Transfer Learning for Facial Emotion Recognition using Vision Transformers

Won Kim
Luddy School of Informatics, Computing, and Engineering
Indiana University Bloomington
wk17@iu.edu

Tripp Ix
Luddy School of Informatics, Computing, and Engineering
Indiana University Bloomington
tredix@iu.edu

Brian Lee
Luddy School of Informatics, Computing, and Engineering
Indiana University Bloomington
bl35@iu.edu

Shun Funeno
Luddy School of Informatics, Computing, and Engineering
Indiana University Bloomington
sfuneno@iu.edu

1 Introduction

Human emotion is a fundamental pillar of communication, social interaction, and decision-making. The ability to interpret subtle emotional cues allows individuals to understand intentions, recognize distress, and respond appropriately within various social environments. As artificial intelligence continues to integrate into daily life via virtual assistants, therapeutic tools, and educational software, it becomes essential for machines to approximate this human capability. Facial expressions, being one of the most direct and universally understood channels of emotional communication, offer a practical pathway toward building systems that can automatically interpret affective states.

The primary goal of this project is to engineer a machine learning system capable of recognizing human emotions from facial images in real time. While our initial exploration began with traditional Convolutional Neural Networks (CNNs), this finalized research evolves toward Vision Transformer architectures. By leveraging the Real-world Affective Faces Database (RAF-DB), a collection of 15,000 real-world images, we examine how modern transformer-based models can overcome the challenges of pose variation, lighting inconsistency, and occlusion that typically hinder traditional systems. This paper details the full engineering process, from the definition of the problem space to a comparative empirical analysis of cutting-edge architectures including ViT-Base, DeiT-Small, and CvT-21.

2 Problem Space

The computational performance and reliability of an emotion recognition system are determined by two main components: face detection and emotion classification. In a real-time pipeline, the system must process video frames at or near 30 frames per second, meaning each frame must be analyzed within roughly 33 milliseconds.

Facial Emotion Recognition (FER) is inherently challenging due to several factors. Expressions vary dramatically across individuals based on age, anatomy, and cultural background. Environmental conditions such as inconsistent lighting, low camera resolution, and off-angle poses (exceeding 45 degrees) can distort visual cues. Furthermore, emotion categories are often ambiguous; human annotators frequently disagree on labels for subtle expressions, and datasets are often imbalanced, with emotions like "fear" or "disgust" being underrepresented.

To address these challenges, we utilized the RAF-DB dataset. Unlike controlled laboratory datasets, RAF-DB provides diverse, real-world images that make classification significantly harder. The dataset consists of 15,000 images, specifically partitioned into 9,800 training samples, 2,500 validation samples, and 2,700 testing samples. This rigorous split ensures that the models are evaluated on their true ability to generalize to unseen, noisy, and unconstrained environments.

3 Techniques and Preprocessing Pipeline

The implemented system integrates a multi-stage preprocessing and detection pipeline followed by a deep learning classification stage.

3.1 Preprocessing Workflow

Before classification, each raw input frame undergoes a rigorous preprocessing sequence to ensure consistency:

- **Face Detection:** We utilize the OpenCV DNN module to locate the face within the image. This module provides higher precision (approx. 90%) than traditional Haar Cascades, especially in low-light or side-profile scenarios.
- **Cropping and Alignment:** Once a face is detected, the region of interest is isolated and cropped. This removes background noise that might confuse the transformer's attention mechanism.
- **Resizing:** Images are resized to 224×224 pixels to match the input specifications of pre-trained transformer models. For baseline comparisons with CNNs, a smaller 48×48 grayscale format was also considered.
- **Normalization and Scaling:** Pixel values are normalized to a standard range (0 to 1) and intensity-scaled based on ImageNet mean and standard deviation values.
- **Data Augmentation:** During training, we apply horizontal flipping and minor rotations to increase the model's robustness to varying facial orientations.

3.2 Model Architectures

- **ViT-Base (google/vit-base-patch16-224):** Utilizing 86 million parameters, this model treats image patches as tokens. It excels at global reasoning but often requires massive data to overcome its lack of local inductive bias.
- **DeiT-Small (facebook/deit-small-patch16-224):** A compact model with 22 million parameters. It introduces a distillation token to learn from a CNN teacher, effectively gaining "convolutional bias" while remaining highly data-efficient.
- **CvT-21 (microsoft/cvt-21):** A hybrid with 32 million parameters. It integrates convolutional layers directly into the transformer blocks to improve spatial locality and feature extraction.

3.3 Training Configuration

All models were trained using PyTorch with the AdamW optimizer for 15 epochs. The learning rate was set to 5×10^{-5} with a linear warmup applied for the first 10% of training steps. We used a batch size of 16 and applied Label Smoothing (0.1) and Dropout (0.1) to prevent overfitting and improve robustness against the label noise present in real-world datasets.

For the best-performing model, DeiT-Small, we analyze the dynamics of the training and the performance in class through the accuracy of the training and validation and the loss curves (Fig. 1–2):

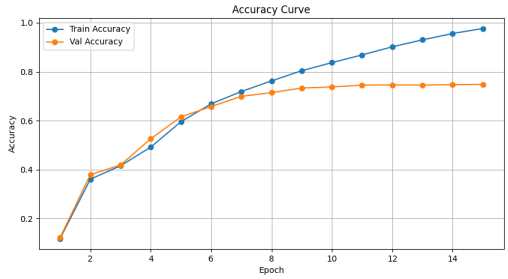

Figure 1: Training and validation accuracy for DeiT-Small over 15 epochs.

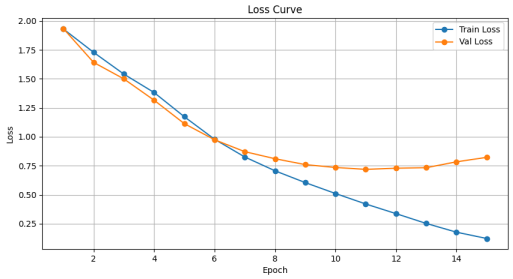

Figure 2: Training and validation loss for DeiT-Small over 15 epochs.

With this setting, the training behavior of DeiT-Small shows fast convergence and a controlled generalization gap. The training and validation accuracy in Fig. 1 both increase sharply in the first 6-8 epochs, showing an efficient transfer of features from the pre-trained model. After that, the validation accuracy plateaus around 74–75%, while the training accuracy keeps increasing steadily toward the perfect performance. This indicates the emergence of mild overfitting, which was expected on medium-scale datasets with class imbalance, like RAF-DB.

This behavior is further supported by the corresponding loss trends shown in Fig. 2. While the training loss decreased monotonically throughout all epochs, the validation loss reaches its minimum within a window of epochs 9 to 11, after which it stabilizes with a slight increase toward the end of training. Notably, the smooth nature of the validation curve, with no sharp oscillations or sudden spikes, provides evidence for stable optimization processes rather than unstable training.

By combining these findings from Fig. 1 and Fig. 2, it can be said that the chosen learning rate, warm-up schedule, and regularization methods effectively control overfitting yet allow the model to fully exploit its representational capacity. Optimal generalization is achieved around mid-to-late training epochs.

4 Human Thought Processes

The computational processing of facial information mirrors several aspects of human perception. When humans interpret emotions, they do not analyze an entire face at once; instead, they focus on localized, salient features such as the furrowing of a brow or the widening of the eyes. This selective attention is modeled by the "Self-Attention" mechanism in transformers, which assigns higher weights to the most informative patches of the face.

Early layers in these models detect low-level features like edges and textures, similar to the human primary visual cortex, while deeper layers combine these into abstract representations of high-level emotional states. Furthermore, both humans and AI models learn through accumulated patterns from experience. The misclassification patterns observed in the AI, such as the frequent confusion between "fear" and "surprise", parallel human cognitive challenges. These two emotions share overlapping visual cues, and without temporal context, even human experts occasionally struggle to resolve the ambiguity.

5 Empirical Analysis

The models were evaluated based on Validation Accuracy, Precision, Recall, and F1-Score. The following results were obtained from the RAF-DB test set:

Model	Parameters	Validation Accuracy	Precision	Recall	F1-Score
DeiT-Small	22 million	0.746	0.621	0.483	0.496
ViT-Base	86 million	0.718	0.542	0.515	0.508
CvT-21	32 million	0.620	0.525	0.421	0.436
DeiT-Small (Weighted Sampling)	22 million	0.668	0.559	0.605	0.577

Table 1: Performance comparison of Vision Transformer models on RAF-DB using macro-averaged metrics. Weighted sampling improves recall and macro F1 at the cost of overall accuracy.

For our best model (DeiT-Small), we examine the 7-class confusion matrices to see whether weighted random sampling actually had an effect on prediction distributions.

Figs. 3 and 4 demonstrate the class-wise performance of DeiT-Small for different training strategies. In the case of standard training (Fig. 3), strong performance is demonstrated on majority classes such as *happy* and *neutral*, since most of the predictions for these categories are concentrated along the diagonal. On the other hand, minority classes such as *disgust* and *fear* are characterized by a high misclassification rate to *a priori* visually similar or high-frequency emotions represented by *happy*, *sad*, and *neutral*. This can be explained by the class imbalance in RAF-DB, which forces the model to favor capturing dominant global features instead of subtle discriminative features of underrepresented emotions.

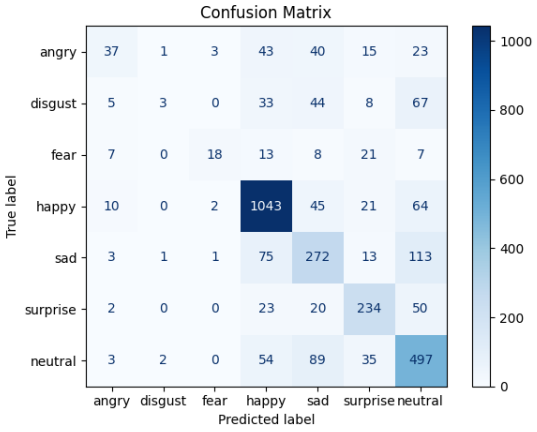

Figure 3: Confusion matrix for DeiT-Small evaluated on the RAF-DB test set.

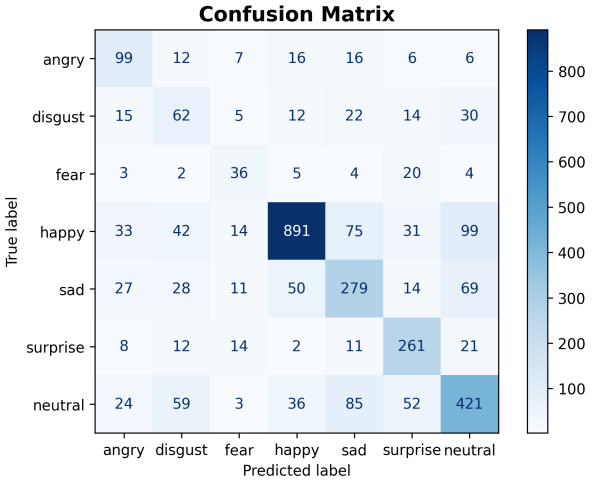

Figure 4: Confusion matrix for DeiT-Small using weighted random sampling.

Note that when weighted random sampling is used during training (Fig. 4), recall noticeably increases for minority classes such as *angry*, *disgust*, and *fear*, as evidence by stronger diagonal presence for these categories. This happens due to more balanced optimization on all emotion classes. Simultaneously, increased confusion among majority classes, especially *happy* and *neutral*, results in reduced overall accuracy. Therefore, weighted sampling sacrifices peak performance on dominant classes while simultaneously yielding higher macro-averaged recall and F1-score. These findings uncover a fundamental trade-off existing in facial emotion recognition between maximizing overall accuracy and providing class-balanced performance, highlighting the importance of sampling strategies when reliable recognition of rare emotions is a priority.

6 Discussion of Findings

A critical finding from this analysis is that model size does not directly correlate with performance. DeiT-Small, despite being four times smaller than ViT-Base, achieved higher accuracy (74.6% vs 71.8%). This highlights the importance of architectural inductive bias; DeiT's ability to inherit spatial knowledge from CNN teachers through distillation makes it far more effective for medium-scale datasets.

Furthermore, the results confirm that "Recall" remains the greatest challenge in FER. The models struggled to correctly identify underrepresented classes such as "disgust" and "fear" because the dataset is dominated by "happy" and "neutral" samples. This suggests that future work must focus on addressing class imbalance through weighted loss functions or oversampling techniques.

The confusion matrix in Fig. 3 reveals clear class-dependent performance differences. Dominant emotions such as *happy* and *neutral* achieve the highest true positive rates, reflecting their strong representation in the training data. In contrast, minority classes including *disgust* and *fear* exhibit substantial misclassification, most frequently being predicted as visually similar or high-frequency categories such as *happy*, *sad*, and *neutral*. This behavior indicates that while DeiT-Small learns robust global facial patterns, it struggles to capture subtle discriminative cues for underrepresented emotions, leading to consistently lower recall for these classes.

7 Outside Code

This project relied on several mature and widely adopted open-source frameworks to ensure both reproducibility and practical deployability. The HuggingFace Transformers library was used for accessing pre-trained Vision Transformer architectures and managing model configuration, allowing rapid experimentation while maintaining architectural consistency across evaluations. PyTorch served as the core deep learning framework, providing flexibility in model training, gradient optimization, and performance monitoring throughout the development process.

OpenCV was utilized for real-time video capture and face detection, enabling efficient preprocessing within a live inference pipeline. The OpenCV DNN module, in particular, allowed the system to maintain high detection accuracy while minimizing latency, which is critical for real-time applications. Each component of the software stack was customized and integrated to ensure that the end-to-end pipeline remained non-blocking, computationally efficient, and suitable for deployment in real-world environments.

8 Expansion and Improvement

There are several promising directions for extending this work and addressing the limitations observed in the current system. One of the most impactful improvements would be the transition from static image-based recognition to video-based emotion analysis through temporal modeling. Facial expressions often evolve over a short sequence of frames, with critical information conveyed through subtle changes in muscle movement known as micro-expressions. Incorporating temporal architectures such as Long Short-Term Memory (LSTM) networks, Temporal Convolutional Networks, or Video Vision Transformers would allow the model to capture motion dynamics and temporal consistency that are not present in single-frame inputs. By modeling how expressions form and decay over time, the system could better disambiguate visually similar emotions such as fear and surprise, which frequently differ more in temporal progression than in spatial appearance alone.

Another important avenue for improvement lies in increasing robustness to real-world occlusions. In unconstrained environments, faces are often partially obscured by accessories such as masks, glasses, hats, or by hands and hair. Training the model with synthetic occlusions, such as randomly masking regions of the face or overlaying realistic occluding objects, would encourage the network to rely on multiple redundant cues rather than overfitting to a single facial region. This form of structured augmentation would align well with transformer attention mechanisms, enabling the model to dynamically shift focus to visible and informative regions of the face when others are unavailable. Improved occlusion robustness is particularly critical for deployment in public or clinical settings, where full facial visibility cannot be guaranteed.

Finally, expanding and diversifying the training data remains essential for improving both performance and fairness. Facial expressions vary significantly across different ethnicities, age groups, and cultural contexts, and models trained on limited demographic distributions risk encoding systematic bias. Increasing dataset diversity, either through the inclusion of additional FER datasets or through carefully designed augmentation strategies, would help ensure that the learned representations generalize more equitably across populations. Moreover, addressing class imbalance through weighted loss functions, focal loss, or targeted oversampling of underrepresented emotions could substantially improve recall for critical but rare classes such as disgust and fear. Through these

efforts, we also learned that architectural efficiency and inductive bias often outweigh raw model size in real-world settings; data-efficient transformers such as DeiT-Small generalize more reliably than larger models when dataset scale and quality are constrained. This project further emphasized that successful FER systems must be designed as complete pipelines, where data quality, preprocessing reliability, and real-time constraints are treated as equally important as classification accuracy, reinforcing the need to balance performance with robustness and deployability.

9 Conclusion

This study demonstrates that effective Facial Emotion Recognition in real-world settings depends more on architectural efficiency and inductive bias than on sheer model scale. Through a controlled evaluation of ViT-Base, CvT-21, and DeiT-Small on the RAF-DB dataset, we show that smaller, data-efficient transformers can outperform significantly larger models when trained on medium-scale, noisy data. DeiT-Small achieved the highest validation accuracy despite having four times fewer parameters than ViT-Base, reinforcing the conclusion that parameter count alone does not determine performance. Its use of knowledge distillation enables the model to inherit convolutional priors, allowing it to better capture local facial features that are critical for emotion recognition under varying pose, lighting, and occlusion.

At the same time, the results highlight persistent challenges that remain unresolved in FER. All evaluated models exhibited strong performance on dominant emotion classes while struggling with underrepresented categories such as fear and disgust, where recall remained consistently low. This imbalance reflects both dataset limitations and the inherent ambiguity of subtle facial expressions, underscoring the need for class-aware training strategies and more representative data. Despite these challenges, the alignment between transformer self-attention mechanisms and human selective perception provides a compelling foundation for future systems. By extending this work toward temporal modeling, stronger augmentation, and broader demographic coverage, FER systems can move closer to reliable, ethical, and human-centered deployment in real-world applications.

References

Dev-ShuvoAlok. "RAF-DB Dataset." Kaggle, 20 Sept. 2023, www.kaggle.com/datasets/shuvoalok/raf-db-dataset.

"Facebook/Deit-Small-Patch16-224 · Hugging Face." Facebook/Deit-Small-Patch16-224 · Hugging Face, huggingface.co/facebook/deit-small-patch16-224. Accessed 19 Dec. 2025.

"Google/VIT-Base-Patch16-224 · Hugging Face." Google/Vit-Base-Patch16-224 · Hugging Face, huggingface.co/google/vit-base-patch16-224. Accessed 19 Dec. 2025.

Hosna, Asmaul, et al. "Transfer Learning: A Friendly Introduction." Journal of Big Data, vol. 9, no. 1, Oct. 2022. Crossref, https://doi.org/10.1186/s40537-022-00652-w.

"Introduction to Transfer Learning." GeeksforGeeks, www.geeksforgeeks.org/machine-learning/ml-introduction-to-transfer-learning/. Accessed 19 Dec. 2025.

Li, Yanghao, et al. "Benchmarking Detection Transfer Learning with Vision Transformers." Version 1, arXiv, 2021, https://doi.org/10.48550/ARXIV.2111.11429.

"Microsoft/CVT-21 · Hugging Face." Microsoft/Cvt-21 · Hugging Face, huggingface.co/microsoft/cvt-21. Accessed 19 Dec. 2025.

Usman, Mohammad, et al. "Analyzing Transfer Learning of Vision Transformers for Interpreting Chest Radiography." Journal of Digital Imaging, vol. 35, no. 6, July 2022, pp. 1445–62. Crossref, https://doi.org/10.1007/s10278-022-00666-z.