1) Teacher Model Training Report（教師模型訓練報告）

Date: 2025-12-24

Training objective（Stage A / Stage B）

- **Stage A (primary)**: Train high-capacity “teacher” classifiers on cleaned multi-source FER data using an ArcFace-style training protocol, then export checkpoints + calibration artifacts.
- **Stage B (optional / paused)**: Stage-A → Stage-B finetuning at higher resolution / different source filtering for downstream use. In this reconstruction, Stage B is **not required** for student KD/DKD because the student consumes exported softlabels (`softlabels.npz`), not teacher ONNX.

Model architectures trained

Stage A (img224):

- RN18: `resnet18`
- B3: `tf_efficientnet_b3`
- CNXT: `convnext_tiny`

Artifacts (per run):

- `outputs/teachers/<RUN>/best.pt`, `checkpoint_last.pt`, `last.onnx`, `best.onnx`
- `history.json`, `reliabilitymetrics.json`, `calibration.json`, `alignmentreport.json`, `environment.json`

Training settings (from checkpoint args)

Common settings across RN18/B3/CNXT Stage A (same hyperparameters; different backbone):

- Image size: **224**
- Optim/LR:
  - LR: **3e-4**
  - Min LR: **1e-5**
  - Weight decay: **0.05**
  - LR warmup: **2 epochs**, then cosine-style schedule (as implemented in `scripts/train_teacher.py`)
- Epochs: **60**
- Batch size: **64** (accum_steps=1)
- DataLoader workers: **8**
- Augmentation / preprocessing:
  - **CLAHE enabled** (`--clahe`)
  - Standard resize+crop+normalize pipeline (see training code)
- ArcFace protocol:
  - `arcface_m=0.35`, `arcface_s=30.0`
  - Plain-logits warmup: 5 epochs; margin ramp 5 → 15
- Source filter configuration (Stage A): `ferplus, rafdb_basic, affectnet_full_balanced, expw_hq`

Evaluation dataset used (for the metrics below)

From each teacher run’s `alignmentreport.json`:

- Manifest: `Training_data_cleaned/classification_manifest.csv`
- After applying the source filter, the actual rows used are:
  - Total after filter: **225,629**
  - Train rows: **182,960**

  - Val rows (used for the recorded evaluation metrics): **18,165**
  - Sources present after filter: **affectnet_full_balanced (71,764), ferplus (138,526), rafdb_basic (15,339)**
- Note: although the include list contains `expw_hq`, these Stage A runs' filtered dataset does not include ExpW rows (see `source_counts_after_filter` in `alignmentreport.json`).

Reference implementation: `scripts/train_teacher.py`

Training curves (loss / accuracy / macro-F1)

- Stored as JSON time series in `history.json` for each teacher run.
- Recommended plots:
  - `epoch` vs `train_loss`
  - `epoch` vs `val.accuracy`, `val.macro_f1`

Final metrics (Stage A teachers)

All metrics below are from each run's `reliabilitymetrics.json`.

Epoch indexing note:

- In teacher `reliabilitymetrics.json`, `epoch` is **59** for the final checkpoint; this corresponds to a 60-epoch run with epoch indices **0–59**.

RN18 (Stage A img224)

- Accuracy: **0.7862** | Macro-F1: **0.7808**
- Per-class F1: Angry 0.7357, Disgust 0.6940, Fear 0.7635, Happy 0.8970, Sad 0.7415, Surprise 0.8186,

Neutral 0.8155

- Calibration:
   - Raw: NLL **4.0259**, ECE **0.2053**
   - Temp-scaled (global T=5.0): NLL **0.8803**, ECE **0.1489**

Run dir: `outputs/teachers/RN18_resnet18_seed1337_stageA_img224/`

B3 (Stage A img224)

- Accuracy: **0.7961** | Macro-F1: **0.7910**
- Per-class F1: Angry 0.7521, Disgust 0.7156, Fear 0.7576, Happy 0.9197, Sad 0.7479, Surprise 0.8042, Neutral 0.8399
- Calibration:
   - Raw: NLL **3.2219**, ECE **0.1988**
   - Temp-scaled (global T=5.0): NLL **0.7871**, ECE **0.0839**

Run dir: `outputs/teachers/B3_tf_efficientnet_b3_seed1337_pretrained_true_v1_stageA_img224/`

CNXT (Stage A img224)

- Accuracy: **0.7941** | Macro-F1: **0.7890**
- Per-class F1: Angry 0.7687, Disgust 0.7194, Fear 0.7395, Happy 0.9135, Sad 0.7313, Surprise 0.8064, Neutral 0.8439
- Calibration:
   - Raw: NLL **3.1014**, ECE **0.2009**
   - Temp-scaled (global T=5.0): NLL **0.7700**, ECE **0.0817**

Run dir: `outputs/teachers/CNXT_convnext_tiny_seed1337_stageA_img224/`

Calibration metrics note

- Teachers currently record **NLL** and **ECE** (and temperature-scaled versions).
- **Brier score** is not present in teacher `reliabilitymetrics.json` at this moment (TBD if needed).

Observations (based on recorded metrics)

- **Most stable overall (acc/macro-F1):** B3 slightly leads (macro-F1 0.7910).
- **Overconfidence (raw calibration):** all three teachers show high raw NLL and ECE; temperature scaling (often T=5.0) significantly improves NLL/ECE.
- **Minority / hard classes:**
  - Disgust: CNXT best (F1 ~0.7194)
  - Fear: RN18 best (F1 ~0.7635)
  - Angry: CNXT best (F1 ~0.7687)

Next steps

- If Stage B is needed for deployment (higher-res or domain finetune), run `scripts/train_teacher.py --init-from <StageA best.pt> ...`.
- Otherwise keep Stage A teachers and focus on ensemble + student KD/DKD (softlabels-based).