3) Student Model Training Report（學生模型訓練報告）

Date: 2025-12-24

Student architecture

- Student backbone: **MobileNetV3-Large** (`mobilenetv3_large_100` via timm)
- Input: img224
- Preprocessing: **CLAHE enabled** (to match teacher/ensemble pipeline)

Primary run outputs:

- CE: `outputs/students/mobilenetv3_large_100_img224_seed1337_CE_20251223_225031/`
- KD: `outputs/students/mobilenetv3_large_100_img224_seed1337_KD_20251223_225031/`
- DKD: `outputs/students/mobilenetv3_large_100_img224_seed1337_DKD_20251223_225031/`
- Logs: `outputs/students/_logs_20251223_225031/`

KD settings

Teacher supervision source:

- Softlabels directory:
  - `outputs/softlabels/_ens_hq_train_rn18_0p4_b3_0p4_cnxt_0p2_logit_clahe_20251223_152856/`
- Files:
  - `softlabels.npz`
  - `softlabels_index.jsonl`

KD hyperparameters (from run command/log):

- Temperature **T=2**
- KD weight **α=0.5**
- Epochs: **20**

DKD settings

- Started from KD `best.pt` (resume)
- DKD hyperparameters:
  - Temperature **T=2**
  - α=0.5
  - **β=4**
- Intended DKD schedule: **10 additional epochs after KD**
  - Fixed a resume/epochs bug (see Week 4 log and runner patch) so DKD actually trains instead of exiting early.

Training curves

- Stored per-run in `history.json`:
  - `train_loss`, `epoch_sec`, LR, and validation metrics
- Recommended plots:
  - epoch vs train_loss
  - epoch vs val accuracy / val macro-F1

Final metrics (HQ-train manifest, img224, CLAHE+AMP, seed=1337)

From each run’s `reliabilitymetrics.json`.

HQ-train manifest split sizes (from `Training_data_cleaned/classification_manifest_hq_train.csv`):

- Train: **213,144**
- Val: **18,020**
- Test: **27,840**
- Total: **259,004**

Evaluation scope note:

- The CE/KD/DKD metrics reported in this file are evaluated on the HQ-train manifest used by the run (not on `test_all_sources.csv`).

CE baseline

- Accuracy: **0.750174** | Macro-F1: **0.741952**
- Per-class F1:
  - Angry 0.726340
  - Disgust 0.642839
  - Fear 0.764029
  - Happy 0.801425
  - Sad 0.716981
  - Surprise 0.787086
  - Neutral 0.754961
- Calibration:
  - Raw: NLL **1.315335**, ECE **0.131019**

- Temp-scaled (global T=3.228): NLL **0.777757**, ECE **0.049897**

KD

- Accuracy: **0.734688** | Macro-F1: **0.733351**
- Per-class F1:
  - Angry 0.723717
  - Disgust 0.678227
  - Fear 0.744691
  - Happy 0.760978
  - Sad 0.723361
  - Surprise 0.780076
  - Neutral 0.722405
- Calibration:
  - Raw: NLL **2.093148**, ECE **0.215289**
  - Temp-scaled (global T=5.000): NLL **0.768196**, ECE **0.027764**

DKD (resume from KD best)

- Accuracy: **0.737432** | Macro-F1: **0.737511**
- Per-class F1:
  - Angry 0.725522
  - Disgust 0.682833
  - Fear 0.756052
  - Happy 0.759617
  - Sad 0.728567
  - Surprise 0.791491

  - Neutral 0.718493

- Calibration:

  - Raw: NLL **1.511788**, ECE **0.209450**

  - Temp-scaled (global T=3.348): NLL **0.765203**, ECE **0.026605**

Comparison vs teacher (Stage A)

- Teachers (Stage A img224) reach ~**macro-F1 0.781–0.791** on their recorded evaluation.

- Student (CE/KD/DKD) currently reaches **macro-F1 ~0.733–0.742** on HQ-train evaluation.

Observations

- **Accuracy / macro-F1:** CE slightly outperforms KD/DKD in this initial run.

- **Calibration:** KD/DKD show much better temperature-scaled ECE (≈0.027) than CE (≈0.050).

- **Trade-off:** current KD/DKD hyperparameters may prioritize calibration over raw macro-F1.

Next steps

- Tune KD/DKD (`temperature`, `alpha`, `beta`) and/or train longer to seek macro-F1 gains without losing calibration.

- Optionally evaluate student on `test_all_sources.csv` for a more deployment-realistic comparison.