 # Report — Teacher Ensemble (v2 restart)

Week 4 (Dec 2025) summary

Week-4 work (Dec 20–23, 2025) is fully captured in this document; the day-by-day log is in:

- `research/process_log/Dec_week4_process_log.md`

basic

The v2-restart ensemble pipeline is now reproducible, alignment-safe, and benchmarked on both single-source and mixed-source tests.

RN18+B3 alone is strong, but adding CNXT improves both RAFDB-test and the larger mixed test.

Softlabels exported in logit-space with CLAHE match legacy preprocessing and provide stable KD/DKD targets.

All exported runs include manifest hashing + alignment proofs to prevent split/ordering mistakes.

Purpose

This report documents the reconstructed teacher-ensemble workflow in the v2-restart repo, the evaluation results we obtained under controlled settings, and how these results inform KD/DKD hyperparameters (especially ensemble weights and KD temperature).

The key deliverables of this ensemble step are:

- reproducible ****teacher ensemble evaluation**** on chosen benchmark splits

- ****exported softlabels**** (aligned to a manifest split) for student KD/DKD

- **metadata + alignment proof** to prevent “wrong split / wrong files / wrong label order” mistakes

Teachers used

Stage A (224×224) checkpoints:

- Teacher A (RN18):
`outputs/teachers/RN18_resnet18_seed1337_stageA_img224/best.pt`

- Teacher B (B3):
`outputs/teachers/B3_tf_efficientnet_b3_seed1337_pretrained_true_v1_stageA_img224/best.pt`

- Teacher C (CNXT):
`outputs/teachers/CNXT_convnext_tiny_seed1337_stageA_img224/best.pt`

Notes:

- B3 Stage A is the “pretrained true” retrain (fixing the earlier parity mismatch where B3 Stage A used `pretrained=false`).

Ensemble implementation (exporter)

Scripts:

- `scripts/export_ensemble_softlabels.py` (2-teacher)

- `scripts/export_multi_ensemble_softlabels.py` (N-teacher; used for 3-model ensembles)

Key options:

- `--ensemble-space prob|logit`

  - `prob`: softmax each teacher → weighted sum of probabilities → store

`log(p)`

- `logit`: weighted average of **raw logits** (preferred for KD “logits/T” semantics)

- `--use-clahe` (+ `--clahe-clip`, `--clahe-tile`) to match the legacy protocol where contrast normalization materially changes results.

Exported artifacts (per run):

- `softlabels.npz` (logits)

- `softlabels_index.jsonl` (row alignment index)

- `classorder.json` (canonical 7 order)

- `hash_manifest.json` (manifest SHA256)

- `alignmentreport.json` (full settings + alignment checks)

- `ensemble_metrics.json` (acc, macro-F1, per-class F1, NLL, ECE, Brier)

Operational utilities (run management):

- `tools/triage_softlabel_runs.py` produces a consolidated leaderboard:

  - `outputs/softlabels/_ensemble_triage.md`

- `outputs/softlabels/_ensemble_bad_list.txt` contains paths to archive

- `scripts/archive_softlabel_runs.py` moves listed runs into `outputs/softlabels/_archive/<tag>_<timestamp>`

Benchmarks evaluated (cleaned data)

A) RAF-DB basic test (cleaned)

Manifest:

- `Training_data_cleaned/rafdb_basic_test_only.csv` (3068 rows)

Settings:

- CLAHE enabled

- `--ensemble-space logit`

Results (RAFDB-basic test):

- RN18 0.3 / B3 0.7: accuracy ≈ 0.8514, macro-F1 ≈ 0.7708

  - Disgust F1 ≈ 0.5683, Fear F1 ≈ 0.6364

- RN18 0.5 / B3 0.5: accuracy ≈ 0.8563, macro-F1 ≈ 0.7775 (best among these three)

  - Disgust F1 ≈ 0.6102, Fear F1 ≈ 0.6154

- RN18 0.7 / B3 0.3: accuracy ≈ 0.8347, macro-F1 ≈ 0.7484

  - Disgust F1 ≈ 0.5392, Fear F1 ≈ 0.5714

Interpretation:

- On RAF-DB, shifting more weight toward B3 generally helps ****Disgust/Fear****; 0.5/0.5 is the best overall among the tested weights.

- If your immediate goal is to strengthen minority classes (Disgust/Fear) **even at some cost elsewhere**, 0.3/0.7 is a reasonable candidate to try next in student KD.

A2) RAFDB test (cleaned) — 3-teacher comparison

This was the “bigger/less RAFDB-basic-only” RAFDB test group used in the softlabel runs (see triage output).

Best observed runs (RAFDB test):

- RN18/B3/CNXT weights 0.4/0.4/0.2 (logit+CLAHE):

  - macro-F1 = 0.791023, acc = 0.865711

  - run: `outputs/softlabels/_ens_rn18_0p4_b3_0p4_cnxt_0p2_rafdb_test_logit_clahe_20251223_1`

- B3/CNXT 0.5/0.5 (logit+CLAHE):

  - macro-F1 = 0.788881, acc = 0.860169

  - run: `outputs/softlabels/_ens_b3_0p5_cnxt_0p5_rafdb_test_logit_clahe_20251223_1`

- RN18/B3 0.5/0.5 (logit+CLAHE):

  - macro-F1 = 0.777514, acc = 0.856258

  - run: `outputs/softlabels/_ens_rn18_0p5_b3_0p5_rafdb_test_logit_clahe_20251220_154146`

Interpretation:

- CNXT is a net positive in this regime; the 3-teacher 0.4/0.4/0.2 becomes the default “best overall” teacher ensemble.

B) Full unified test (cleaned)

Manifest:

- `Training_data_cleaned/classification_manifest.csv --split test` (49,457 rows)

Settings:

- CLAHE enabled

- `--ensemble-space logit`

Observed results:

- RN18 0.3 / B3 0.7 (best on fulltest): accuracy ≈ 0.6824, macro-F1 ≈ 0.6534
  - raw ECE ≈ 0.3008, raw NLL ≈ 4.246
- RN18 0.5 / B3 0.5: accuracy ≈ 0.6809, macro-F1 ≈ 0.6521
  - raw ECE ≈ 0.2934, raw NLL ≈ 4.408
- RN18 0.7 / B3 0.3: accuracy ≈ 0.6651, macro-F1 ≈ 0.6366
  - raw ECE ≈ 0.3112, raw NLL ≈ 5.018
- temperature scaling fit on this output hit the cap at T* = 5.0:
  - calibrated ECE ≈ 0.1410
  - calibrated NLL ≈ 1.102
  - accuracy/macro-F1 unchanged (expected; argmax unchanged)

Interpretation:

- The multi-source test is *harder* than RAF-DB-only. Old-report "~0.8" numbers are not directly comparable unless we match the same benchmark split and preprocessing.

C) Mixed-source "bigger test" (cleaned)

To reduce run-to-run variance and RAFDB-only overfitting concerns, we built a merged test manifest from multiple sources:

Manifest:

- `Training_data_cleaned/test_all_sources.csv` (48,928 rows)

Best observed runs (test_all_sources):

- RN18/B3/CNXT weights 0.4/0.4/0.2 (logit+CLAHE):
  - macro-F1 = 0.659608, acc = 0.687255
  - run: `outputs/softlabels/_ens_test_all_sources_rn18_0p4_b3_0p4_cnxt_0p2_logit_clahe_20251223_111523`
- RN18/B3 weights 0.3/0.7 (logit+CLAHE):
  - macro-F1 = 0.654132, acc = 0.682186
  - run: `outputs/softlabels/_ens_test_all_sources_rn18_0p3_b3_0p7_logit_clahe_20251223_091041`
- RN18/CNXT weights 0.5/0.5 (logit+CLAHE):
  - macro-F1 = 0.649794, acc = 0.677383
  - run: `outputs/softlabels/_ens_test_all_sources_rn18_0p5_cnxt_0p5_logit_clahe_20251223_111122`

Interpretation:

- The larger mixed-source test lowers absolute scores (expected) but is more stable for model comparisons.
- Proceed with the 3-teacher ensemble as the default teacher for student KD/DKD.

Recommended softlabels folder(s) for student training

Because the “best” teacher depends on what you want the student to generalize to, the simplest recommendation is:

If training the student on unified multi-source data (default “overall” choice)

Pick the best `fulltest` run (largest, most representative evaluation):

- `outputs/softlabels/_ens_rn18_0p3_b3_0p7_fulltest_logit_clahe_20251220_161909`
  - fulltest: accuracy ≈ 0.6824, macro-F1 ≈ 0.6534

If you want the best “overall” teacher we’ve measured so far (including CNXT), prefer the 3-teacher config validated on `test_all_sources`:

- `outputs/softlabels/_ens_test_all_sources_rn18_0p4_b3_0p4_cnxt_0p2_logit_clahe_20251223_111523`

If the student is specifically targeting RAF-DB performance

Pick the best RAFDB-test run:

- `outputs/softlabels/_ens_rn18_0p5_b3_0p5_rafdb_test_logit_clahe_20251220_154146`
  - RAFDB test: accuracy ≈ 0.8563, macro-F1 ≈ 0.7775

If CNXT is allowed in the student’s teacher set, use:

- `outputs/softlabels/_ens_rn18_0p4_b3_0p4_cnxt_0p2_rafdb_test_logit_clahe_2

0251223_1`

- RAFDB test: accuracy ≈ 0.8657, macro-F1 ≈ 0.7910

Answer to “Are the AffectNet runs best?”

For AffectNet-full-balanced specifically, yes — RN18 0.7 / B3 0.3 is best among the current runs:

- `outputs/softlabels/_ens_affectnet_full_balanced_rn18_0p7_b3_0p3_logit_clahe_20251220_145024`

- affectnet_full_balanced: accuracy ≈ 0.7960, macro-F1 ≈ 0.7954

Benchmarks evaluated (uncleaned data)

RAF-DB basic test (uncleaned)

Problem encountered:

- Uncleaned RAFDB-basic images are stored under `Training_data/RAFDB-basic/basic/Image/aligned/aligned/`.

- The initial manifest generator wrote paths under `.../Image/aligned/`, causing 3068/3068 test rows to appear missing.

Fix:

- Patched `scripts/build_uncleaned_manifests.py` to auto-detect `aligned/aligned` nesting.

Rebuilt manifest:

- `Training_data/uncleaned_manifests/rafdb_basic_manifest.csv`

- Rows: 15,339

- Missing aligned images: 0

Uncleaned RAFDB-basic test ensemble (0.5/0.5, logit+CLAHE):

- Output folder:
`outputs/softlabels/_uncleaned_rafdb_basic_test_rn18_0p5_b3_0p5_logit_clahe_20251220_v2`

- `ensemble_metrics.json`:

  - accuracy ≈ 0.8563

  - macro-F1 ≈ 0.7775

  - per-class F1: Disgust ≈ 0.6102, Fear ≈ 0.6154

Interpretation:

- Uncleaned RAFDB-basic now evaluates correctly and matches the cleaned RAFDB benchmark numbers under the same protocol.

KD/DKD tuning guidance (weights and temperature)

Ensemble weights

- For RN18+B3-only, RAF-DB prefers 0.5/0.5 among (0.3/0.7, 0.5/0.5, 0.7/0.3).

- With CNXT available, RN18/B3/CNXT 0.4/0.4/0.2 becomes the default “best overall” teacher.

Temperature (T)

We measured softlabel sharpness (teacher confidence) using `scripts/inspect_softlabels.py`.

- On RAF-DB ensemble outputs, T=1 produces extremely sharp targets (mean max-prob ≈ 0.983; p99 max-prob = 1.0).

- Increasing T to ~4–6 noticeably softens targets.

Practical recommendation:

- Start student KD/DKD sweeps with `T ∈ {4, 5, 6}`.

- Keep `T=1` only as a baseline; it is often too sharp to add useful "dark knowledge".

HQ-train softlabels export status (for student KD/DKD)

Target:

- Manifest: `Training_data_cleaned/classification_manifest_hq_train.csv`

- Split: `train` (209,661 rows)

Chosen teacher ensemble (default):

- RN18/B3/CNXT weights 0.4/0.4/0.2

- logit-space fusion + CLAHE

- dtype float16

Current export run folder:

- `outputs/softlabels/_ens_hq_train_rn18_0p4_b3_0p4_cnxt_0p2_logit_clahe_20251223_152856`

Notes:

- For very large manifests, path existence verification can dominate startup time; the exporter supports skipping this when the manifest is known-good.

What might still be missing vs the old interim report numbers

If the goal is to reproduce a specific old-table row like “RN18+B3 (0.7/0.3), T*=1.2 → acc≈0.80 / macro-F1≈0.79”, the likely missing controls are:

- exact benchmark split (very likely RAF-DB test or another narrow/easier slice)

- exact preprocessing (CLAHE parameters, crop pipeline)

- exact fusion method (prob-space vs logit-space) + whether any extra post-processing was used

- whether metrics were computed after applying a fixed temperature scaling (T*) during evaluation

Next steps (recommended)

1) Finish HQ-train softlabels export and confirm `softlabels.npz` exists in the run folder.

2) Start student KD/DKD using the exported HQ-train softlabels (3-teacher default).

3) “One by one” per-source evaluation: run ensemble metrics on each source’s test subset (cleaned), then repeat on uncleaned where applicable.

4) Student KD/DKD grid (minimal):

    - weights: {0.5/0.5} first, then {0.3/0.7} if targeting Disgust/Fear

    - temperature: {4, 5, 6}

    - alpha/beta: keep conservative at first (avoid overpowering CE)

5) Document any benchmark alignment needed to fairly compare with the old interim report.