Confidence-Gated Calibration for Video-based Facial Expression Recognition in-the-Wild

Andrey V. Savchenko[1, 2]

[1]Sber AI Lab, Moscow, Russia and [2]ITMO University, Saint Petersburg, Russia

Introduction & Motivation

- In-the-wild facial expression recognition (FER) suffers from **frame-level noise**, occlusions, pose/lighting shifts, and **severe class imbalance** (Neutral/Happiness dominate; Fear/Disgust are rare).
- Heavy temporal models (Transformers, TCNs) are computationally expensive and often overfit to frequent classes.
- **Goal:** A lightweight, calibration-aware pipeline that maximizes macro F1 without complex spatiotemporal architectures.

Key Idea

"Trust strong pretrained predictions when confident; fall back to a calibrated, task-specific classifier when ambiguous. Smooth locally to remove noise."

Proposed Method Overview

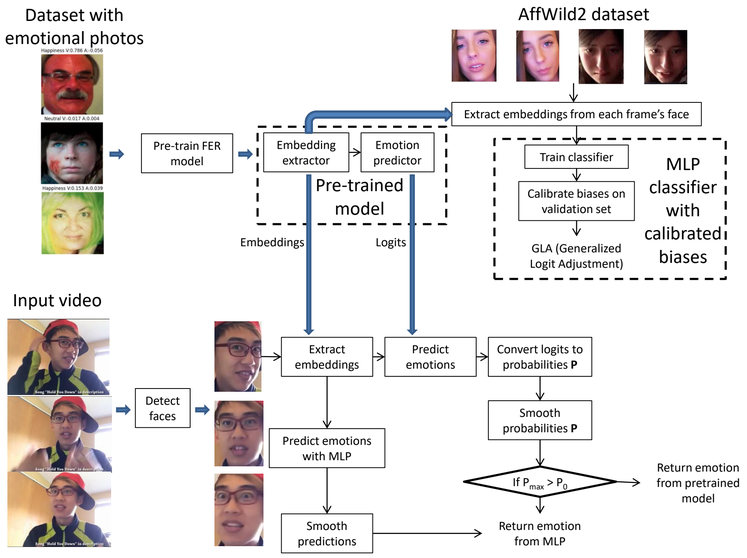

Figure 1. Overview of the proposed pipeline: training (top) and inference (bottom).

Training:

1. Pre-train the lightweight neural network to recognize expressions from static photos using an external dataset (e.g., AffectNet);
2. Use the pre-trained model to extract emotional embeddings $\mathbf{x}_n$ for faces cropped from the frames of the AffWild2's training set,
3. Train MLP for frame-level prediction of facial expressions;
4. Leverage logit adjustment to calibrate biases of the trained MLP to take into account the imbalance of emotional classes.

Inference:

1. Detect facial regions in each frame of an input video;
2. Feed it into pre-trained neural network to extract embeddings and posterior probabilities of expressions from AffectNet
3. Compare the maximal probability with a certain threshold to verify the reliability of decision of a pre-trained model;
4. Feed the facial embeddings into trained MLP if the pre-trained model is not certain;
5. Smooth outputs for several consequent frames to mitigate the noise in frame-level predictions.

Core Technical Components

1. **Feature Extraction:** Pretrained `EmotiEffNet-B0` (AffectNet) $\rightarrow$ facial embeddings + raw probabilities.
2. **Confidence-Gated Calibration:**
   - If $\max(p_{\text{smoothed}}) > p_0$ $(0.8\text{–}0.9)$ $\rightarrow$ use pretrained prediction.
   - Else $\rightarrow$ MLP trained on Aff-Wild2 with **Generalized Logit Adjustment (GLA)** to correct class priors: $\mathbf{p}^{\text{MLP}}(t) = \text{softmax}(\mathbf{z}(\mathbf{x}(t))) \cdot \mathbf{b}^*$
3. **Temporal Smoothing:** Sliding window averaging ($T+1$ frames) reduces frame-level volatility.
4. **Optional Audio Fusion:** Late blending with `acoustic` features: $\mathbf{p}(t) = w \cdot \mathbf{p}^{\text{feat}}(t) + (1-w) \cdot \mathbf{p}^{\text{audio}}(t)$

Why It Works?

- **GLA Calibration:** Shifts decision boundaries for minority classes without retraining the backbone.
- **Confidence Gating:** Exploits strong visual priors on "easy" frames, avoids MLP overcorrection.
- **Short-Range Smoothing:** Removes annotation noise without RNNs/Transformers.
- **Modular & Efficient:** $\sim 10\times$ faster than temporal ensembles, easy to deploy/debug.

Experimental Setup

- **Dataset:** Aff-Wild2 EXPR challenge data (248 train / 70 val videos) provided by ABAW-10 organizers.
- **Labels:** 8 target classes (Neutral, Anger, Disgust, Fear, Happiness, Sadness, Surprise, or Other); the AffectNet class "Contempt" is excluded during direct gating to Aff-Wild2 predictions.
- **Backbone:** EfficientNet-based emotion recognizer from EmotiEffLib, pretrained on AffectNet for 8 basic expressions.
- **Audio embeddings:** wav2vec 2.0.

Experimental Results (Validation Set)

Method	Modality	F1-score	Accuracy
Baseline VGGFACE (MixAugment)	Faces	25.0	-
CLIP+LSTM	Video	34.1	-
CLIP	Faces	36.0	-
EfficientNet-B0	Faces	40.2	-
V-NAW	Video	41.81	-
SSL + Temporal+ Post-process	Audio/video	44.43	-
Best EfficientNet + wav2vec	Audio/video	44.59	55.32
CLIP+TCN	Video	46.51	-
MAE+Transformer feature fusion	Audio/video	55.55	-
wav2vec 2.0, focal loss	Audio	30.50	40.08
wav2vec 2.0, focal loss, GLA	Audio	32.33	42.15
wav2vec 2.0, focal loss, GLA, smoothing	Audio	38.75	51.74
EmotiEffNet	Faces	38.68	47.59
EmotiEffNet, calibration	Faces	41.40	51.62
EmotiEffNet, confidence-gated calibration	Faces	42.50	52.22
EmotiEffNet, focal loss	Faces	38.99	50.22
EmotiEffNet, focal loss, calibration	Faces	40.18	51.90
wav2vec 2.0+EmotiEffNet, calibration	Audio/video	43.11	53.40
EmotiEffNet, calibration, smoothing	Faces	44.85	55.41
EmotiEffNet, confidence-gated calibration + smoothing	Faces	45.79	55.69
EmotiEffNet + wav2vec 2.0, calibration, smoothing	Audio/video	46.04	57.50
EmotiEffNet + wav2vec 2.0, confidence-gated calibration + smoothing	Audio/video	47.40	57.98

Each component contributes: calibration (+2.72), confidence gate (+1.10), smoothing (+4.35), audio fusion (+1.61). The final multimodal system achieves **47.40% F1**, significantly outperforming the challenge baseline.

Leaderboard (ABAW-10 Test Set)

Team	Test F1 (%)
EagleonPamir1	39.1
Our approach	**38.6**
USTC-IAT-United	36.0
IMLAB	32.0
Baseline VGGFACE	22.5

- **2nd Place** (Test F1: 38.6%), trailing winner by only 0.5% despite a much simpler architecture.
- Our lightweight pipeline (EfficientNet + MLP + calibration + smoothing) achieves competitive performance with $2\times$–$10\times$ fewer parameters than large transformer or ensemble methods.

Handling Class Imbalance

Weighted softmax loss during training:

$$\mathcal{L}_{\text{EXPR}} = -\log \text{softmax}(z(\mathbf{x}_n)_{y_n}) \cdot \max\left(1, \frac{N_{y_n}}{N}\right)$$

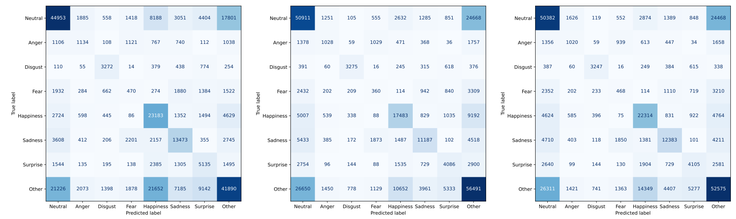

Figure 2. Confusion matrices: (a) EmotiEffNet; (b) EmotiEffNet, calibration; (c) EmotiEffNet, confidence-gated calibration

- Calibration visibly reduces confusion between frequent and minority classes.
- Temporal smoothing cuts frame-level volatility, the main error source.

Conclusion & Takeaways

- A **simple, modular pipeline** can match or exceed complex spatiotemporal models when calibration and temporal smoothing are applied.
- Bias calibration alone yields a clear macro-F1 improvement over the raw pretrained baseline.
- Confidence gating helps exploit easy frames without forcing the MLP to handle every case.
- Short temporal smoothing delivers one of the largest gains, indicating that local prediction instability is a major error source.
- Simple late audio fusion remains beneficial when visual evidence is weak or ambiguous.

Acknowledgments: Ministry of Economic Development of the Russian Federation (IGK 000000C313925P4C0002).