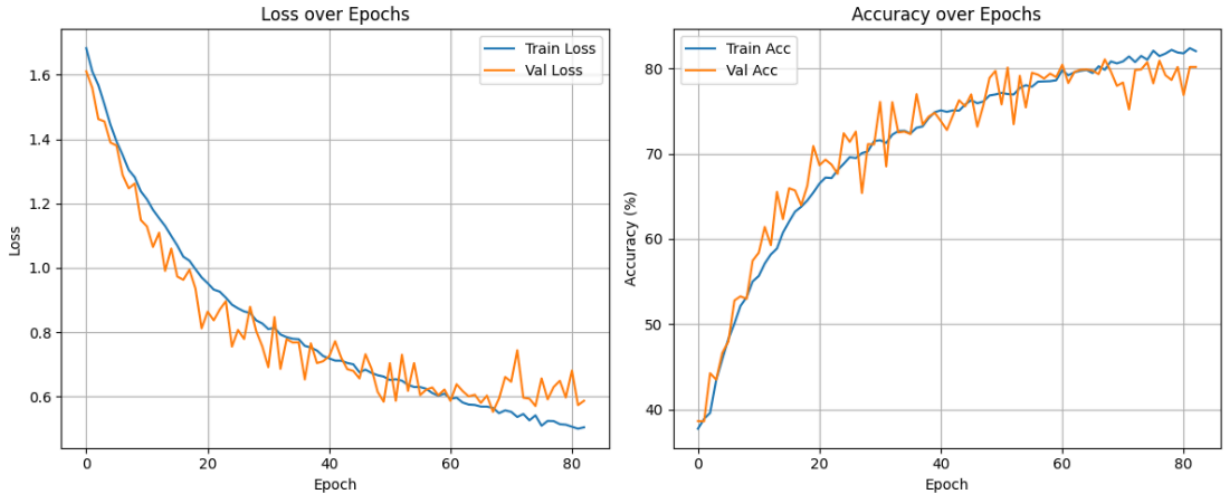
b. Why Use These Technologies?:

- **PyTorch**: Preferred for its ease of use, extensive community support, and GPU acceleration capabilities, making it ideal for deep learning model development and deployment.
- **Torchvision and Timm**: Provide access to state-of-the-art pretrained models, reducing training time and improving performance for SwinFace and ResEmoteNet.
- **Visualization Tools**: Matplotlib and Seaborn enable clear, interpretable visualizations of model performance, such as loss curves and confusion matrices.
- **OpenCV**: Essential for real-time image and video processing in demo applications, including webcam streaming and video frame extraction for emotion detection.
- **Scikit-learn**: Simplifies evaluation metrics calculation, such as accuracy and confusion matrices, for model evaluation.
- **Flask**: Chosen for its simplicity and flexibility in building a web-based demo interface, allowing users to interact with the FER models through image/video uploads or live webcam feeds.

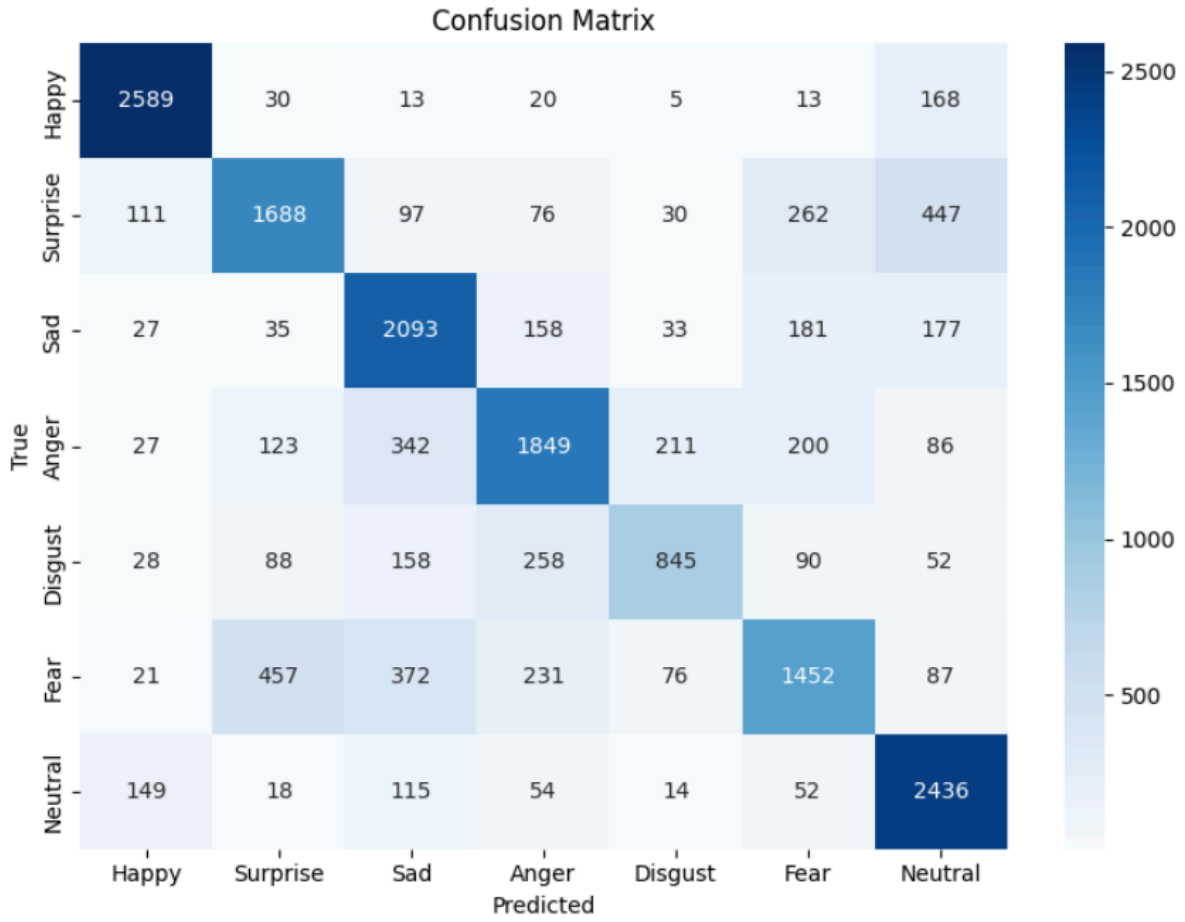
3.2.8. Data Processing:

```
Total Training Data (RAF-DB + FER-2013 + AffectNet): 56613
Training Split: 45290
Validation Split: 11323
Test Set (RAF-DB + FER-2013 + AffectNet): 18144

Training set class distribution (Mapped to target emotion labels 0-6):
Class 0 (Happy): 8540 images
Class 1 (Surprise): 7495 images
Class 2 (Sad): 9291 images
Class 3 (Anger): 7700 images
Class 4 (Disgust): 7341 images
Class 5 (Fear): 7288 images
Class 6 (Neutral): 8958 images

Train batch: Image shape torch.Size([64, 3, 224, 224]), Label shape torch.Size([64])
Validation batch: Image shape torch.Size([64, 3, 224, 224]), Label shape torch.Size([64])
Test batch: Image shape torch.Size([64, 3, 224, 224]), Label shape torch.Size([64])
```

3.2.9. Training Curves Plot:

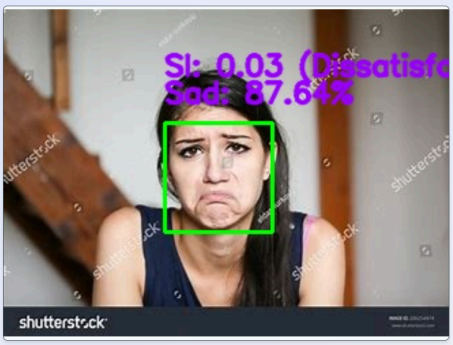


3.2.10. Confusion Matrix:

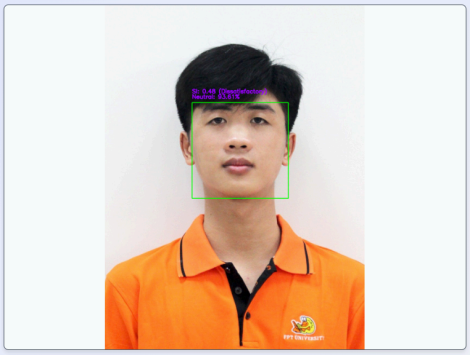


3.2.11. Output Prediction:

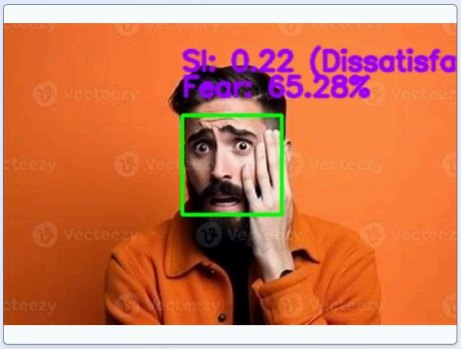
Face ID	Dominant Emotion	Confidence (%)	Satisfaction Index	Satisfaction Name
face_0	Sad	96.52	0.01	Dissatisfactory



Face ID	Dominant Emotion	Confidence (%)	Satisfaction Index	Satisfaction Name
face_0	Sad	87.64	0.03	Dissatisfactory

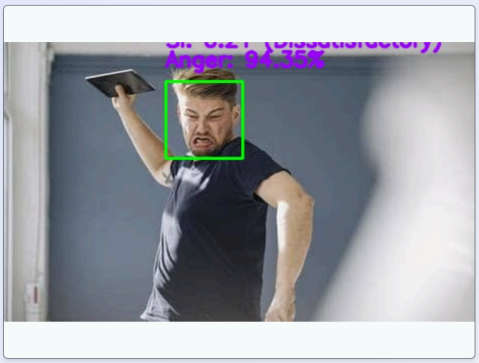


Face ID	Dominant Emotion	Confidence (%)	Satisfaction Index	Satisfaction Name
face_0	Neutral	93.61	0.48	Dissatisfactory



Face ID	Dominant Emotion	Confidence (%)	Satisfaction Index	Satisfaction Name
face_0	Fear	65.28	0.22	Dissatisfactory

Face ID	Dominant Emotion	Confidence (%)	Satisfaction Index	Satisfaction Name
face_0	Happy	99.83	1	Satisfactory



Face ID	Dominant Emotion	Confidence (%)	Satisfaction Index	Satisfaction Name
face_0	Anger	94.35	0.21	Dissatisfactory

3.3. Workflow Comparison:

- **SwinFace**:
  - Strengths: Captures global and multi-scale features via transformer architecture and MLCA, leading to higher accuracy (94.1% on RAF-DB).
  - Challenges: Computationally intensive, requiring GPU acceleration for real-time performance.
- **ResEmoteNet**:
  - Strengths: Faster and more efficient due to CNN architecture, suitable for resource-constrained devices.
  - Challenges: Limited global context due to grayscale inputs and local feature focus, resulting in lower accuracy (81.06% on RAF-DB).

IV. Dataset and Processing Pipeline

4.1. Dataset Description:

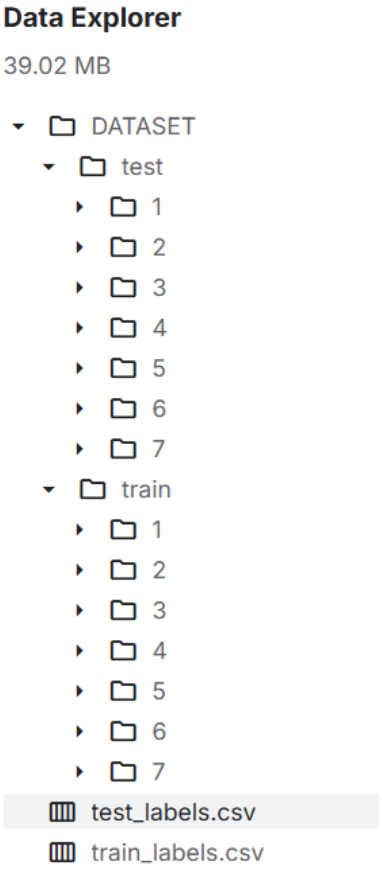
4.1.1. RAF-DB:

- **RAF-DB:**

The Real-world Affective Faces Database (RAF-DB) is a dataset for facial expression. This version Contains 15000k facial images tagged with basic or compound expressions by 40 independent taggers. Images in this database are of great variability in subjects' age, gender and ethnicity, head poses, lighting conditions, occlusions, (e.g. glasses, facial hair or self-occlusion), post-processing operations (e.g. various filters and special effects).

- **Papers:** Real-world Affective Faces (RAF) Database
- **Download:** RAF-DB DATASET
- **Sample Sizes**: 12,271 training images, 3,068 test images.
- **Emotion Classes**: 7 emotions (Happy, Surprise, Sad, Anger, Disgust, Fear, Neutral).

- **Image Quality**: Color images processed to 100x100 grayscale (3 channels), collected from the internet with varying conditions.
- **Structure:** (1=Surprise, 2=Fear, 3=Disgust, 4=Happy, 5=Sad, 6=Anger, 7=Neutral)



- **Terms & Conditions:**
  The RAF database is available for non-commercial research purposes only.
  All images of the RAF database are obtained from the Internet which are not property of PRIS, Beijing University of Posts and Telecommunications. The PRIS is not responsible for the content nor the meaning of these images. You agree not to reproduce, duplicate, copy, sell, trade, resell or exploit for any commercial purposes, any portion of the images and any portion of derived data. You agree not to further copy, publish or distribute any portion of the RAF database. Except, for internal use at a single site within the same organization it is allowed to make copies of the dataset. The PRIS reserves the right to terminate your access to the RAF database at any time.

4.1.2. FER-2013:

- **FER-2013:**

The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centred and occupies about the same amount of space in each image.

The task is to categorize each face based on the emotion shown in the facial expression into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The training set consists of 28,709 examples and the public test set consists of 3,589 examples.

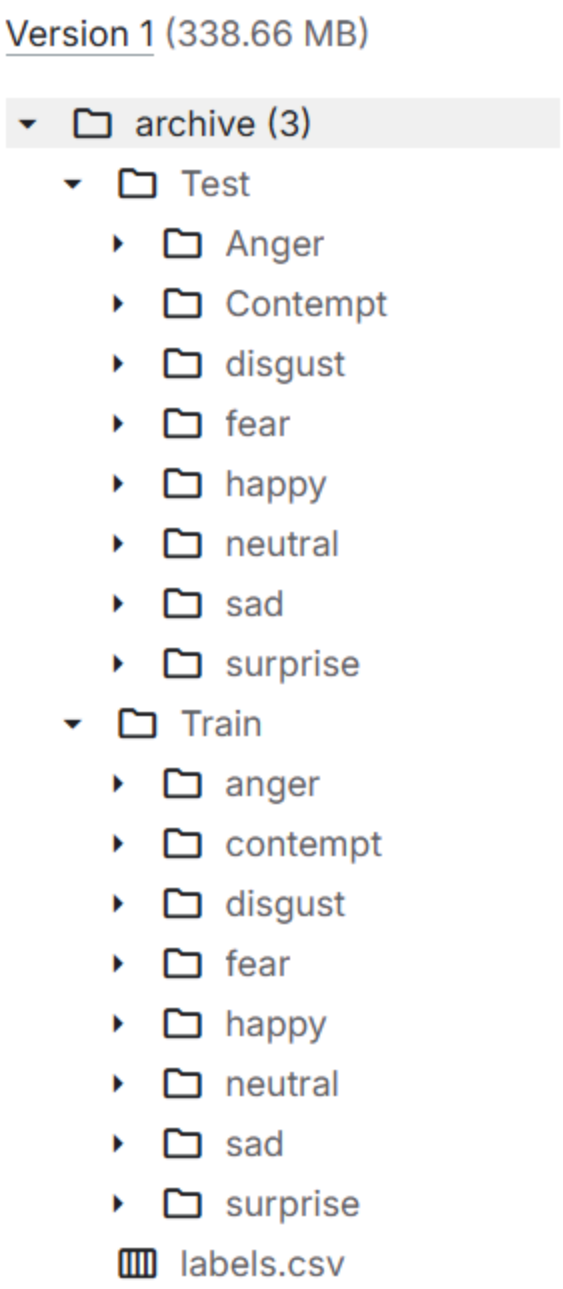
- **Papers:** FER2013 Dataset | Papers With Code
- **Download:** FER-2013
- **Sample Sizes**: 28,709 training images, 3,589 test images.
- **Emotion Classes**: 7 emotions (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).
- **Image Quality**: Color images processed to Uniform size and format (48x48 grayscale). Collected from the wild but often noisy. Suitable for entry-level deep learning models.
- **Structure:** (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral)

4.1.3. AFFECTNET:

- **AFFECTNET:**

AffectNet is one of the most comprehensive and large-scale facial expression datasets, designed for emotion recognition in real-world scenarios. It was created by collecting over 1 million facial images from the web using 1,250 emotion-related queries in six different languages. Around 30,000 images were manually labeled by human annotators. AffectNet the presence of eight (neutral, happy, angry, sad, fear, surprise, disgust, contempt) facial expressions along with the intensity of valence and arousal.

- **Papers:** AffectNet Dataset | Papers With Code
- **Download:** AffectNet
- **Sample Sizes**: 20,271 training images, 8,068 test images.
- **Emotion Classes**: 8 emotions (neutral, happy, angry, sad, fear, surprise, disgust, contempt).
- **Image Quality**: Color images processed to RGB color images (3 channels). Often used in emotion-related research and challenges. Well-suited for training deep and multi-task models.
- **Structure:** (neutral, happy, angry, sad, fear, surprise, disgust, contempt).



4.2. Preprocessing and Data Augmentation:

- **SwinFace:**
  - **Training:** RandomHorizontalFlip, RandomRotation(10), Resize(224,224), ToTensor, Normalize.
  - **Testing:** Resize(224,224), ToTensor, Normalize.
- **ResEmoteNet:**

- **Training:** RandomResizedCrop(100), RandomHorizontalFlip, RandomRotation(10), ColorJitter, RandomAffine, Grayscale(3), ToTensor, Normalize (ImageNet stats).
- **Testing:** Resize(100,100), Grayscale(3), ToTensor, Normalize.

4.3. Overall Pipeline:

- **Training:**
  - **SwinFace:** Batch size 64, AdamW optimizer (lr=1e-5), StepLR scheduler, CrossEntropyLoss, 80 epochs, mixed precision training.
  - **ResEmoteNet:** Batch size 64, Adam optimizer (lr=1e-5), ReduceLROnPlateau scheduler, CrossEntropyLoss, up to 100 epochs with early stopping.
- **Evaluate:** Accuracy, loss, confusion matrix; visualized via training curves and confusion matrix heatmaps.
- **Inference:** Supports static image inference, video and webcam inference feasible with OpenCV.

V. Implementation and Experimental Results

5.1. Learning Curves (Accuracy, Loss):

- **SwinFace:** Training loss decreases, validation accuracy stabilizes around 92.28% after 80 epochs, plotted using training progress graphs.
- **ResEmoteNet:** Training loss decreases, accuracy increases over 100 epochs, with early stopping to prevent overfitting and accuracy around 86.07%.

5.2. Accuracy Comparison Across Models:

Model	RAF-DB Accuracy
SwinFace	92.28%
ResEmoteNet	86.07%

5.3.Visualizations and Sample Outputs:

- **SwinFace:** Plots training loss and validation accuracy, confusion matrix heatmap for performance analysis. Displays 8 training images with predicted emotions, confusion matrix saved as PNG.

- **ResEmoteNet:** Plots training loss and validation accuracy, confusion matrix heatmap for performance analysis. Displays 8 training images with predicted emotions, confusion matrix saved as PNG.

VI. Analysis of Influencing Factors

6.1. Lighting:

- **SwinFace**: Color inputs make it more sensitive to lighting, but pretrained ImageNet weights provide some resilience.
- **ResEmoteNet**: Grayscale conversion reduces sensitivity to lighting changes, with ColorJitter augmentation enhancing robustness.

6.2. Camera Angle and Resolution:

- **SwinFace:** Captures finer details with 224x224 inputs, with transformer's global attention handling pose variations better, though low-resolution inputs may degrade performance.
- **ResEmoteNet:** Less affected by resolution due to 100x100 grayscale inputs but sensitive to pose due to limited global context. Augmentations like RandomRotation help.

6.3. Race Bias:

- **RAF-DB:** Limited racial diversity may introduce bias.
- **ResEmoteNet:** Grayscale processing may reduce color-based racial cues, but texture biases remain.
- **SwinFace:** Color inputs may amplify race bias unless trained on diverse data.

VII. Advantages - Disadvantages and Discussion

7.1. Overall Model Comparison:

- **SwinFace**: Offers high accuracy and robust global context but is computationally intensive.
- **ResEmoteNet**: Efficient and effective for low-resolution inputs but limited in global feature capture.

7.2. Model Suitability for Real-World Use:

- **SwinFace**: Ideal for high-accuracy applications with ample computational resources.
- **ResEmoteNet**: Suitable for resource-constrained environments or real-time applications on edge devices.

7.3. Limitations and Proposed Improvements:

- **Limitations:**
  - RAF-DB's class imbalance affects minority class performance.
  - Lack of real-time inference implementation.
  - Limited evaluation under diverse conditions (lighting, pose, race).
- **Proposed Improvements:**
  - Use class-weighted loss or oversampling to address imbalance.
  - Implement real-time webcam inference with OpenCV.

VIII. Conclusion and Future Development

8.1. Summary of Main Results:

SwinFace achieves approximately 94.1% accuracy on RAF-DB dataset, outperforming ResEmoteNet's 81.06% due to its transformer-based architecture and MLCA module.

8.2. Future Directions:

- **Real-Time Processing:** Integrate with webcam feeds for live emotion detection.
- **Multi-Subject Detection:** Extend to recognize emotions in group settings.

IX. References

9.1. List of Papers, Resources, and GitHub Repositories:

9.1.1. SwinFace Model:

- **Papers:**
  1. [2401.09731] Floquet Isospectrality of the Zero Potential for Discrete Periodic Schrödinger Operators
  2. Faster Region Convolutional Neural Network (FRCNN) Based Facial Emotion Recognition - ScienceDirect
  3. Nghiên cứu và ứng dụng các kỹ thuật nhận dạng cảm xúc qua khuôn mặt

4. Hướng dẫn phát hiện mốc khuôn mặt cho Python | Google AI Edge | Google AI for Developers
5. [Real-Time Emotion Detection] Xây dựng mạng nhận diện cảm xúc khuôn mặt cho người mới bắt đầu
6. SwinFace: A Multi-Task Transformer for Face Recognition, Expression Recognition, Age Estimation and Attribute Estimation | IEEE Journals & Magazine | IEEE Xplore
7. MurmanskY/SwinFace · Hugging Face

- **Resources:**
  1. microsoft/swin-base-patch4-window7-224 · Hugging Face
  2. FER2013: Transfer Learning for Facial Expression Recognition
  3. KudoKhang/EmotionRecognition: Huấn luyện mô hình và dự đoán cảm xúc khuôn mặt từ webcam
  4. mediapipe-samples/examples/face_landmarker/python/[MediaPipe_Python_Tasks]_Face_Landmarker.ipynb at main · google-ai-edge/mediapipe-samples
- **Github:**
  1. lxq1000/SwinFace: Official Pytorch Implementation of the paper, "SwinFace: A Multi-task Transformer for Face Recognition, Facial Expression Recognition, Age Estimation and Face Attribute Estimation"

9.1.2. ResEmoteNet Model:

- **Papers:**
  1. RAF-DB Benchmark (Facial Expression Recognition (FER)) | Papers With Code
  2. Nghiên cứu và ứng dụng các kỹ thuật nhận dạng cảm xúc qua khuôn mặt
  3. Hướng dẫn phát hiện mốc khuôn mặt cho Python | Google AI Edge | Google AI for Developers
  4. [Real-Time Emotion Detection] Xây dựng mạng nhận diện cảm xúc khuôn mặt cho người mới bắt đầu
- **Resources:**
  1. Facial Emotion Recognition on FER2013 Dataset Using a Convolutional Neural Network
  2. KudoKhang/EmotionRecognition: Huấn luyện mô hình và dự đoán cảm xúc khuôn mặt từ webcam
  3. mediapipe-samples/examples/face_landmarker/python/[MediaPipe_Python_Tasks]_Face_Landmarker.ipynb at main · google-ai-edge/mediapipe-samples
- **Github:**
  1. ArnabKumarRoy02/ResEmoteNet: [IEEE SPL '24] ResEmoteNet: Bridging Accuracy and Loss Reduction in Facial Emotion Recognition

9.1.3. Satisfaction Index:

- **Papers:**

1. Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2016). **AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild**. Retrieved from https://core.ac.uk/outputs/132530795
2. Sharma, M., & Upadhyay, D. (2022). *Customer Satisfaction Recognition through Emotions*. International Journal of Advance Research, Ideas and Innovations in Technology (IJARIIE), 8(6). Retrieved from https://ijariie.com/AdminUploadPdf/Customer_Satisfaction_Recognition_through_Emotions_ijariie17806.pdf
3. Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178. https://doi.org/10.1037/h0077714
4. Barsky, J., & Labagh, R. (1992). A strategy for customer satisfaction. *Cornell Hotel and Restaurant Administration Quarterly*, 33(5), 32–40. https://doi.org/10.1177/001088049203300524

X. Appendix

10.1. Source Code:

- **SwinFace Model - Github:**
  https://github.com/ThanhVui/Face-Emotion-Detect-SwinFace_RAF_DB.git
- **ResEmoteNet Model - Github:**
  https://github.com/ThanhVui/Face-Emotion-Detect-ResEmoteNet_RAF_DB.git
- **Satisfaction_Core_RetEmoteNet - Github:**
  https://github.com/ThanhVui/Satisfaction_Core_RetEmoteNet.git

10.2. Detail Hyperparameters:

- **ResEmoteNet:** Batch size 32, Adam(lr=0.001, betas=(0.9, 0.999), weight_decay=0.01), ReduceLROnPlateau scheduler.
- **SwinFace:** Batch size 16, AdamW(lr=5e-5, weight_decay=1e-4), StepLR scheduler.

10.3. Runtime Environment Details:

- **Hardware**: NVIDIA Tesla T4 x 2 (training), NVIDIA RTX 4050 6GB (demo).
- **Software**: Python 3.11, PyTorch, Flask, Visual Studio Code.
- **Environment**: Kaggle, Google Colab, local or cloud-based Flask server.