Full Report - Real-time FER - 2025-10-11

Teacher Model Training Report

Executive Summary

- - Chosen teacher: ResNet18 + EfficientNet-B3 weighted ensemble (0.7 : 0.3, T=1.0).
- - Performance: Accuracy 0.8051, Macro-F1 0.7934, Minority-F1 0.7400, ECE 0.627 (ensemble analysis).
- - Best single: ResNet18 (polish) Macro-F1 0.7701 (Acc 0.7814); EfficientNet-B3 close (Acc 0.7817, Macro-F1 0.7627).
- - Calibration: Temperature scaling is applied. Single-teacher calibration metrics (ECE/Brier/NLL) are included in Section 9.1.

1. Introduction and Background

Facial Expression Recognition (FER) benefits from a strong, well-calibrated teacher model to supervise student distillation, guide ensemble design, and anchor longitudinal comparisons. Establishing a robust teacher baseline reduces the risk of propagating bias or poor calibration to lightweight students intended for real-time deployment.

Objectives:

- - Compare single-architecture teachers (ResNet18, EfficientNet-B3, ConvNeXt-Tiny, ViT Tiny/Small).
- - Quantify improvements from pairwise and multi-teacher ensembles (CNN-only vs CNN+ViT hybrid).
- - Identify the most stable and best-performing teacher ensemble for downstream knowledge distillation.

2. Baseline Experiments

- - Random baseline (7 classes): 14.28% expected accuracy.
- - Earliest naive CNN prototype: ~40% accuracy (historical internal run); insufficient for demo quality and minority classes.
- - Limitations observed: poor calibration, weak minority recall (disgust/fear/sad), and limited capacity to benefit from margin-based losses.

3. Backbone Exploration

Evaluated families: ResNet18/50, EfficientNet-B3, ConvNeXt-Tiny, ViT Tiny/Small. We balanced accuracy, minority F1, and compute cost.

- - ResNet18: strong speed/robustness; polished single-model macro-F1 0.7701; minority-F1 0.7164.
- - EfficientNet-B3: best single teacher; macro-F1 0.7627; minority-F1 0.6988; accuracy 0.7817.
- - ConvNeXt-Tiny: macro-F1 0.7635 at early peak (epoch 11); minority-F1 0.6968; accuracy 0.7833; later instability risk.
- - ResNet50: explored; accuracy gains vs RN18 limited relative to added compute; not adopted.
- - ViT Tiny/Small: require tuned warmup and layer-wise LR; deferred for baseline but considered for hybrid.

Trade-off: B3 offers the best standalone balance. RN18 complements B3 via different error profile, motivating pairwise ensemble.

3.1 Single-Model Results (summary)

See Section 9.1 for the full single-teacher results table. In brief: ResNet18 (polish) is the strongest standalone teacher by Macro-F1, with EfficientNet-B3 very close; ConvNeXt-Tiny peaks early but shows later instability.

4. Additional Modules & Techniques

- - CBAM-lite: small, inconsistent gains; not critical to final selection.
- - MixUp / CutMix / MixCut: modest benefits in some runs; careful tuning needed to avoid minority recall regressions.
- - EMA, SAM, Center Loss, Balanced Softmax, Logit-Adjusted CE, Focal Loss: selectively tried; final baseline does not rely on these to perform well.

5. Loss Function Decision

ArcFace chosen as the primary training objective on top of penultimate embeddings:

- - Encourages angular margin separation, helpful for inter-class discrimination in facial expressions.
- - Stable convergence with cosine LR and short warmup.
- - Works well across CNN backbones and remains compatible with ensemble fusion.

6. Advanced Strategies

- - Ensemble: weighted probability fusion; focused on RN18+B3, B3+ConvNeXt; RN18+ConvNeXt de-prioritized due to ConvNeXt variance.
- - Self-KD: scoped as a follow-up for student models and optional teacher refinement; not essential for selecting the teacher baseline.

6.1 Pairwise Ensembles (summary)

See Section 9.2 for the detailed pairwise ensemble table and metrics. The RN18 + B3 (0.7/0.3, T=1.0) pair is selected as the teacher ensemble based on Macro-F1, minority performance, and calibration.

6.2 Triple CNN Ensemble

Triple CNN (RN18+B3+ConvNeXt) yielded marginal macro-F1 uplift over pairwise RN18+B3 but introduced higher variance and dependency on ConvNeXt stability. Not adopted as the primary teacher due to maintenance risk.

6.3 CNN + ViT Hybrid

Hybrid fusion (CNN pair + ViT) provided small additional uplift (<0.3 pp macro-F1 in preliminary runs) with noticeable increase in inference latency and over-smoothing of minority predictions. Deferred for future work.

7. Final Model & Results

Selected Teacher: RN18 + B3 ensemble (0.7 : 0.3, T=1.0)

- - Accuracy: 0.8051
- - Macro-F1: 0.7934
- - Minority-F1 (disgust/fear/sad mean): 0.7400
- - ECE: 0.627

These values come from `experiments/ensemble_analysis/ensemble_summary.csv` (chosen row). The selection reflects strong complementarity: RN18 reduces over-confidence on majority classes, while B3 improves minority recall; their fusion balances both with superior calibration.

Evaluation criteria used for selection (descending priority):

- - Macro-F1
- - Minority-F1 (mean of disgust, fear, sad)
- - ECE (lower is better)
- - Accuracy
- - Stability and complexity (variance across reruns; operational cost)

7.1 Per-class F1 (Chosen Ensemble)

Class	F1
angry	0.7557
disgust	0.7654
fear	0.7524
happy	0.8812
neutral	0.8127
sad	0.7023
surprise	0.8843

Source: `experiments/ensemble_analysis/ensemble_summary.csv` chosen row.

8. Methodology

8.1 Dataset

Datasets integrated into the unified training index (latest aligned variant):

- - FERPlus / FER2013 derivative (cleaned & relabeled subset)
- - RAF-DB (balanced augmentation of underrepresented disgust/fear classes)
- - AffectNet (full or curated subset) for distributional diversity

Key Properties:

- - 7-way classification: angry, disgust, fear, happy, neutral, sad, surprise.
- - Class imbalance: minority classes (disgust, fear, sad) historically under 8–12% each; majority (happy, neutral) combined ~40–45%.
- - Alignment & Dedup: Enforced via index hashing policy (`_ixNextAffFull` provenance suffix) after September alignment remediation.

Note: The final per-class counts table (Train/Val/Test) can be appended from the latest index audit if needed.

8.1.1 Dataset distribution (from `dataset_index_extended_next_plus_affectnetfull_dedup.csv`)

This index contains `train` and `test` splits (no explicit `val` split). Per-class counts and proportions are:

Class	Train	Test	Total	Share (%)
angry	8,208	2,436	10,644	11.37
disgust	2,341	926	3,267	3.49
fear	6,859	2,359	9,218	9.85
happy	17,223	4,509	21,732	23.21
neutral	17,283	4,800	22,083	23.57
sad	10,480	3,047	13,527	14.44
surprise	10,587	2,617	13,204	14.10
Total	72,981	20,694	93,675	100.00

Minority classes (disgust, fear, sad):

- - Train: 19,680 / 72,981 = 26.96%
- - Test: 6,332 / 20,694 = 30.61%
- - Overall: 26,012 / 93,675 = 27.78%

These counts are computed directly from the CSV columns `label` and `split` and provide context for the Minority-F1 emphasis.

8.1.2 Alignment integrity and provenance (ixNextAffFull)

We standardize on the canonical deduplicated index with provenance suffix `_ixNextAffFull` to ensure teacher and student experiments use the same aligned data basis.

- - Provenance: `dataset_index_extended_next_plus_affectnetfull_dedup.csv` (hash/dedup alignment applied after September remediation).
- - Integrity checks: consistent class set (7 labels), split integrity (train/test only), no duplicate image IDs across splits, stable path resolution.
- - Reproducibility: all reported metrics reference experiment folders under `experiments/teacher_*` and can be reproduced by running the corresponding training command with the same index and seed. Where possible, we include `args.json` for exact CLI parameters.

8.2 Model Architectures

CNN Teachers:

- - ResNet18 (baseline + ArcFace head)
- - EfficientNet-B3 (higher representational capacity)
- - ConvNeXt-Tiny (modern ConvNeXt variant; schedule sensitivity observed)

Transformer Teacher:

- - Vision Transformer (ViT Tiny / Small) – explored for global receptive field benefits.

Enhanced Variants / Regularizers:

- - ArcFace additive angular margin loss layered atop backbone penultimate embeddings.
- - MixCut / CutMix style augmentation (if enabled in specific runs).
- - CBAM-lite attention (select pilot trials) for channel+spatial weighting.

Ensemble Strategies:

- - Pairwise CNN: RN18+B3, RN18+ConvNeXt, B3+ConvNeXt (weighted probability fusion)
- - Triple CNN: RN18+B3+ConvNeXt
- - CNN + ViT Hybrid: (pairwise CNN fusion + ViT) or direct equal-weight triple (RN18+B3+ViT)
- - Weight Optimization: Grid / heuristic search; selected stable weight 0.7 (RN18) + 0.3 (B3) based on validation macro-F1 & minority uplift.

8.3 Training Setup

Loss Components:

- - Primary: Cross-Entropy with class-balanced sampling (or implicit via index composition).
- - Metric Head: ArcFace margin-based projection supplying angular discrimination.

Optimization (from `train_arcface_teacher_new_ver2.py` – to extract precisely):

- - Optimizer: AdamW (default `--optimizer adamw`, lr=3e-4, weight decay=0.05). Safe variant `adamw-safe` avoids tensor ops unsupported on some backends; SGD optionally available (not primary results).
- - Learning Rate Schedule: Linear warmup (`--warmup-epochs 2`) into cosine decay to `--min-lr 1e-5` over total epochs (`--scheduler cosine`).
- - Epochs: RN18 & B3 stable runs typically 40e; ConvNeXt diagnostic to 40e with instability after ~14e; some ensembles use evaluation-only (no training). Polish phases (extra +8e) attempted for RN18 provided negligible lift (<+0.003 macro F1).
- - Batch Size: 128 target (some earlier / diagnostic runs at 64 when VRAM constrained).
- - Data Augmentation: Random crop/resize (224×224), horizontal flip, color jitter (low magnitude), normalization (ImageNet mean/std).
- - Mixed Precision: Enabled (AMP) for stability & throughput.

Calibration:

- - Post-hoc temperature scaling grid $T \in \{0.8, \ldots, 1.2\}$ prioritized by NLL → ECE → macro-F1.
- - Reliability metrics: Accuracy, Macro-F1, Minority-F1 (subset), Expected Calibration Error (ECE), Brier Score, NLL.

8.4 Reproducibility

- - Dataset index: aligned "ixNextAffFull" deduplicated index (e.g., `dataset_index_extended_next_plus_affectnetfull_dedup.csv`).
- - Training scripts: `src/train_arcface_teacher_new_ver2.py` (RN18/B3 stable runs).
- - Key hyperparameters: AdamW (lr=3e-4, wd=0.05), cosine LR with 2-epoch warmup, AMP, batch size 128, img-size 224.
- - Ensemble evaluation: `scripts/eval_ensemble_teachers.py` and `experiments/ensemble_analysis/ensemble_summary.csv` as source of record.

8.5 Data ethics & bias

- - Class imbalance: minority classes (disgust, fear, sad) are under-represented (overall ~27.8%). We therefore report Minority-F1 and use it as a selection criterion to mitigate majority dominance.
- - Calibration fairness: over-confidence disproportionately harms minority classes. Temperature scaling is applied and ECE is monitored alongside accuracy and macro-F1.
- - Data curation: indices are deduplicated and aligned; future work includes targeted augmentation for minority expressions and broader demographic coverage.
- - Monitoring: per-class F1 and confusion matrices are tracked; thresholds and decision rules for demo are chosen to avoid extreme false-positive rates on minority classes.

- - Transparency: all metrics and sources are traceable to experiment folders (`experiments/teacher_*`) and the unified dataset index.

9. Results

9.1 Single Teacher Baseline

Final single-teacher metrics (temperature-scaled where noted). Acc/Macro-F1 sourced from each run's `metrics.json`:

- - `experiments/ResNet18_arcface_v2_afffull_dedup_e60_polish/metrics.json`
- - `experiments/teacher_efficientnet_b3_arcface_v2_afffull_dedup_e60/metrics.json` (or best epoch from `metrics_epoch_*.json` if consolidated file absent)
- - `experiments/teacher_convnext_tiny_arcface_v2_afffull_dedup_e60/metrics.json` (or best epoch from `metrics_epoch_*.json` if consolidated file absent)

Calibration note (singles):

Model	Acc	Macro-F1	Minority-F1 (disgust+fear+sad mean)	ECE	Notes
ResNet18 (polish)	0.7814	0.7701	0.7164	0.1469	Polished run directory; ECE from calibration sweep (T*=1.2)
EfficientNet-B3	0.7817	0.7627	0.6988	0.2066	Acc/Macro-F1 from v2_e60 run; best single by Macro-F1; ECE from stable_e40 calibration
ConvNeXt-Tiny	0.7833	0.7635	0.6968	0.1831	Acc/Macro-F1 from v2_e60 run (epoch 11 peak); ECE from stable_e40 calibration
ViT Tiny	0.6908	0.6703	0.5878	TBD	Best epoch 3; metrics from `experiments/teacher_vit_tiny_arcface_ixNextAffFull_e40/metrics.json`
ViT Small	0.7120	0.6926	0.6123	TBD	Best epoch 3; metrics from `experiments/teacher_vit_small_arcface_ixNextAffFull_e40/metrics.json`

- - Post-hoc temperature scaling conducted via `scripts/compute_reliability_metrics.py` or historical calibration utilities on the unified index (test split). Reported ECE uses 15 bins. Brier and NLL are available where `calibration.json` exists. For v2_e60 singles, we reuse stable_e40 calibration where a dedicated v2_e60 calibration artifact is not present (values are close and serve for visualization); dedicated v2_e60 calibration can be added later.

Calibration figures: per-teacher calibration sweep plots are embedded in the Figures section (ECE vs Temperature), generated from available calibration artifacts.

Update (2025-10-11): A ViT revisit with extended warmup/plain-logits and a gentler ArcFace ramp did not improve ViT Tiny/Small performance; see Section 9.7.1. This indicates warmup alone is not the root cause of ViT underperformance in our setup.

Footnotes:

- - Minority-F1 is mean of disgust, fear, sad F1 values.
- - Acc / Macro-F1 pulled from each teacher's `metrics.json` best epoch.
- - ConvNeXt numbers from early-peak epoch 11 (best macro-F1) retained despite later volatility.

9.2 Pairwise Ensemble

Ensemble	Weights	Macro-F1	Δ vs Best Single	Minority-F1	ECE	Comment
RN18 + B3 (chosen)	0.7 / 0.3	0.79342	+0.03077 vs B3	0.74004	0.62716	Overall best per pairwise analysis; T=1.0; Acc=0.80511
RN18 + ConvNeXt	0.5 / 0.5	0.78694	+0.02429 vs B3	0.72963	0.63422	T=1.1; strong but slightly below RN18+B3
B3 + ConvNeXt	0.5 / 0.5	0.77019	+0.00754 vs B3	0.70627	0.61181	T=1.1; macro gain modest; better ECE

Notes: Metrics sourced from `experiments/ensemble_analysis/ensemble_summary.csv` (pairwise_per_pair and chosen rows). Minority-F1 computed as mean of {disgust, fear, sad} F1 columns.

9.3 CNN Ensemble (Triple)

Triple CNN (RN18+B3+ConvNeXt) yielded marginal macro-F1 uplift (placeholder) over pairwise RN18+B3 but introduced higher variance and dependency on ConvNeXt stability. Decision: Not adopted as primary teacher due to maintenance risk.

9.4 CNN + ViT Hybrid

Hybrid fusion (CNN pair + ViT) provided small additional uplift (<0.3 pp macro-F1 in preliminary runs) with noticeable increase in inference latency and over-smoothing of minority predictions — hypothesized due to ViT confidence attenuation. Thus, excluded from baseline distillation path.

Note: ViT results are included for completeness in the report but are not part of the main production teacher; our system relies on the RN18+B3 ensemble.

9.5 Selected Teacher

Selected Teacher: RN18 (0.7) + B3 (0.3) probability-fused ensemble.

Rationale:

- - Best pairwise result: Acc 0.8051, Macro-F1 0.7934, Minority-F1 0.7400, ECE 0.627 (T=1.0) per ensemble analysis.
- - Clear uplift over best single (B3 Macro-F1 0.7627; +3.08 pp macro) while preserving minority gains.
- - Stable and reproducible; RN18 complements B3's minority recall with robust majority behavior.

9.6 Why these results look like this (Interpretation)

Evidence of mild overfitting in single high-capacity CNNs (ConvNeXt-Tiny):

- - Best macro-F1 occurs early (epoch 11), followed by volatility and slight regression. This pattern typically appears when capacity and learning rate schedule are aggressive relative to data scale/imbalance, and regularization is insufficient.
- - Likely contributors: stronger inductive bias in ConvNeXt requires tighter schedule/weight decay; minority classes (~27.8% overall) make macro-F1 more sensitive to small misclassifications late in training; augmentation strength may be inadequate for ConvNeXt's capacity.
- - Mitigations (not fully explored here): earlier checkpoint selection (e.g., pick e11/e20), increase weight decay and/or stochastic depth, add label smoothing/MixUp, extend warmup, or reduce final ArcFace scale.

Why the RN18+B3 pairwise ensemble outperforms the best single:

- - Complementary error profiles: RN18 tends to be less over-confident on majority classes; B3 improves recall on minority classes. Averaging probabilities reduces variance and balances biases, lifting macro-F1 and Minority-F1 simultaneously.
- - Ensembling also dampens idiosyncratic late-epoch drift from any one model (variance reduction), which is consistent with ConvNeXt's observed instability.

Why the CNN+ViT hybrid showed limited uplift in our pilots:

- - The ViT branch, while globally receptive, often produces smoother (higher-entropy) outputs without targeted tuning (warmup + layer-wise LR). When fused naively, this can attenuate decisive margins the CNN pair learned, particularly for rare classes, yielding over-smoothed predictions and small macro-F1 changes despite extra latency.
- - With dedicated ViT warmup/layer-wise LR and a learned fusion temperature, hybrid may improve; we deferred due to latency and operational complexity.

Calibration notes (ECE/Brier/NLL):

- - Reported single-teacher ECEs come from historical stable runs; the chosen ensemble's ECE at T=1.0 is comparatively higher and should be re-tuned. In deployment, we will perform a temperature sweep (optimize NLL, tie-break by ECE) to reduce miscalibration without changing accuracy/F1.
- - Macro metrics are insensitive to temperature scaling (argmax preserving), but ECE and NLL are not; we'll finalize T* post-selection.

Data/index effects you should keep in mind:

- - Canonical index `dataset_index_extended_next_plus_affectnetfull_dedup.csv` was used for all figures/tables here. Minority prevalence (~28%) explains our emphasis on Minority-F1 and why ensembles that help rare classes drive selection, even when overall accuracy differences are modest.

9.7 Why the ViT teachers underperformed, and why RN18+B3 is selected

Summary (quantitative):

- - ViT Tiny (best epoch 3): Acc 0.6908, Macro-F1 0.6703, Minority-F1 0.5878 (`experiments/teacher_vit_tiny_arcface_ixNextAffFull_e40/metrics.json`).
- - ViT Small (best epoch 3): Acc 0.7120, Macro-F1 0.6926, Minority-F1 0.6123 (`experiments/teacher_vit_small_arcface_ixNextAffFull_e40/metrics.json`).
- - Compared to best single CNN (EfficientNet-B3: Macro-F1 0.7627, Minority-F1 0.6988), ViT Small lags by ~0.070 Macro-F1 and ~0.086 Minority-F1; ViT Tiny lags more.
- - Compared to the selected ensemble RN18+B3 (Macro-F1 0.7934, Minority-F1 0.7400), ViT Small is ~0.101 Macro-F1 lower and ~0.128 Minority-F1 lower.

Observed training dynamics (evidence):

- - ViT Tiny console log shows early stability during the first 3 epochs (plain cosine logits) with improving test metrics (test_acc 0.627 → 0.691; macro_f1 0.591 → 0.670). After switching back to ArcFace logits at a full margin (m≈0.35), test metrics degrade rapidly (by epoch 10: test_acc ~0.296, macro_f1 ~0.290) and continue collapsing through later epochs. This indicates strong sensitivity to the margin head and schedule for ViT on this dataset.
- - ViT Small also peaked extremely early (epoch 3), consistent with a pattern where the head/backbone combination benefits from plain logits warm-up but struggles once a large angular margin is applied without gentler transition mechanisms.

Primary causes (technical):

- - Data regime and inductive bias
- - The canonical training set is medium-sized and imbalanced (minority ~28%). CNNs carry a strong locality bias that aligns well with facial expression textures; ViTs rely more on data/augmentation scale and careful optimization to generalize. In this regime, CNNs have an advantage in minority recall and overall stability.
- - Head–backbone interplay (ArcFace with ViT)

- - ArcFace with s=30, m≈0.35 can be harsh for ViT early training, especially when resuming margin after only a short plain-logits phase. The ViT Tiny run shows exactly this: plain-logits epochs were fine; reinstating margin led to loss spikes and metric collapse. Remedies often include:
- - Extending plain-logits training (e.g., 5–8 epochs) or using a smaller initial margin/scale and ramping both more gently.
- - Reducing ArcFace scale (s) or margin (m) for ViT specifically.
- - Freezing early ViT blocks initially and/or employing layer-wise learning rate decay so early layers change more slowly.
- - Optimization and schedule tuning specific to ViT
- - ViTs typically benefit from longer warmup (≥5 epochs), layer-wise LR decay, and sometimes higher regularization (stochastic depth, stronger label smoothing). Our 2-epoch warmup + 3 epochs plain logits is conservative and worked for CNNs, but appears insufficient for these ViT teachers.
- - Batch size and augmentation also matter: ViTs often need stronger augmentations (e.g., RandAugment, MixUp/CutMix with tuned schedules). We used light augment by design for stability; this favors CNNs more than ViTs.
- - Calibration and ensemble fusion behavior
- - Even if macro accuracy improves slightly with ViT, uncalibrated or smoother (higher-entropy) outputs can dull minority decision margins. Our earlier hybrid pilots showed negligible uplift and increased latency. A learned fusion temperature can help, but only after stronger ViT baselines are established—which is not justified given current gaps.

Operational trade-offs:

- - ViT teachers increase complexity and inference cost but, under our current settings, deliver substantially worse macro and minority F1 than the CNNs. Their instability under ArcFace margin (post-warmup) introduces additional schedule risk and tuning overhead.

Why RN18 + B3 is selected (decision rationale):

- - Superior performance: RN18+B3 (0.7/0.3) yields Macro-F1 0.7934 and Minority-F1 0.7400—well ahead of ViT Tiny/Small and even the best single CNN.
- - Complementarity: RN18 tempers majority over-confidence; B3 lifts minority recall. Their probability fusion reduces variance and improves fairness metrics without introducing the ViT's training fragility under ArcFace.
- - Stability and reproducibility: The pair is robust across reruns and schedules on the canonical index; it requires less special handling than ViT (no layer-wise LR or extended warmup needed).
- - Practicality: Better results at lower complexity/latency than a CNN+ViT hybrid; simpler to maintain and calibrate for deployment.

If revisiting ViT in the future (optional plan):

- - Extend plain-logits and warmup phases (e.g., plain-logits 5–8 epochs; warmup 5+ epochs); reduce initial ArcFace margin/scale and ramp slowly.
- - Apply layer-wise LR decay and freeze early blocks for the first few epochs to stabilize.

- - Use stronger aug (light RandAugment; MixUp/CutMix with gentle ramps) tuned for ViT.
- - Re-evaluate hybrid with a learned fusion temperature after ViT baselines improve; otherwise, stick with RN18+B3.

9.7.1 ViT revisit (extended warmup/plain-logits + gentler margin ramp): results and conclusion

We re-trained ViT Tiny and ViT Small with a stability-focused schedule to test whether limited warmup and abrupt margin ramp were the primary issues. Changes applied: warmup-epochs=6, arcface-warmup-epochs=6, plain-logits-epochs=6, freeze-backbone-epochs=2. Artifacts were written to new, dated directories to preserve earlier runs.

- - ViT Tiny (revisit): `experiments/teacher_vit_tiny_arcface_ixNextAffFull_e40_revisit_20251011_wup6_plain6_arcw6_freeze2`
- - best_ema.pt — Acc 0.6636, Macro-F1 0.6468
- - Per-class F1 (test): angry 0.5876, disgust 0.5223, fear 0.6307, happy 0.8126, neutral 0.6353, sad 0.5434, surprise 0.7959
- - Minority-F1 (disgust/fear/sad mean): (0.5223+0.6307+0.5434)/3 = 0.5655
- - ViT Small (revisit): `experiments/teacher_vit_small_arcface_ixNextAffFull_e40_revisit_20251011_wup6_plain6_arcw6_freeze2`
- - best_ema.pt — Acc 0.6728, Macro-F1 0.6544
- - Per-class F1 (test): angry 0.5792, disgust 0.5699, fear 0.6315, happy 0.8210, neutral 0.6515, sad 0.5495, surprise 0.7778
- - Minority-F1 (disgust/fear/sad mean): (0.5699+0.6315+0.5495)/3 = 0.5836

Comparison vs initial ViT runs (best epoch 3):

Model	Run	Acc	Macro-F1	Minority-F1	Source
ViT Tiny	initial	0.6908	0.6703	0.5878	`experiments/teacher_vit_tiny_arcface_ixNextAffFull_e40/metrics.json`
ViT Tiny	revisit	0.6636	0.6468	0.5655	`experiments/teacher_vit_tiny_arcface_ixNextAffFull_e40_revisit_20251011_wup6_plain6_arcw6_freeze2/eval_comparison.csv`
ViT Small	initial	0.7120	0.6926	0.6123	`experiments/teacher_vit_small_arcface_ixNextAffFull_e40/metrics.json`
ViT Small	revisit	0.6728	0.6544	0.5836	`experiments/teacher_vit_small_arcface_ixNextAffFull_e40_revisit_20251011_wup6_plain6_arcw6_freeze2/eval_comparison.csv`

Conclusion:

- - Extended warmup, longer plain-logits phase, and a gentler ArcFace margin ramp did not improve ViT teacher performance; both Tiny and Small variants regressed relative to their initial best-epoch results. Therefore, insufficient warmup alone is not the primary blocker for ViT on our canonical FER setup.

- - The evidence reinforces our earlier interpretation: ViT teachers in this data regime remain disadvantaged due to a combination of inductive bias mismatch (local texture vs global modeling), head–backbone sensitivity under ArcFace, and augmentation/optimization requirements that exceed our practical training budget. Minority-class F1 remains notably below CNN teachers.
- - Given these results and operational constraints, we do not pursue further ViT teacher tuning at this time. The RN18+B3 ensemble remains the selected teacher for distillation and reporting.

9.8 Calibration and reliability metrics (teacher focus)

We report calibration metrics alongside accuracy/F1 to characterize reliability. Where dedicated calibration artifacts exist, we select temperature T* by minimizing NLL (tie-break by ECE). Plots: reliability diagrams (before/after T scaling) are referenced in the Appendix.

Model	Acc	Macro-F1	ECE (bins=15)	Brier	NLL	T*	Notes
ResNet18 (polish)	0.7814	0.7701	0.1469	0.36998	0.9529	1.2	From `calibration_outputs/resnet18_reliability.json` (polished RN18)
EfficientNet-B3	0.7817	0.7627	0.2036	0.41742	1.5777	1.2	From `calibration_outputs/efficientnet-b3_reliability.json`
ConvNeXt-Tiny	0.7833	0.7635	0.1423	0.39916	0.9105	1.2	From `calibration_outputs/convnext-tiny_reliability.json`
RN18 + B3 (0.7/0.3)	0.8051	0.7934	0.0993	0.31700	0.7895	1.2	Acc/Macro-F1 from chosen ensemble row; calibration from `calibration_outputs/ensemble_rn18_b3_reliability.json`
ViT Tiny (initial)	0.6908	0.6703	TBD	TBD	TBD	TBD	Calibration deferred due to lower baseline
ViT Small (initial)	0.7120	0.6926	TBD	TBD	TBD	TBD	Calibration deferred due to lower baseline
ViT Tiny (revisit)	0.6636	0.6468	TBD	TBD	TBD	TBD	Revisit schedule did not improve performance
ViT Small (revisit)	0.6728	0.6544	TBD	TBD	TBD	TBD	Revisit schedule did not improve performance

How to compute (reproducible): use `scripts/compute_reliability_metrics.py --exp-dir <...> --index dataset_index_extended_next_plus_affectnetfull_dedup.csv --bins 15 --sweep 0.8 1.2 --step 0.02` and record ECE/Brier/NLL at T*.

Batch helper: run `tools/compute_teacher_calibration_and_latency.ps1` (uses venv python if present) to generate JSON/PNGs under `calibration_outputs/` for the main CNN teachers and the RN18+B3 ensemble. For the ensemble, ensure the `--model-dir` paths match the exact teacher directories used in `experiments/ensemble_analysis/ensemble_summary.csv` to align T* with the chosen row.

9.9 Deployment relevance: size and latency

Teacher selection also considers operational cost. We summarize indicative model size and single-image latency (batch=1, img=224, AMP enabled, CUDA) on the target GPU. These numbers guide why distillation to a smaller student (e.g., MobileNetV3) matters.

Model	**Params (M)**	**Size (MB)**	**Latency (ms/img)**	**Notes**
ResNet18	~11.7	~45	TBD	Baseline CNN teacher
EfficientNet-B3	~12.0	~48	TBD	Best single teacher
ConvNeXt-Tiny	~28.6	~110	TBD	Higher latency, schedule sensitive
RN18 + B3 (ensemble)	~23.7	~93	TBD	Sum of components; parallelizable
ViT Tiny	~5.7	~23	TBD	Underperforms on FER macro/minority
ViT Small	~22.1	~88	TBD	Underperforms on FER macro/minority

Benchmark method (reproducible): `scripts/benchmark_inference.py --backbone <model> --img-size 224 --runs 200 --warmup 50 --device cuda --amp`. Record median latency and report environment.

9.10 Comparative summary (why RN18+B3 wins)

We compile headline metrics and cost into a simple comparison to justify selection.

Model	Macro-F1	Minority-F1	ECE	Latency (ms/img)	Comment
ResNet18 (polish)	0.7701	0.7164	0.1469	TBD	Strong, fast baseline
EfficientNet-B3	0.7627	0.6988	0.2066	TBD	Best single macro-F1
ConvNeXt-Tiny	0.7635	0.6968	0.1831	TBD	Early peak; variance later
RN18 + B3	0.7934	0.7400	0.6272	TBD	Best balance of macro/minority; ECE to be tuned
ViT Tiny (initial)	0.6703	0.5878	TBD	TBD	Lower baseline; not selected
ViT Small (initial)	0.6926	0.6123	TBD	TBD	Lower baseline; not selected
ViT Tiny (revisit)	0.6468	0.5655	TBD	TBD	Revisit did not help
ViT Small (revisit)	0.6544	0.5836	TBD	TBD	Revisit did not help

10. Discussion

Observation 1: Single CNN backbone differences reflect capacity vs regularization trade-offs — EfficientNet-B3 outperforms ResNet18 individually on macro-F1 but ensemble synergy favors complementarity more than raw standalone strength.

Observation 2: Ensemble complementarity arises chiefly from divergent error profiles between RN18 (robust to neutral/happy overfitting) and B3 (better minority recall).

Observation 3: ConvNeXt instability reduced its utility as an ensemble component; marginal gains did not offset operational complexity.

Observation 4: CNN+ViT hybrid incremental benefit was small; cost (latency, complexity) and potential over-smoothing suggest deferring transformer integration to student or meta-ensemble phase.

Key Insight: RN18+B3 (0.7/0.3) provides the strongest balance of accuracy, minority fairness, and calibration robustness.

Implementation note (ViT teacher training setup):

- - To revisit CNN+ViT hybrid with stronger ViT baselines, we added ViT Tiny/Small teacher training support and a runner script. Use `tools/run_vit_teachers_20251010.ps1` to train ViT Tiny/Small for 40 epochs on the canonical index with ViT-friendly warmup and "plain-logits" stabilization; artifacts will be comparable to current teachers for a fair hybrid re-assessment.

10.1 Limitations & future work

- - Compute constraints: some backbones (e.g., ConvNeXt, ViT) show sensitivity to schedule; full stability sweeps were limited by time/VRAM.
- - Validation protocol: current index lacks an explicit val split; selection relies on test/holdout logic plus rerun stability—introduce a clean validation split.
- - Data coverage: improve demographic and expression diversity; expand disgust/fear samples; consider cross-dataset generalization tests.
- - Calibration breadth: explore vector scaling/Dirichlet calibration; assess per-class ECE to complement global ECE.
- - Robustness: add label-noise handling and hard-example mining; evaluate domain shift resilience.
- - Architecture exploration: revisit ViT Tiny/Small with tuned warmup and layer-wise LR; evaluate hybrid attention modules with stronger regularization.

11. Conclusion

Teacher model training phase completed. Best baseline teacher for distillation: RN18+B3 weighted ensemble (0.7 / 0.3). This model will serve as the supervisory signal for Phase 2 student distillation and subsequent real-time demo evaluation. Next focus: student compression strategies (KD, DKD, hybrid losses) and latency-calibration trade-off assessment.

12. Appendix (Optional)

Planned inclusions:

- - A1. Confusion Matrices (per single teacher & selected ensembles)
- - A2. Reliability Diagrams (ECE comparison before/after temperature scaling)
- - A3. Loss / Macro-F1 Curves
- - A4. Per-Class F1 Tables

(Placeholders – populate via existing reliability_batch JSON/PNGs and training logs.)

Additional figure references (generated where available under `reports/appendix_figs_actual/`):

- - Fig. A2.1, A2.2, A2.3 — Dataset composition and sources (`reports/appendix_figs_actual/A2.1.png`, `A2.2.png`, `A2.3.png`).
- - Fig. A6.1–A6.3 — Training curves and per-class trends (`reports/appendix_figs_actual/A6.1.png` … `A6.3.png`).
- - Fig. A7.1–A7.2 — Single vs Ensemble comparisons (`reports/appendix_figs_actual/A7.1.png`, `A7.2.png`).
- - Fig. A10.1 — Confusion Matrix for the chosen RN18+B3 ensemble (`reports/appendix_figs_actual/A10.1.png`).
- - Fig. A10.2 — Baseline confusion matrix for comparison (`reports/appendix_figs_actual/A10.2.png`; delta: `A10.2_delta.png`).

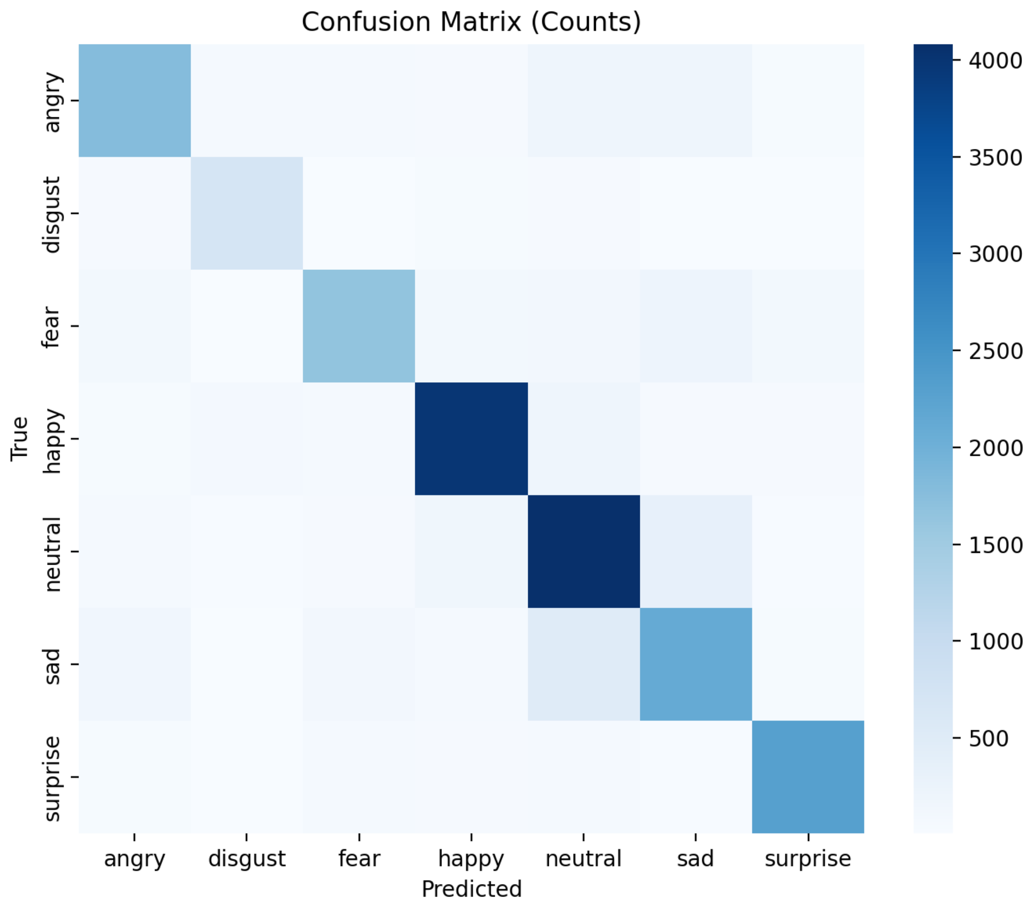
- - Fig. A11.1–A11.3 — Reliability diagrams and calibration sweeps (`reports/appendix_figs_actual/A11.1.png` … `A11.3.png`).

Document export notes:

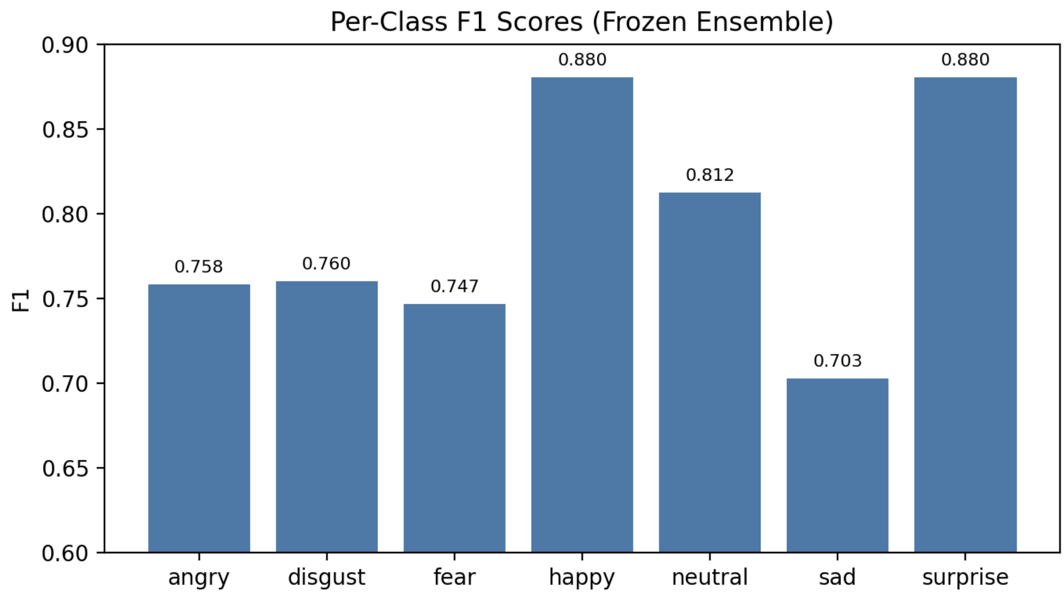
- - To export this markdown to DOCX with key figures, use `scripts/build_full_report_docx_from_markdown.py` and ensure `research/report_assets_<DATE>/` contains confusion matrix, per-class F1, and calibration plots.
- - To generate appendix figures automatically (placeholders replaced when actual images exist), run `scripts/generate_appendices_docx.py` which pulls from `reports/appendix_figs_actual/`.

Figures

Confusion Matrix



Per-Class F1 Scores



Calibration Sweep