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 Problem Statement

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 not only accurate but also 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.

While modern deep learning approaches have improved upon 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 Data

To address the challenge of real-world facial expression recognition, we will utilize three standard, publicly available datasets. The datasets will allow us to train robust models and evaluate their cross-dataset generalization capabilities.

2.1 RAF-DB (Real-world Affective Face Database)

We will use an augmented RAF-DB dataset as our primary dataset for training and evaluation [4]. RAF-DB is a large-scale, "real-world" database containing 62,916 facial images of 75x75 pixels sourced from the internet, stored as JPEG format [1]. Each image was independently annotated by about 40 human coders, ensuring high-quality labels for seven basic emotions. The total size of this dataset is approximately **130 MB**, making it highly convenient for faster training iterations.

2.2 FER-2013

This dataset, introduced as part of a Kaggle challenge, will serve as a foundational benchmark and for supplementary training [2]. It consists of 35,887 grayscale images of size 48x48 pixels, categorized into the same seven basic emotions as RAF-DB's. The data is formatted within a single CSV file, with pixels separated by spaces. While also captured in real world, its images are of lower resolution and have noisier labels compared to RAF-DB. The total size of this dataset is approximately **660 MB**. We will leverage it to validate our model's robustness.

2.3 DFEW (Dynamic Facial Expression in-the-Wild)

We will utilize DFEW dataset for evaluating our temporal modeling approaches, containing over 16,000 video clips from movies with challenging conditions including extreme illumination, occlusions, and pose changes [6]. It enables validation of our landmark-motion modeling and temporal consistency objectives for dynamic expression recognition from onset through apex to offset.

3 Method

We propose a five phases deep learning framework that progressively addresses subject variability, temporal dynamics, regional contradictions, and geometric robustness to improve facial emotion recognition accuracy over standard CNN baselines. Each phase introduces specialized neural architectures—subject-conditional routing, spatio-temporal graph networks, adversarial fusion consensus, and geometric feature learning—that can operate independently or be integrated for maximum performance.

3.1 Phase I: Instance-Aware, Dataset-Conditioned Recognition

We characterize each detected face using a compact set of geometry and appearance features (e.g., inter-ocular distance, face aspect ratio, eye–mouth distances, and CNN embeddings from facial regions), normalized by face size to ensure scale invariance. These descriptors are denoised with PCA and projected with LDA to enhance separability among morphology/appearance regimes. A lightweight routing classifier (e.g., SVM or nearest-prototype) assigns the instance to the closest group within our training distribution, after which the emotion recognizer—the baseline CNN head equipped with group-aware parameters or adapters—operates on the routed sample. By reducing intra-class variance and mitigating cross-domain mismatch with negligible runtime overhead, this instance-aware conditioning improves accuracy while preserving a near real-time budget.

3.2 Phase II: Landmark–Motion Temporal Modeling for Subtle Expressions

We model short videos (or image series) by tracking facial landmarks over a lookback window and learning dynamics that disambiguate low-amplitude expressions (e.g., neutral↦happy with a slight smile). For each frame we detect and normalize K landmarks (similarity transform on eye centers), form per-frame geometry features (relative coordinates, inter-point distances, aspect ratios) and first/second-order motion features (Δ, Δ^2), and build a spatio-temporal graph whose spatial edges follow the facial topology (brows/eyes/nose/mouth) and temporal edges connect the same landmark across adjacent frames. A lightweight *spatio-temporal GCN* (or a Temporal Transformer with landmark positional encodings) encodes these trajectories into a clip embedding; optionally, we fuse a small appearance branch (CNN on a face crop) via late fusion to retain texture cues while keeping identity effects minimal. The model outputs clip-level emotion logits with standard cross-entropy, plus auxiliary objectives that strengthen subtle dynamics: (i) phase consistency (onset/apex/offset) estimated via velocity magnitude to encourage smooth progression, and (ii) temporal contrastive loss that pulls frames from the same clip/emotion phase together and pushes different emotions apart. By leveraging motion direction and velocity—signals missing from single images—this approach reduces confusion in near-neutral cases while preserving near real-time efficiency (small K, shallow layers, and short windows).

3.3 Phase III: Adversarial Fusion Consensus with Compositional Emotion Primitives

We propose Adversarial Fusion Consensus (AFC), where fusion is framed as a multi-agent game in which each facial region maintains a critic network that learns to identify when fusion predictions are inconsistent with that region's reliable patterns. Specifically, during training, each critic observes countless examples of when its region is reliable versus unreliable—the eye critic learns that when eyes show sadness (narrowed, downcast) but fusion predicts happiness, the fusion is likely wrong; the mouth critic learns that isolated mouth smiles without eye involvement often indicate forced expressions. These critics are trained to output high scores when fusion aligns with their region's reliable patterns and low scores when fusion ignores important signals from their region. The fusion network must then learn to produce predictions that minimize criticism from all regions simultaneously—if it predicts 'happy' when eyes clearly show sadness, the eye critic penalizes it heavily. Through training critics to better identify inconsistencies stacking training fusion to avoid criticism, the system naturally discovers complex resolution patterns: trusting eyes over mouth for genuine emotion, recognizing when eyebrow tension overrides a smile, or identifying when regional disagreement itself indicates mixed emotions. The critics are implemented as lightweight 3-layer MLPs evaluating (fusion_output, own_region_features) pairs, while the fusion network uses a transformer architecture processing all regional features. As an enhancement if time permits, regions could output learned primitive activations (e.g., 'eye_crinkle,' 'lip_corner_raise') instead of emotion probabilities, allowing the critics to enforce more fine-grained consistency rules, though the core AFC framework operates effectively with standard emotion probabilities from existing region-specific networks.

3.4 Phase IV: Geometric-Based Single Image Emotion Recognition via Facial Landmarks

The goal is to develop a classifier for single image that predicts emotions based exclusively on the relative geometric configuration of facial features, making it inherently robust to photometric variations like illumination and skin texture. To achieve this, we will employ a two-stage pipeline where we first process each facial image to extract a set of key facial landmarks (e.g., corners of the eyes, eyebrows, and mouth), initially leveraging a state-of-the-art pre-trained model [5]. From these landmarks, we will construct a feature vector composed of normalized Euclidean distances between semantically significant pairs, which will be made invariant to face scale and orientation by normalizing against a stable reference like the inter-ocular distance. This normalized geometric vector will then serve as the input to a dedicated neural network, such as a Multi-Layer Perceptron, trained to map these geometric configurations to the final predicted emotion labels. The primary advantage of this approach is its focus on the underlying structural changes of the face, which has the potential to be more robust against the 'real world' challenges, such as lighting and cosmetic differences.

3.5 Phase V: System Integration and Inference Flow

We integrate all prior phases behind a shared front end that performs face detection, landmarking, and geometric normalization (eye-centered similarity transform) to yield consistent inputs for every branch. The assembled system exposes multiple predictors: the instance-aware subject-conditioning branch, the landmark-motion temporal branch for subtle expressions, and supporting image/geometry branches; each outputs calibrated class logits and an uncertainty score. A lightweight learned fusion module (logistic/MLP or a tiny transformer over tokens representing the branches) ingests per-branch logits, confidences, and simple quality indicators (e.g., landmark reliability) to produce final logits; for video, the temporal branch provides clip-level logits while other branches are temporally pooled before fusion. Training proceeds in three steps: (1) pre-train each branch to convergence, (2) calibrate with temperature scaling on a held-out split, and (3) train the fusion module with branches frozen, followed by a brief joint fine-tune with a small learning rate; finally, we distill the fused predictor into a compact student to meet latency targets. The stacker learns fallbacks and routing implicitly—relying more on image-based branches when landmarks are poor and letting the temporal branch dominate subtle-expression clips—while deployment exports both the fused model and the distilled student under a single pre-processing pipeline and batch scheduler for near real-time inference.

4 Evaluation Metrics

We evaluate on FER-2013, RAF-DB, and DFEW using their standard splits, and report cross-dataset transfer (train on RAF-DB, test on FER-2013 and vice versa; split DFEW for training and testing purpose). Primary accuracy metrics are *macro-F1* (main), per-class precision/recall, balanced accuracy, and confusion analysis with emphasis on common confusions (fear↔surprise, sadness↔neutral). For videos, temporal stability is quantified by (i) flip rate between adjacent frames, (ii) label edit distance after simple temporal smoothing, and (iii) onset–apex–offset coherence via correlation between predicted intensity and velocity-based phase estimates. Calibration is measured by Expected Calibration Error (ECE), Negative Log-Likelihood (NLL), and Brier score on a held-out set, with temperature scaling applied once per model/branch and re-checked post-fusion. Robustness is assessed across pose buckets (yaw/pitch), occlusion stress (synthetic regional cutout, landmark dropout), blur/JPEG noise, and illumination jitter; we report macro-F1 deltas relative to clean conditions. We also measure efficiency—per-frame and per-clip latency (ms), model size (MB), and peak memory—targeting near real time (< 50 ms/frame on GPU or < 120 ms/frame on CPU). Ablations include: baseline CNN (image-only); +subject conditioning; +landmark–motion temporal model; +supporting branches; full fusion; plus removals of prototype guidance (class-agnostic attention only) and calibration. Statistical significance is established with 95% bootstrap confidence intervals and paired tests (approximate randomization) on macro-F1 improvements.

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`