2) Teacher Model Ensemble Report（教師模型集成報告）

Date: 2025-12-24

Ensemble method

- Fusion space: **logit fusion** (weighted sum of logits, then softmax for probabilities)
- Preprocessing: **CLAHE enabled** to match training-time preprocessing
- Weight selection: grid-style trials over (RN18/B3) and (RN18/B3/CNXT) weightings

Core scripts/tools:

- `scripts/export_ensemble_softlabels.py`, `scripts/export_multi_ensemble_softlabels.py`
- Triage summary: `outputs/softlabels/_ensemble_triage.md`

Models used

- Stage A teachers (img224):
  - RN18: `resnet18`
  - B3: `tf_efficientnet_b3`
  - CNXT: `convnext_tiny`

Benchmark results (selected runs)

Below are the recorded `ensemble_metrics.json` outputs (accuracy, macro-F1, per-class F1, NLL, ECE, Brier).

A) Multi-source benchmark: test_all_sources (48,928 images)

This benchmark is used to reduce RAF-only bias and better reflect mixed-domain performance.

- **Best 3-teacher ensemble (selected): RN18/B3/CNXT = 0.4/0.4/0.2**

  - acc=0.687255 | macro-F1=0.659608 | NLL=4.077156 | ECE=0.287694 | Brier=0.590869

  - Per-class F1: Angry 0.6304, Disgust 0.6034, Fear 0.5350, Happy 0.8389, Sad 0.5924, Surprise 0.7205, Neutral 0.6967

  - Artifact: `outputs/softlabels/_archive/bad_list_20251223_121501/_ens_test_all_sources_rn18_0p4_b3_0p4_cnxt_0p2_logit_clahe_20251223_111523/ensemble_metrics.json`

- RN18/B3 = 0.3/0.7

  - acc=0.682186 | macro-F1=0.654132 | NLL=4.254531 | ECE=0.301064 | Brier=0.612558

  - Artifact: `outputs/softlabels/_archive/bad_list_20251223_121501/_ens_test_all_sources_rn18_0p3_b3_0p7_logit_clahe_20251223_091041/ensemble_metrics.json`

- RN18/CNXT = 0.5/0.5

  - acc=0.677383 | macro-F1=0.649794 | NLL=4.513956 | ECE=0.296576 | Brier=0.610196

  - Artifact: `outputs/softlabels/_archive/bad_list_20251223_121501/_ens_test_all_sources_rn18_0p5_cnxt_0p5_logit_clahe_20251223_111122/ensemble_metrics.json`

- RN18/B3 = 0.5/0.5

  - acc=0.680755 | macro-F1=0.652920 | NLL=4.415604 | ECE=0.293571 | Brier=0.604839

  - Artifact: `outputs/softlabels/_ens_test_all_sources_rn18_0p5_b3_0p5_logit_clahe_20251223_121224/ensemble_metrics.json`

B) RAF-DB basic test (3,068 images)

- Example (RN18/B3 = 0.7/0.3)

  - acc=0.834746 | macro-F1=0.748370 | NLL=1.711080 | ECE=0.146590 | Brier=0.300635

  - Artifact:
`outputs/softlabels/_archive/bad_list_20251223_121501/_ens_rafdb_basic_rn18_0p7_b3_0p3_logit_clahe_20251220_145349/ensemble_metrics.json`

C) RAFDB test (ensemble run, logit+CLAHE)

- RN18/B3 = 0.5/0.5

  - acc=0.856258 | macro-F1=0.777514 | NLL=1.444302 | ECE=0.126894 | Brier=0.265329

  - Artifact:
`outputs/softlabels/_ens_rn18_0p5_b3_0p5_rafdb_test_logit_clahe_20251220_155054/ensemble_metrics.json`

D) “Full unified test” (legacy-style “fulltest”)

- RN18/B3 = 0.3/0.7

  - acc=0.682391 | macro-F1=0.653374 | NLL=4.246412 | ECE=0.300790 | Brier=0.611916

  - Artifact:
`outputs/softlabels/_archive/bad_list_20251223_121501/_ens_rn18_0p3_b3_0p7_fulltest_logit_clahe_20251220_161909/ensemble_metrics.json`

Best ensemble selection

Selected ensemble for HQ-train softlabels export:

- **RN18/B3/CNXT = 0.4/0.4/0.2**, logit fusion + CLAHE

- Softlabels export dir (used for student KD/DKD):

  - `outputs/softlabels/_ens_hq_train_rn18_0p4_b3_0p4_cnxt_0p2_logit_clahe_20251223_152856/`

  - Contains: `softlabels.npz`, `softlabels_index.jsonl`, `hash_manifest.json`, `classorder.json`, `alignmentreport.json`

Where the selected softlabels live (confirmed)

Export directory:

- `outputs/softlabels/_ens_hq_train_rn18_0p4_b3_0p4_cnxt_0p2_logit_clahe_20251223_152856/`

Files present in that folder:

- `softlabels.npz`
- `softlabels_index.jsonl`
- `alignmentreport.json`
- `hash_manifest.json`
- `classorder.json`
- `ensemble_metrics.json`

Why this ensemble is chosen (evidence-based)

- **Stability (multi-source test):** highest macro-F1 among compared ensembles on `test_all_sources`.
- **Minority/hard classes:** improves Fear/Disgust/Sad F1 relative to comparable 2-teacher baselines on the same benchmark.

- **Calibration metrics available:** ensembles record NLL/ECE/Brier consistently; selected run is competitive (not necessarily best ECE, but best macro-F1 target).

Notes / limitations

- Some benchmark runs were archived during triage; their metrics remain valid, but the folder location is under `_archive/`.

- Future: re-run the best test_all_sources 0.4/0.4/0.2 benchmark into a non-archived folder for clarity (optional).