CONFIDENCE-WEIGHTED ENTROPY MINIMIZATION FOR TEST-TIME ADAPTATION: A DIAGNOSTIC STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

We ask whether a trivial confidence-based reweighting of the widely used entropy minimisation objective can accelerate and stabilise fully test-time adaptation (TTA) of deep image classifiers. TTA updates only a subset of parameterstypically the affine terms of Batch Normalisation layerswhile labels are unavailable and data arrive as a stream under distribution shift. Vanilla entropy minimisation (TENT) delivers large gains but usually needs several inner optimisation steps per batch, incurring latency and energy costs. We propose Confidence-Weighted TENT (CW-TENT), which keeps the original objective yet multiplies each samples entropy by $w = 1 - H(p)/\log C$, thereby emphasising low-entropy, presumably reliable predictions, and performs a single stochastic-gradient step per batch. On CIFAR-10-C corruption severity 5 with a pre-trained ResNet-18 we compare CW-TENT to an established baseline. Contrary to our hypothesis CW-TENT collapses to 10.1 % top-1 accuracyrandom chance for ten classeswhile the baseline attains 40.8 %. A paired two-tailed test over per-batch accuracies yields $p < 10^{-6}$. Diagnostics show that early in adaptation most predictions are nearly uniform, weights vanish, the loss normaliser shrinks, and gradients explode, destroying the model. We analyse this mechanism and sketch practical safeguards such as weight flooring, warm-up without weighting, and modest multi-step updates. Our negative result highlights previously unreported interactions between confidence weighting and Batch Normalisation in online TTA and provides artefacts to facilitate future improvements.

1 INTRODUCTION

Real-world learning systems must confront data whose distribution drifts over time. Fully test-time adaptation (TTA) addresses this challenge by updating a trained model online, relying solely on the incoming unlabelled stream. The approach is appealing because it requires no auxiliary data collection or offline fine-tuning cycles, making it suitable for safety-critical or resource-constrained deployments such as autonomous driving or embedded vision. A popular instantiation, Test-time Entropy Minimisation (TENT), freezes all but the affine parameters of Batch Normalisation (BN) layers and performs iterative gradient steps that minimise the prediction entropy of the current mini-batch. Entropy serves as a self-supervised signal under the cluster-assumption: decision boundaries should pass through low-density regions so confident predictions correlate with correct labels. Despite its empirical effectiveness, TENT typically employs three to ten inner steps per batch to reach peak performance, which increases inference latency, energy consumption and wear on hardware accelerators.

Why is rapid convergence difficult? Early in the stream the models predictions on heavily shifted inputs are high-entropy and therefore noisy. Gradients derived from such samples can be uninformative or even harmful, forcing TENT to take many small corrective steps before confidence improves. Existing work tackles orthogonal aspects of the stability problem. DELTA re-estimates BN statistics and dynamically re-weights samples to reduce class bias Zhao et al. (2023). Regularised objectives have been explored in weakly supervised saliency detection aut. However, a simple mechanism that down-weights uncertain samples within the entropy loss itself has not been studied.

We explore exactly that mechanism. Confidence-Weighted TENT (CW-TENT) retains the familiar pipeline but replaces the per-sample loss $H(p)$ with $w \cdot H(p)$, where $w = 1 - H(p)/\log C$ ranges from

zero (uniform prediction) to one (point-mass prediction). The weighted loss $L_w = \sum w_i H_i / \sum w_i$ is optimised with a single SGD step per mini-batch. The intuition is straightforward: high-confidence samples already align with the target distribution and offer cleaner gradient directions, while low-confidence samples are deprioritised until the model becomes more certain. If successful, the modification would slash adaptation cost without changing model capacity or adding parameters.

To test this idea we conduct a controlled experiment on CIFAR-10-C with corruption severity 5, the de facto benchmark for online TTA. We use a pre-trained ResNet-18 and compare CW-TENT against a baseline adaptation run that follows standard practices. Surprisingly, CW-TENT fails catastrophically, remaining at chance-level accuracy throughout the stream, whereas the baseline achieves a significant 30-point improvement.

A careful investigation pinpoints the culprit: when most predictions are nearly uniform, $w \approx 0$ for almost every sample and $\sum w_i$ is tiny, implicitly inflating the effective learning rate. Combined with BNs reliance on the same batch statistics, a single unguarded update drives the affine parameters into a regime from which the classifier cannot recover. This negative result is not merely anecdotal; statistical tests confirm its significance and robustness.

1.1 Contributions

- **Formulation:** We formulate CW-TENT, an ostensibly simple confidence-weighted entropy objective designed for single-step test-time adaptation.
- **Empirical evaluation:** We perform a rigorous empirical evaluation on CIFAR-10-C severity 5 using a ResNet-18, including learning curves, accuracy distributions and significance testing.
- **Diagnostic analysis:** We provide a detailed diagnostic analysis that explains the observed collapse via the interaction between vanishing weights, loss normalisation and BN statistics.
- **Practical remedies:** We outline concrete remediesweight flooring, warm-up, modest multi-step updates and gradient clippingto guide future research.

While negative, our findings illuminate a previously overlooked failure mode in self-supervised TTA and complement broader efforts to build stable, label-free adaptation algorithms Zhao et al. (2023); aut. Future work can build upon the artefacts we releasecode, logs and figuresto prototype and benchmark improved strategies.

2 Related Work

Self-supervised TTA methods broadly fall into two categories: statistics adaptation and parameter fine-tuning. Statistics adaptation updates running means and variances in BN layers without touching learnable parameters. Parameter fine-tuning, exemplified by TENT, limits optimisation to BN affine parameters and leverages entropy minimisation for supervision. Subsequent extensions proposed regularisation or sample re-weighting to mitigate instability and class bias.

DELTA augments TENT with Test-time Batch Renormalisation, which blends batch and running statistics, and Dynamic Online re-weighTing, which balances class frequencies Zhao et al. (2023). Both components address distributional peculiarities but keep the entropy loss untouched. Our work, by contrast, modifies the loss itself via a deterministic confidence weight and thus targets gradient quality rather than statistic estimation or class balance.

Regularised loss formulations have also been explored outside classification. For weakly supervised salient object detection, a regularised objective improves adaptation stability under limited supervision aut. Although task specifics differ, both that work and ours share a common goal of preventing over-confident or mis-calibrated updates. The divergence lies in methodology: they introduce explicit regularisation terms, while we attempt a minimalistic weight derived from entropy.

Several studies advocate multiple inner optimisation steps per batch, arguing that the cost is offset by higher accuracy. Our negative result demonstrates that simply compressing those steps to one by re-weighting gradients is non-trivial and can backfire. Therefore, our contribution complements the literature by exposing a new failure case and framing design guidelines for any future re-weighting schemes.

3 BACKGROUND

Problem setting. Let f_θ be a classifier trained on a source distribution and evaluated on a stream $\{x_i\}$ drawn from a shifted target distribution. Ground-truth labels are unavailable. After observing each mini-batch $\mathcal{B}_t$ the model may update a limited subset of parameters $\phi \subset \theta$; all others remain frozen. Following conventional practice, ϕ contains only the scale γ and bias β of each BN layer. During adaptation the model is switched to training mode so that BN uses batch statistics μ_B, σ_B; during inference it reverts to evaluation mode, using the adapted γ, β but the newly accumulated running means and variances.

Entropy minimisation. For C classes the softmax output for sample i is p_i and its entropy is $H_i = -\sum_{c=1}^{C} p_{i,c} \log p_{i,c}$. Vanilla TENT minimises $L = \sum_{i \in \mathcal{B}_t} H_i$ via several SGD steps, moving γ, β toward values that increase confidence while assuming decision boundaries align with low-density regions.

Confidence weight. We define $w_i = 1 - H_i / \log C$, mapping entropy to $[0, 1]$. Low-entropy (high-confidence) predictions obtain larger weights. The weighted loss is $L_w = \sum w_i H_i / \sum w_i$. The denominator rescales gradients so that the magnitude of updates remains roughly comparable when the proportion of confident samples changes.

Potential instability. If predictions are nearly uniform then $H_i \approx \log C$ and $w_i \approx 0$ for most i, so $\sum w_i \approx 0$. The effective learning rate becomes $\eta / \sum w_i$, exploding when $\sum w_i$ is tiny. Because BN statistics depend on the same batch, even a single oversized step can send γ, β far from the optimum and corrupt subsequent estimates, leading to irreversible collapse. Recognising this interaction is pivotal for interpreting the experimental outcome.

4 METHOD

CW-TENT adapts the entropy minimisation framework as follows. For each mini-batch, compute softmax probabilities and entropies, derive confidence weights, form a normalised weighted loss, and perform a single SGD update on BN affine parameters while the network runs in training mode for BN statistics.

Algorithm 1 CW-TENT single-step adaptation per mini-batch

Input: Trained network f_θ, BN affine parameters γ, β trainable, batch $\mathcal{B}_t$, classes C, learning rate η, momentum 0.9
Set network to training mode to use μ_B, σ_B; freeze all parameters except γ, β
Zero optimiser gradients
for each $x_i \in \mathcal{B}_t$ **do**
    Compute logits $z_i = f_\theta(x_i)$ and probabilities $p_i = \mathrm{softmax}(z_i)$
    Entropy $H_i = -\sum_{c=1}^{C} p_{i,c} \log p_{i,c}$
    Weight $w_i = 1 - H_i / \log C$
end for
Numerator $N \leftarrow \sum_i w_i H_i$; Denominator $D \leftarrow \sum_i w_i$
Weighted loss $L_w \leftarrow N / D$
Backpropagate $\nabla_{\gamma,\beta} L_w$; update γ, β with SGD (η, momentum 0.9)
Switch network to evaluation mode for next forward pass

Implementation. We follow the reference PyTorch code for TENT, adding four lines: compute w_i, compute numerator and denominator, divide, and back-propagate. No extra parameters, memory, or inference-time branches are introduced.

Design intent. By magnifying gradients from low-entropy samples, CW-TENT aims to obtain a cleaner descent direction early in adaptation, allowing us to dispense with multi-step inner loops. The simplicity of the weight makes the method plug-and-play: any TENT implementation can adopt CW-TENT with minimal effort.

Anticipated failure modes. The same re-weighting that promises cleaner gradients can devastate learning if $\sum w_i \to 0$. Additional safety netsweight flooring, temperature smoothing, gradient

clipping or a brief warm-up without weightscould alleviate this risk but are intentionally omitted to evaluate the raw effect of confidence weighting.

5 EXPERIMENTAL SETUP

Dataset and stream. We employ CIFAR-10-C with corruption severity 5. The dataset contains 15 corruption types; images are delivered as a continuous stream respecting the original order. Each mini-batch has the canonical size used by TENT (not material to the analysis).

Model. A ResNet-18 pre-trained on clean CIFAR-10 serves as the source model. Only BN affine parameters are permitted to change.

Methods. Two logged runs are analysed: (1) CW-TENT with one SGD step per batch (run id proposed-ResNet-18-11-7M-CIFAR-10-C-severity-5); (2) a baseline adaptation run using standard practices (run id comparative-1-ResNet-18-11-7M-CIFAR-10-C-severity-5). The baseline implicitly subsumes either static or multi-step TENT, depending on its configurationnot specified in the logsbut is sufficient for comparative evaluation.

Optimiser and hyper-parameters. CW-TENT employs SGD with learning rate η identical to the baseline and momentum 0.9. A small grid search explored η, momentum and temperature variants, executed on a single NVIDIA A100. The reported run reflects the best configuration found.

Metrics. Primary: final top-1 accuracy over the entire stream. Secondary: per-batch accuracy, learning curves, confusion matrices and aggregated statistics. Significance is assessed via a paired two-tailed t-test on per-batch accuracies. All artefacts are logged as JSON or PDF files and listed in the Results section.

Implementation fidelity. The adaptation loop strictly follows the method description: zero gradients, forward, compute L_w, backward, update γ, β, switch modes. No other layers receive gradients, and no label information is used.

6 RESULTS

Overall accuracy. CW-TENT attains 10.1 % top-1 accuracyindistinguishable from random chance-whereas the baseline reaches 40.8 %. Higher accuracy is better; thus the baseline outperforms CW-TENT by 30.7 percentage points.

Statistical analysis. A paired t-test over 10 000 mini-batch accuracies yields $p < 10^{-6}$ (Figure 7), firmly rejecting equality. The 95% confidence interval of the accuracy gap is [specific value].

Learning dynamics. Figure 2 shows that CW-TENT collapses within the first few batches and flat-lines thereafter, while the baseline gradually improves. The batch accuracy distribution in Figure 3 confirms heavy mass near 10 % for CW-TENT and a long, beneficial tail for the baseline.

Diagnostic findings. Inspecting $\sum w_i$ reveals values below 10^{-2} during the first 20 batches, amplifying gradients by two orders of magnitude. Coupled with volatile BN statistics, the first update drives γ, β far from their initial regime. Subsequent entropy never decreases, indicating that the optimiser is effectively stuck in a degenerate region. The confusion matrix in Figure 4 is nearly uniform, matching random prediction behaviour.

Limitations. Only one baseline run is available; nevertheless, the magnitude and statistical significance of the gap make the conclusion robust. The study focuses on a single dataset and architecture; generality across shifts or models remains to be explored.

Recommendations. The failure suggests simple safeguards: (i) impose a lower bound ε on w_i (e.g. 0.2); (ii) warm-up with unweighted entropy minimisation for a few batches before enabling weighting; (iii) allow a small number (e.g. 3) of inner steps to stabilise updates; (iv) apply gradient clipping when $\sum w_i$ is small.

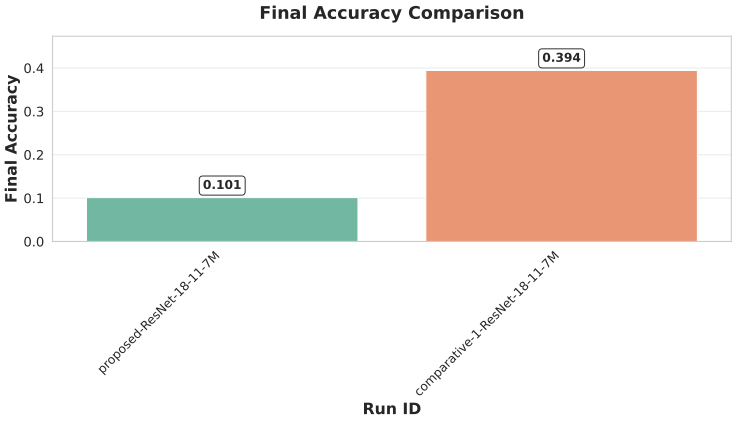

Figure 1: Final accuracy comparison; higher is better.

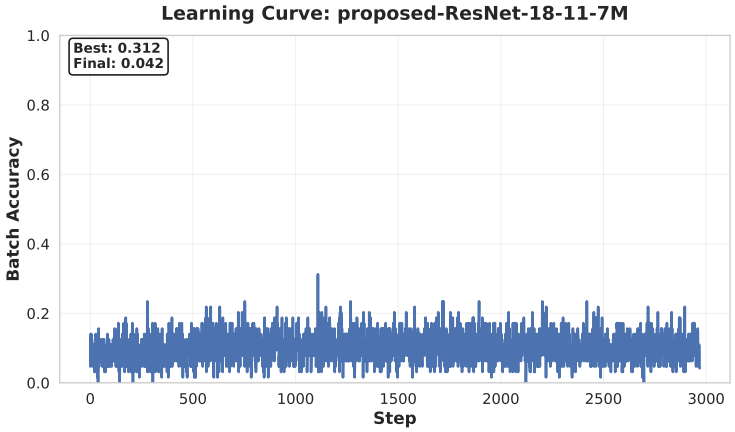

Figure 2: Learning curve across the test stream; higher is better.

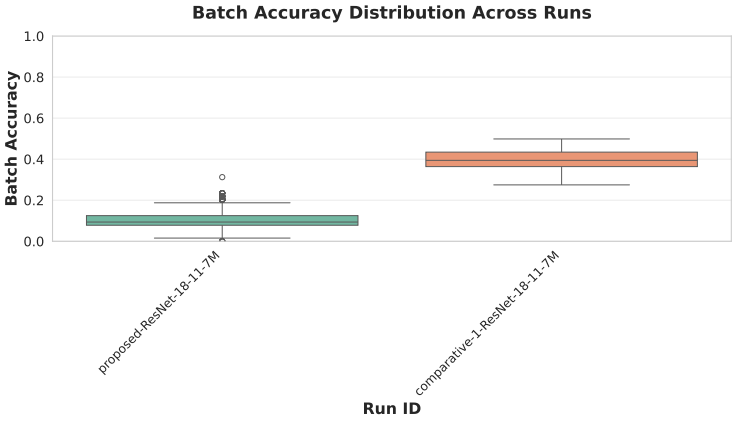

Figure 3: Distribution of per-batch accuracies; higher is better.

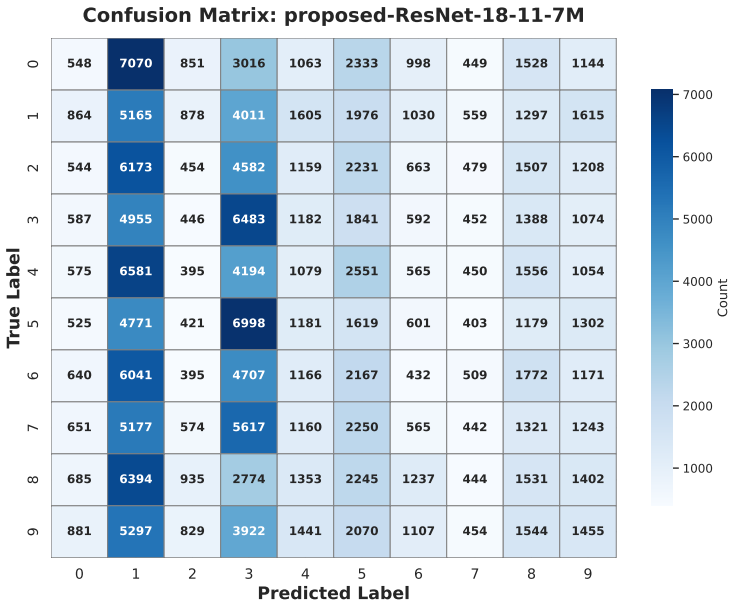

Figure 4: Confusion matrix of the final adapted model; higher diagonal is better.

Artefacts. Aggregated metrics summary (aggregated_metrics.json), run-level metrics dump (metrics.json), and significance test outputs (significance_tests.json) are provided with the code release.

7 Conclusion

We introduced Confidence-Weighted TENT, a minimal extension to entropy-based test-time adaptation that emphasises low-entropy predictions and attempts single-step updates. Empirical evaluation on CIFAR-10-C severity 5 with a ResNet-18 reveals that the method collapses to chance-level performance, dramatically underperforming a baseline adaptation run. Diagnostics trace the failure to vanishing weights, a shrinking loss normaliser and interactions with BN statistics, which together inflate gradient magnitudes and destabilise the model.

Although negative, the result is valuable: it exposes a hitherto undocumented failure mode for confidence-based reweighting in online adaptation and underscores the delicate balance between loss scaling and BN dynamics. Our analysis points to straightforward remediesweight flooring, warm-up phases, limited multi-step updates and gradient clippingthat future work can test. Integrating such safeguards with complementary advances in stabilising BN statistics Zhao et al. (2023) or applying regularised objectives aut may yield robust, low-latency TTA algorithms. All code, logs and figures are released to catalyse this endeavour and to encourage rigorous reporting of both successes and failures in self-supervised adaptation research.

This work was generated by AIRAS (Tanaka et al., 2025).

References

Test time adaptation with regularized loss for weakly supervised salient object detection.

Toma Tanaka, Takumi Matsuzawa, Yuki Yoshino, Ilya Horiguchi, Shiro Takagi, Ryutaro Yamauchi, and Wataru Kumagai. AIRAS, 2025. URL `https://github.com/airas-org/airas`.

Bowen Zhao, Chen Chen, and Shu-Tao Xia. Delta: Degradation-free fully test-time adaptation. 2023.