# Robust Facial Emotion Recognition with Subject Conditioning and Continuous Facial Motion Detection

**Zhenghao Jin**
ECE
zhenghao@andrew.cmu.edu

**Putian Wang**
ECE
putianw@andrew.cmu.edu

**Veronica Zhao**
ECE
veronicz@andrew.cmu.edu

## 1 Introduction

Facial emotion recognition (FER) is a critical technology for next-generation human-computer interaction, enabling applications from adaptive tutoring systems to driver safety monitoring. For these systems to be effective, their predictions must be accurate and stable and reliable over time. However, real-world conditions present significant challenges: faces are often partially occluded, with constant variations in pose and lighting. Furthermore, existing models struggle to generalize across different individuals and often fail to capture subtle, low-amplitude expressions.

Although modern deep learning approaches have improved on the classical methods, they still have key limitations [3]. CNN-based models, though powerful, often produce unstable frame-by-frame predictions and perform poorly when faced with data from new subjects or environments. Sequence models like RNNs can improve temporal consistency but at a high computational cost, making them less suitable for real-time applications.

Two primary gaps remain in current research: (1) inadequate adaptation to the unique facial morphology of each input, and (2) insufficient use of short-term facial dynamics, which are crucial for distinguishing near-neutral expressions. Our project directly addresses these gaps. We propose to build upon strong CNN baselines by introducing two key innovations: subject-specific conditioning and a temporal model based on facial landmark motion, aiming to create a more robust and reliable FER system for real-world applications.

## 2 Related Work

We adopted our baseline model from a previous study to ground our image-based expression recognition (FER) experiments. We will also adopt a video-based FER technique to further boost the performance of our model mentioned in another study.

### 2.1 Image-based FER on RAF-DB

For static images, we follow Stoychev and Gunes' setup in The Effect of Model Compression on Fairness in Facial Expression Recognition and use their RAF-DB baseline [7]. The authors implemented a compact CNN classifier without architectural bells and whistles and then studied compression and fairness effects on top of this backbone. The model architecture has only basic layers, such as the convolutional layer, the pooling layers, and the dropout layers. We will reproduce this baseline and its train/validation protocol as our image classifier, treating the uncompressed model as our baseline model.

### 2.2 Video-based FER on DFEW

For dynamic expressions, we will not adopt the DFEW paper's models as baselines [6]. Instead, we will treat DFEW as a technical reference for accuracy-improving design choices. The DFEW work benchmarks spatiotemporal CNNs under a five-fold protocol (splits fd1–fd5) and evaluates

with WAR and UAR. It further shows that an Expression-Clustered Spatiotemporal Feature Learning module (EC-STFL) improves both the C3D and 3D ResNet-18 baselines, with gains visible in class-wise recalls (e.g., happy, sad, neutral) and modest transfer benefits when pretraining in DFEW and fine-tuning on AFEW 7.0. In our study, we maintain our own video baselines and use DFEW insights, such as expression-sensitive feature clustering and spatio-temporal aggregation, as optional enhancements to improve recognition accuracy, rather than as baseline architectures or evaluation protocols.

# 3 Methods

Relating to our problem statement, we target two issues: (1) inadequate adaptation to the unique facial morphology of each input; (2) insufficient use of short-term facial dynamics, which are crucial for distinguishing near-neutral expressions.

For the first issue, we hypothesize that incorporating relative distances among facial landmarks (L2 norms) as explicit features can improve the robustness to person-specific morphology. Because expressions are ultimately manifested by the geometric configuration of facial components, landmark locations and their pairwise distances form a direct, discriminative representation for FER.

For the second issue, we note that single-image analysis lacks temporal context. The neutral face of a subject can resemble a mild smile in a single frame, leading to misclassification. With short sequences, frame-to-frame changes provide additional cues that help the model learn the neutral baseline of an individual and thus classify expressions more reliably.

For our current method, we focus first on quantifying the determinative power of landmark-distance features in single-image FER.

## 3.1 Our Current Method

### 3.1.1 Data Engineering

First, we perform landmark extraction. We use the Python face alignment package to extract facial landmarks. To keep the model compact and efficient, we currently test five landmarks per face image. We will consider extending to 68 landmarks based on subsequent experiments. Second, we perform image-clarity stratification. We computed the Laplacian variance of each image and divided the dataset into three clarity levels with equal sample counts, ensuring a balanced distribution per clarity class. Third, we perform relabeling with metadata. For all training images, we augment the labels with the extracted landmarks and the assigned clarity level.

### 3.1.2 Modeling and Training

For each clarity level, we train a separate MLP classifier on landmark-distance features. We perform cross-validation hyperparameter tuning, including optimizer selection and learning-rate tuning, to maximize validation accuracy.

### 3.1.3 Inference Pipeline

Given a test image, our pipeline: extracts five landmarks using face-alignment; assigns a clarity level using the same Laplacian variance criterion as in training; routes the sample to the corresponding MLP for the final FER prediction.

## 3.2 Our Method's Performance Compared to Baseline Method

We deployed and evaluated our image-based FER baseline. The training and validation losses on RAF-DB are shown in Figure 1. On the RAF-DB test set, this baseline achieves an accuracy of 82.46%.

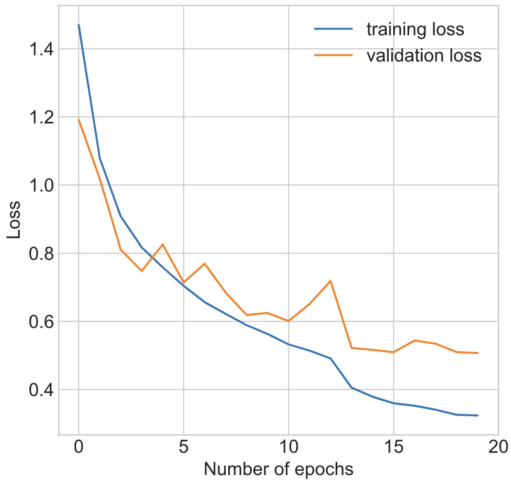
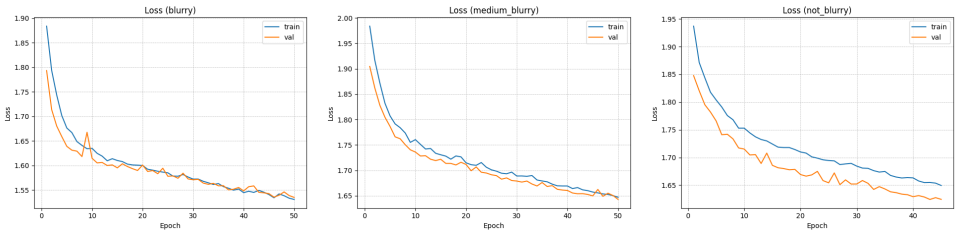



Figure 1: Train and validation loss of the baseline model for RAF-DB

We also evaluated our proposed pipeline. Its training and validation losses on RAF-DB are shown in Figure 2, and the corresponding accuracies are shown in Figure 3. On the RAF-DB test set, the pipeline achieves an accuracy of 37.20%.

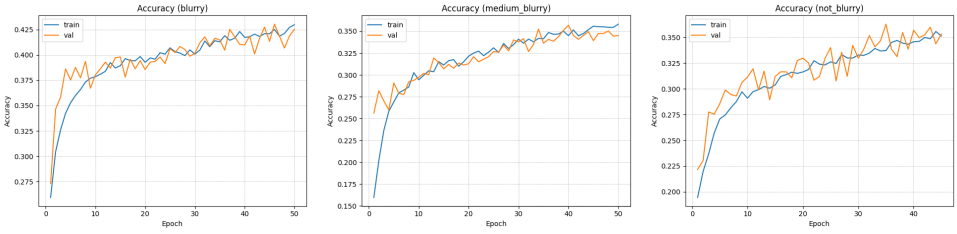
Overall, the current pipeline underperforms the baseline, which is expected: for this diagnostic study, we deliberately restrict the input to a single feature family, pairwise distances among only five facial landmarks, to probe the determinative power of landmark geometry in single-image FER. Despite this constraint, the observed accuracy is markedly higher than anticipated for such a compact representation, indicating that landmark geometry is indeed highly informative for expression recognition.

These results provide a clear direction for future work. Depending on follow-up experiments, we may de-prioritize integrating landmarks into the original baseline and instead pursue a landmarks-only approach, scaling the number of landmarks (e.g., from 5 to 68) and refining the classifier to obtain a high-efficiency, competitive FER model.




(a) Train-Val Loss (Blurry) (b) Train-Val Loss (Mid Blurry) (c) Train-Val Loss (Not Blurry)

Figure 2: Train-Val Loss for the Models




(a) Accuracy (Blurry) (b) Accuracy (Mid Blurry) (c) Accuracy (Not Blurry)

Figure 3: Train-Val Accuracy for the Models

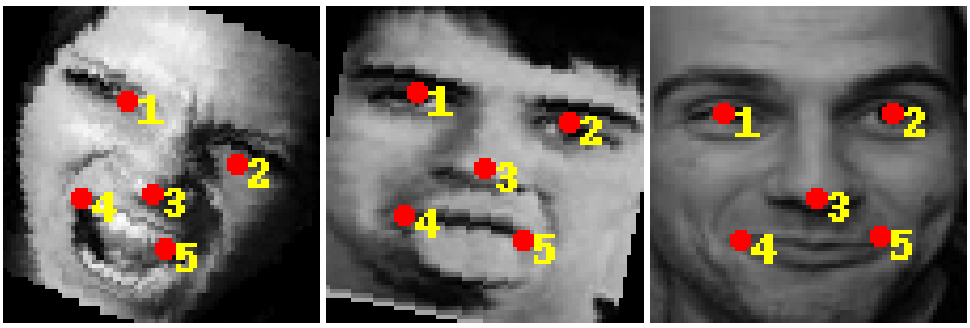
## 4 Results

As shown in Figure 1, the baseline model on RAF-DB exhibits a stable convergence in the training split. Using accuracy as an evaluation metric, it reaches a test accuracy of 82.46%. The qualitative results are provided in Figure 4.

(a) Prediction: Angry (b) Prediction: Disgust (c) Prediction: Happy

Figure 4: Sample Predictions from Baseline Model

As shown in Figure 3, our pipeline achieves an overall test accuracy of 37. 20% on RAF-DB when aggregating predictions in the three clarity strata: blurry, moderately blurry and not blurry. In Figure 2, all three clarity-specific models demonstrate gradual convergence in their respective training subsets. Representative visual outputs for our pipeline are shown in Figure 5.




(a) Prediction: Angry (b) Prediction: Disgust (c) Prediction: Happy

Figure 5: Sample Predictions from Our Models

## 5 Discussion and Conclusion

### 5.1 Efficiency and Latency

In our current experiments, using only landmark-based features yields very low inference latency and high efficiency. The landmark extraction itself takes approximately 0.08 seconds per image, and the downstream classifier adds negligible overhead.

### 5.2 Effectiveness of Landmarks

Although the current landmark-only model underperforms the baseline in accuracy, it relies on only five landmarks. This result indicates that landmark geometry is highly determinative for FER and that there remains substantial headroom for improvement within a landmark-first design.

### 5.3 Next Steps for Single-Image FER

We will increase the number of extracted landmarks (e.g., from 5 to 68), replace the landmark extractor with a lower-latency model, and apply structured pruning to the classifier. Because landmark features exhibit spatial structure, targeted pruning can reduce latency while also improving generalization. Depending on forthcoming ablations, we may de-prioritize integrating landmarks into the original baseline and instead pursue a landmarks-only approach to achieve a high-efficiency, competitive FER model.

### 5.4 Toward Temporal Modeling

After finalizing a strong single-image pipeline, we will extend the approach to short facial sequences. Incorporating short-term facial dynamics should better disambiguate near-neutral expressions from subtle non-neutral ones while retaining landmark-distance features as a core signal.

## References


[1] Li, S., Deng, W. and Du, J. (2017). Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2584–2593.

[2] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. (2013). Challenges in representation learning: A report on the ICML 2013 workshop. In *JMLR: Workshop and Conference Proceedings* (Vol. 27, pp. 1-10).

[3] Ali, M. F., Khatun, M., & Turzo, N. A. (2020). Facial emotion detection using neural network. *International Journal of Scientific & Engineering Research*, 11(8), 1318–1325.

[4] Prajapati, D. (2025). Balanced RAF-DB Dataset (75x75 grayscale). Kaggle. Retrieved from `https://www.kaggle.com/datasets/dollyprajapati182/balanced-raf-db-dataset-7575-grayscale`

[5] Sun, K., et al. (2020). High-Resolution Representations for Labeling Pixels and Regions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7907–7916.

[6] Jiang, X., et al. (2020). DFEW: A Large-Scale Database for Recognizing Dynamic Facial Expressions in the Wild. arXiv preprint arXiv:2008.05924. Retrived from `https://dfew-dataset.github.io/download.html`

[7] Stoychev, S. & Gunes, H. (2023) The Effect of Model Compression on Fairness in Facial Expression Recognition. In J.-J. Rousseau & B. Kapralos (eds.), *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges: Montreal, QC, Canada, August 21–25, 2022, Proceedings, Part IV*.