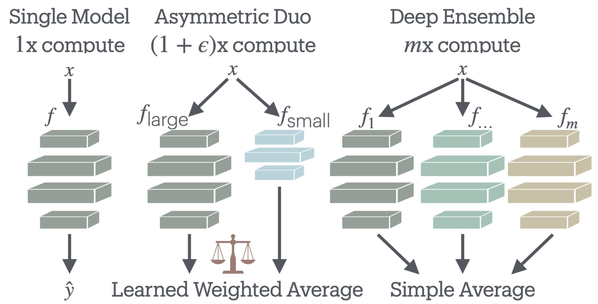
# Dynamic Duos: Robust Duo Adaptation Under Shift

Alexandre St-Aubin  Evan Shelhamer
University of British Columbia, Vector Institute  alxstaub@cs.ubc.ca

## Asymmetric Duos




Asymmetric Duos (ADs) pair a Large Model ($f_\ell$) with a computationally cheaper Small Model ($f_s$), combined via *joint temperature scaling*:

$$\hat{p}(x) = \text{softmax}\left[\frac{1}{2}\left(\frac{z_\ell(x)}{T_\ell} + \frac{z_s(x)}{T_s}\right)\right].$$

$T_\ell, T_s > 0$ minimize Negative Log-Likelihood (NLL) on **clean** validation data and are fixed at test time (**Fixed Temperature Scaling (Fixed TS)**). An oracle sweep over 231 pairs (22 torchvision architectures) selects ViT-B/16 + ResNet-50 as the best duo (+1.4 pp over best single model).

## Problem: Fixed TS Fails Under Shift

Static temperatures calibrated on clean data (Fixed TS) **cannot adapt**. Optimal $(T_\ell^*, T_s^*)$ varies *drastically* across corruption types:

Table 1: Per-corruption optimal naive temperature scaling vs. clean-fitted TS (ViT-B/16 + ResNet-50, severity 5 normalized logits). 80/20 train/test split.

| | | | Accuracy | | | |
|---|---|---|---|---|---|---|
| Corruption | $T_l^*$ | $T_s^*$ | large | small | clean TS | corr. TS |
| clean | 0.877 | 0.606 | 0.813 | 0.807 | **0.827** | **0.827** |
| fog | 0.470 | 0.973 | 0.478 | 0.393 | 0.498 | **0.530** |
| gaussian_noise | 0.450 | 4.850 | 0.359 | 0.105 | 0.272 | **0.357** |
| mean (19 corr.) | 0.636 | 2.138 | 0.393 | 0.269 | 0.375 | **0.413** |

For corruptions where $T_s^* \gg T_\ell^*$, the optimal strategy is to **suppress** $f_s$, something Fixed TS cannot do. At severity 5, Duo + Fixed TS (0.375) falls *below* ViT-B/16 alone (0.393). Corr. TS is fitted on individual shifts.

## Fitting on Corrupted Data

A natural fix: include corrupted data when fitting $(T_\ell, T_s)$. We pool 80 % of the clean ImageNet val set with 80 % of 4 *held-out* shifts (`gaussian_blur`, `speckle_noise`, `spatter`, `saturate`) and refit Fixed TS on this mixture.

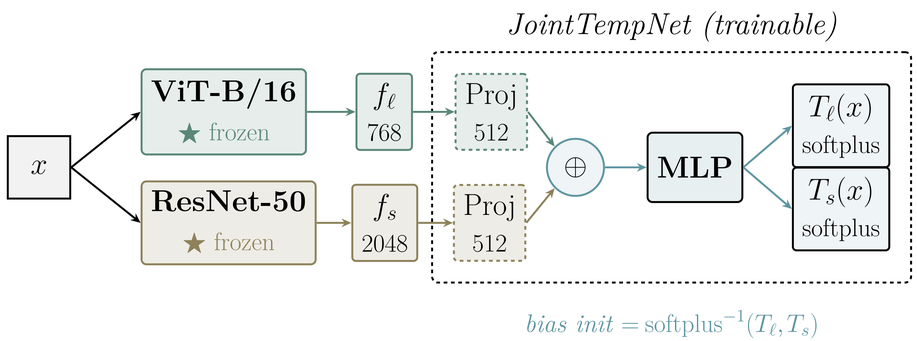
Table 2: Fixed TS fitted on clean only ($T_\ell = 0.877$, $T_s = 0.606$) vs. clean + held-out corruptions ($T_\ell = 0.621$, $T_s = 1.109$). Accuracy, Expected Calibration Error (ECE), and NLL on the held-out 20% of each shift (sev. 5).

| | Clean Fixed TS | | | Mixed Fixed TS | | |
|---|---|---|---|---|---|---|
| Corruption | Acc ↑ | ECE ↓ | NLL ↓ | Acc ↑ | ECE ↓ | NLL ↓ |
| *mean (15 test corr.)* | 0.375 | 0.083 | 3.369 | **0.404** | 0.094 | **3.212** |

The mixed calibrator learns a **higher** $T_s$, down-weighting ResNet-50. This closes roughly 75% of the gap to the per-corruption oracle (0.404 vs. 0.413) *without* test-time observation. Yet a single global $(T_\ell, T_s)$ cannot distinguish noisy from blurred images, motivating per-sample temperatures.

## Method 1: Joint Temperature Network (JointTempNet)

A small trainable MLP predicts per-sample $(T_\ell(x), T_s(x))$ from *frozen* penultimate features of both models. **No gradient computation at test time.** The network is initialized with Fixed TS temperatures, so training starts from the known-good calibrated baseline and a regularizer prevents it from catastrophically degrading.




**Loss:** NLL of the Asymmetric Duo's prediction.
**Training data:** clean ImageNet val + 4 held-out corruptions (`gaussian_blur`, `speckle_noise`, `spatter`, `saturate`), all severities. Each batch is balanced and guaranteed to contain a proportional amount of samples from each corruption.

## Method 2: Duo Test-Time Adaptation (TTA)

Following TENT, we minimize the entropy of the duo's joint prediction at test time. Update only affine parameters of normalization layers at test-time.

$$\mathcal{L} = \mathbb{E}[H(\bar{z})], \qquad \bar{z} = \frac{1}{2}\left(\frac{z_\ell}{T_\ell} + \frac{z_s}{T_s}\right)$$

**Six adaptation modes** differing in *who adapts* and *what entropy signal* drives adaptation:

| Mode | Who adapts | Entropy signal |
|---|---|---|
| Joint | Both | Duo $H(\bar{z})$ |
| Large Adapt | Large only | Duo $H(\bar{z})$ |
| Small Adapt | Small only | Duo $H(\bar{z})$ |
| Indep. large | Large only | Own $H(z_\ell)$ |
| Indep. small | Small only | Own $H(z_s)$ |
| Indep. duo | Both | Own (separate) |

**Key insight:** Using the *duo entropy* (rather than each model's individual entropy) provides a far more stable adaptation target, as shown in the results.

## Joint Temperature Network Results

To test JointTempNet's ability to generalize from corrupted training data, we construct a fixed test set by sampling 40k images from each of the 15 ImageNet-C corruptions. We then vary the training set: clean ImageNet only (−1), clean + 40k samples from each of the four held-out shifts (0), and progressively add {100, 1,000, 10,000} samples per test corruption (disjoint from the test set). We plot the mean accuracy across all 15 test corruptions in Figure 1.
With clean data alone (−1), both methods perform almost identically. The JointTempNet has no signal to learn corruption-aware temperatures and recovers the Fixed TS baseline. As soon as any corrupted data is added, the JointTempNet surpasses Fixed TS, and the gap widens monotonically: +0.5 pp at 0, +1.1 pp at 1,000, and +1.4 pp at 10,000 samples per corruption.

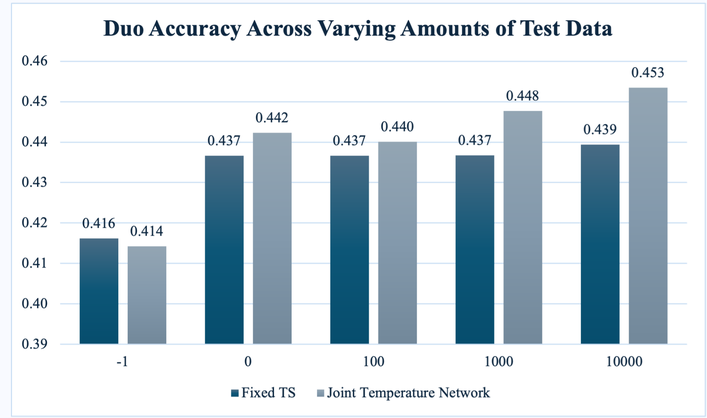



Figure 1: Mean accuracy across 15 ImageNet-C corruptions as a function of training-set composition (avg. over 5 seeds). Fixed TS is refit on the same data; the JointTempNet consistently improves with more ID training samples, while Fixed TS plateaus near 0.437.

Fixed TS, constrained to a single global $(T_\ell, T_s)$, cannot exploit the additional data, while the JointTempNet can.

## Duo TTA Results

Routing TENT through the duo's *joint* entropy, rather than each model's individual entropy, yields a dramatic jump over both Fixed TS and single-model TENT. Large Anchor TENT and Joint TENT both significantly outperform other modes across all metrics.

Table 3: Mean accuracy, ECE and NLL across all ImageNet-C corruptions at severity 5.

| | Variant | Acc ↑ | ECE ↓ | NLL ↓ |
|---|---|---|---|---|
| *Baselines* | Duo + Fixed TS (clean) | 0.375 | 0.082 | 3.353 |
| | $f_\ell$ + TENT | 0.374 | 0.508 | 7.169 |
| *Joint entropy* | Duo + Joint TENT | **0.656** | 0.075 | 1.603 |
| | Duo + Large Adapt TENT | 0.655 | **0.072** | **1.602** |
| | Duo + Small Adapt TENT | 0.538 | 0.120 | 2.548 |
| *Per-model entropy* | Duo + Indep. Large TENT | 0.486 | 0.251 | 2.840 |
| | Duo + Indep. Small TENT | 0.537 | 0.119 | 2.554 |
| | Duo + Indep. TENT | 0.488 | 0.248 | 2.775 |

## ★ Duo prevents catastrophic TENT collapse

Single-model TENT can collapse on unseen shifts since TENT hyperparameters tuned on held-out corruptions do not always generalize. Routing adaptation through the duo's joint entropy sidesteps this: the second model acts as a stable anchor that pulls predictions away from degenerate solutions.

Table 4: Single-model TENT collapses on several shifts (notably Fog with the Large model: 0.487 → 0.056), while Joint TENT remains stable and improves over both models across all four corruptions.

| Corruption | $f_s$ | $f_s$ + TENT | $f_\ell$ | $f_\ell$ + TENT | **Joint TENT** |
|---|---|---|---|---|---|
| Fog | 0.399 | 0.409 | 0.487 | 0.056 | **0.686** |
| Gaussian Noise | 0.105 | 0.113 | 0.357 | 0.200 | **0.329** |
| Glass Blur | 0.085 | 0.071 | 0.215 | 0.308 | **0.450** |
| Motion Blur | 0.178 | 0.129 | 0.332 | 0.510 | **0.554** |
| **Average** | 0.270 | 0.270 | 0.395 | 0.380 | **0.565** |

**Both models must fail simultaneously for the duo to collapse**, far less likely than either alone.