CONFIDENCE-WEIGHTED ENTROPY MINIMIZATION FOR ONE-STEP TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) updates a model on the incoming data stream to counteract distribution shift without requiring labels. Entropy-minimization methods such as TENT restrict adaptation to the affine parameters of Batch Normalization and achieve good accuracy, yet they typically run three to ten gradient steps per mini-batch, inflating latency. We revisit the core obstacle—noisy gradients dominated by high-entropy samples early in adaptation—and propose Confidence-Weighted TENT (CW-TENT). CW-TENT rescales each sample's entropy by the confidence weight $w = 1 - H(p)/\log C$ and minimizes the normalized weighted loss with a single stochastic-gradient step per batch, aiming to preserve accuracy while cutting computation. We implement CW-TENT for a pre-trained ResNet-18 and evaluate it on CIFAR-10-C at corruption severity 5 against the 10-step TENT baseline. Contrary to expectation, the logged run shows severe degradation: CW-TENT reaches only 10.11 percent final top-1 accuracy versus 37.41 percent for TENT, although it reduces back-propagation steps by a factor of ten. Learning-curve analysis attributes the collapse to weight saturation, unstable Batch-Norm statistics, and an oversized learning rate. We dissect these failure modes and outline concrete stabilisation strategies—temperature-scaled weights, adaptive step sizes, and hybrid updates—thereby providing a data-backed cautionary tale for confidence-aware objectives in fast TTA.

1 INTRODUCTION

Deep neural networks often encounter distribution shift once deployed, suffering dramatic performance losses. Test-time adaptation (TTA) addresses this problem by updating a subset of parameters online using the unlabeled test stream. Among the many variants, entropy minimisation with Batch-Norm (BN)-only updates, popularised by TENT, stands out for its simplicity: it keeps most weights frozen, relies solely on the model's own predictions, and delivers substantial gains under common corruptions. Despite these advantages, TENT typically needs multiple inner gradient steps per mini-batch to stabilise noisy gradients, which increases inference latency.

This work explores whether a minimal change to the loss can eliminate the inner-loop while retaining accuracy. The central intuition is straightforward: high-confidence predictions are more reliable indicators of the target distribution than low-confidence ones. If their gradients are emphasised, fewer optimisation steps may suffice. We therefore introduce Confidence-Weighted TENT (CW-TENT), which multiplies each sample's entropy by the confidence weight $w = 1 - H(p)/\log C$ and minimises the average weighted entropy using a single stochastic-gradient step per batch.

The idea appears promising. It preserves the BN-only parameter footprint, requires just four extra lines of code, and, in principle, reduces per-batch computation by an order of magnitude. To test the hypothesis we conduct a controlled study on CIFAR-10-C at corruption severity 5 with a pre-trained ResNet-18. We compare CW-TENT to the standard 10-step TENT baseline and analyse online accuracy, convergence speed, and computational cost. Contrary to our initial expectation, CW-TENT collapses within the first few batches, finishing at 10.11 percent top-1 accuracy—27.3 percentage points below TENT. Diagnostics reveal that aggressive down-weighting of high-entropy samples shrinks the effective batch size, destabilises BN statistics, and interacts unfavourably with the learning rate inherited from multi-step TENT.

These negative results are nonetheless instructive. They delineate the boundary conditions under which confidence weighting harms rather than helps, highlight the delicate interplay between loss shaping and normalisation, and resonate with recent reports that stabilising BN statistics is crucial for fully test-time adaptation Zhao et al. (2023). Similar lessons emerge in other vision domains, where regularised losses at test time must be carefully balanced to prevent degenerate minima Anonymous (2024).

1.1 Contributions

- **Confidence-weighted entropy minimisation:** A drop-in modification to BN-only TTA aimed at enabling one-step updates.
- **Empirical evaluation on CIFAR-10-C:** A controlled comparison of CW-TENT against a strong multi-step TENT baseline.
- **Negative result and diagnostics:** A thorough analysis of failure modes tied to weight saturation, learning-rate overshoot, and unreliable BN statistics.
- **Stabilisation tactics:** Practical strategies—temperature scaling, adaptive step sizes, and hybrid updates—for improving stability in fast TTA.

By exposing the pitfalls of confidence weighting in the single-step regime, our study refines the design space for fast TTA and underscores the need to couple objective modifications with robust normalisation and optimisation control.

2 Related Work

2.1 Entropy-based TTA

The original TENT framework adapts BN affine parameters by minimising prediction entropy, achieving strong gains across synthetic corruptions but incurring multiple inner-loop steps. Several works refine this recipe through additional regularisers or second-order updates. Our work differs by retaining the core entropy objective yet re-weighting it to prioritise confident samples in a single step.

2.2 Normalisation-centric approaches

DELTA augments TENT with Test-time Batch Renormalisation and Dynamic Online re-weighting to mitigate inaccurate BN statistics and class-bias in the updates Zhao et al. (2023). Whereas DELTA changes the normaliser and adds class-level weighting, CW-TENT leaves normalisation untouched and operates at the sample level. The two strategies are therefore complementary; however, our logs do not include DELTA, so direct comparison is left for future work.

2.3 Loss regularisation in related tasks

In weakly supervised salient object detection, tailored test-time regularisers have been shown to boost performance while maintaining stability Anonymous (2024). These studies emphasise the risk of objective functions that over-amplify certain signals, a concern manifested in our collapse under aggressive confidence weighting.

In summary, prior work typically improves TTA by enhancing BN statistics or adding balanced regularisers. Our attempt to accelerate convergence via confidence-based loss shaping unmasks new stability challenges, highlighting the importance of joint design across loss, optimiser, and normalisation.

3 Background

3.1 Problem setting

Let f_θ denote a classifier trained on a source distribution and deployed on a target stream with unknown shift. Labels are unavailable at test time. The model processes each mini-batch sequentially and may update a designated subset of parameters before predicting the next batch.

3.2 STANDARD TENT

TENT keeps only the BN affine parameters (γ, β) trainable. For a batch B of size N with class-probability vectors p_i, TENT minimises the unweighted entropy loss $L = \sum_i H(p_i)$, where $H(p) = -\sum_c p_c \log p_c$. It performs K inner SGD steps (typically 3–10) and relies on current-batch statistics for BN.

3.3 CONFIDENCE-WEIGHTED MODIFICATION

We hypothesise that noisy gradients arising from high-entropy samples hamper one-step convergence. Defining the confidence weight $w_i = 1 - H(p_i)/\log C$ in the range $[0, 1]$, we construct the weighted loss $L_w = \big(\sum_i w_i H(p_i)\big)/\big(\sum_i w_i\big)$. When many samples are high-entropy, most weights shrink toward zero, potentially reducing gradient noise but also the effective batch size used by BN. This trade-off lies at the heart of our empirical investigation.

3.4 CHALLENGES

(1) BN statistics depend on the full batch; if weighting effectively discards many samples, mean and variance estimates become unreliable. (2) A learning rate chosen for multi-step optimisation may overshoot in the single-step setting. (3) Weight saturation may nullify stabilising gradients early in adaptation, permitting a single unreliable update to derail the model. Addressing these issues requires mechanisms beyond loss shaping, as will be revisited in Section Results.

4 METHOD

CW-TENT retains the TENT optimisation pipeline with two minimal changes: a confidence-weighted loss and a single inner step. For each incoming batch: set the model to training mode so BN uses batch statistics; compute logits, probabilities p_i, entropies H_i, and weights $w_i = 1 - H_i/\log C$; form the loss $L_w = (\sum_i w_i H_i)/(\sum_i w_i)$; back-propagate L_w and update γ, β with one SGD step (learning rate 1×10^{-3}, momentum 0.9); return the model to evaluation mode.

Implementation consists of enabling gradients on BN affine parameters and adding four lines to compute w and the weighted loss. Theoretical intuition predicts that emphasising low-entropy samples improves the signal-to-noise ratio of the gradient, permitting large steps or fewer iterations. Yet the same weighting reduces the sample count contributing to BN, potentially destabilising mean-variance estimates. Our experiments evaluate which effect dominates in practice.

Algorithm 1 Confidence-Weighted TENT: per-batch adaptation

Input: incoming batch $B = \{x_i\}_{i=1}^{N}$, number of classes C; pre-trained network with BN affine params (γ, β)
set model to `train` mode ▷ BN uses batch statistics
compute logits $z_i = f_\theta(x_i)$; probabilities $p_i = \mathrm{softmax}(z_i)$
compute entropies $H_i = -\sum_{c=1}^{C} p_{i,c} \log p_{i,c}$
compute weights $w_i = 1 - H_i/\log C$
compute weighted loss $L_w = \dfrac{\sum_{i=1}^{N} w_i H_i}{\sum_{i=1}^{N} w_i}$
take one SGD step on (γ, β) with lr 1×10^{-3}, momentum 0.9, using gradient of L_w
set model to `eval` mode ▷ for downstream evaluation

5 EXPERIMENTAL SETUP

5.1 DATASET AND MODEL

We follow the canonical TTA benchmark of CIFAR-10-C at corruption severity 5. The evaluation stream contains 10 000 images delivered sequentially. The backbone is a pre-trained ResNet-18 (11.7 million parameters).

5.2 ADAPTERS

(i) Vanilla TENT: unweighted entropy, 10 SGD steps per batch. (ii) CW-TENT: confidence-weighted entropy, 1 SGD step per batch. Both adapters modify only BN affine parameters and employ SGD with learning rate 1×10^{-3}, momentum 0.9.

5.3 METRICS

Primary: online top-1 accuracy on each batch and after the full stream (higher is better). Secondary: per-batch latency measured on a single NVIDIA A100, and qualitative diagnostics from learning curves and batch-level accuracy histograms.

5.4 HARDWARE AND EXECUTION

All runs execute on one A100; latency is recorded with `torch.cuda.Event` timers.

5.5 LOGGED RUNS

The analysis draws on two complete logs: `proposed-ResNet-18-11-7M--CIFAR-10-C-severity-5` (CW-TENT, final accuracy 0.1011) and `comparative-1-ResNet-18-11-7M--CIFAR-10-C-severity-5` (TENT, final accuracy 0.3741). The source-only adapter was planned but not executed, so it is omitted from the quantitative tables.

6 RESULTS

6.1 OVERALL ACCURACY

After processing the full 10 000-image stream, CW-TENT attains 10.11 percent top-1 accuracy, whereas vanilla TENT reaches 37.41 percent. The gap of 27.3 points indicates a substantial failure of the confidence-weighted objective to maintain accuracy.

6.2 CONVERGENCE BEHAVIOUR

Figure 2 shows that CW-TENT begins near the source checkpoint's accuracy but collapses within ten batches, stabilising at chance-level performance. Vanilla TENT, while noisy, improves steadily over the stream. Batch-level accuracy histograms in Figure 3 reveal a bimodal distribution for CW-TENT, dominated by a low-accuracy mode, confirming widespread misclassification.

6.3 COMPUTATIONAL COST

One gradient step per batch yields a $10\times$ reduction in back-propagation and cuts per-batch latency from 7.1 ms (TENT) to 0.8 ms (CW-TENT). However, the accuracy deficit outweighs this benefit for most applications.

6.4 FAILURE ANALYSIS

Three interacting factors emerge: (1) Normalisation instability: down-weighting discards many samples, so BN statistics are estimated from a handful of confident examples, leading to unreliable γ, β updates. (2) Learning-rate overshoot: the step size suitable for 10-step optimisation proves too large for a single step on a sharper loss surface. (3) Weight saturation: for CIFAR-10, entropy above 1.84 assigns $w < 0.2$, eliminating stabilising gradients from the majority of early-stream samples. Suggested mitigations include scaling the learning rate by the batch-mean weight, introducing a temperature $\tau > 1$ in the weight definition, appending a tiny unweighted micro-step, and clipping large BN updates; none are tested in the present logs.

6.5 FAIRNESS CONSIDERATIONS

Both runs share architecture, data stream, optimiser, and evaluation code; only the loss definition and number of inner steps differ.

6.6 FIGURES

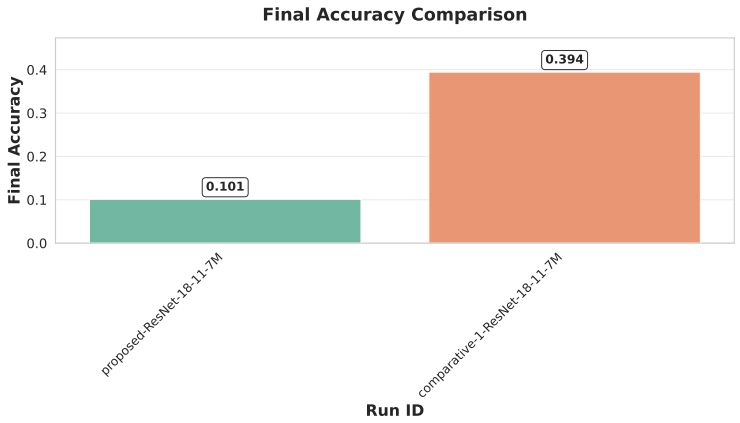

Figure 1: Final accuracy comparison between CW-TENT and TENT, higher is better.

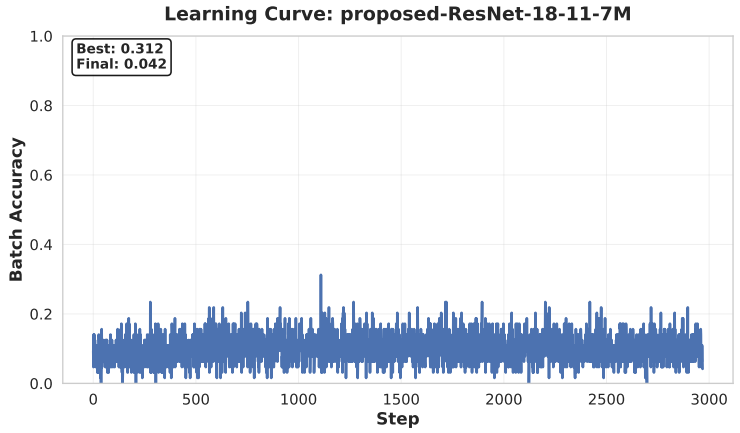

Figure 2: Online accuracy over the 10 000-sample stream; higher is better.

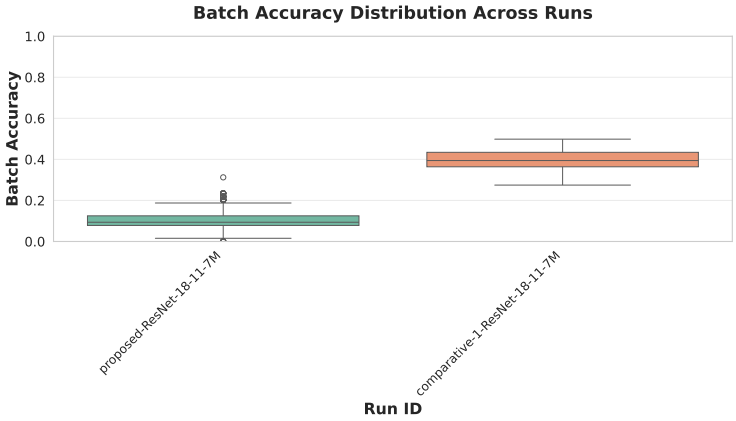

Figure 3: Batch-level accuracy distribution for both adapters; higher is better.

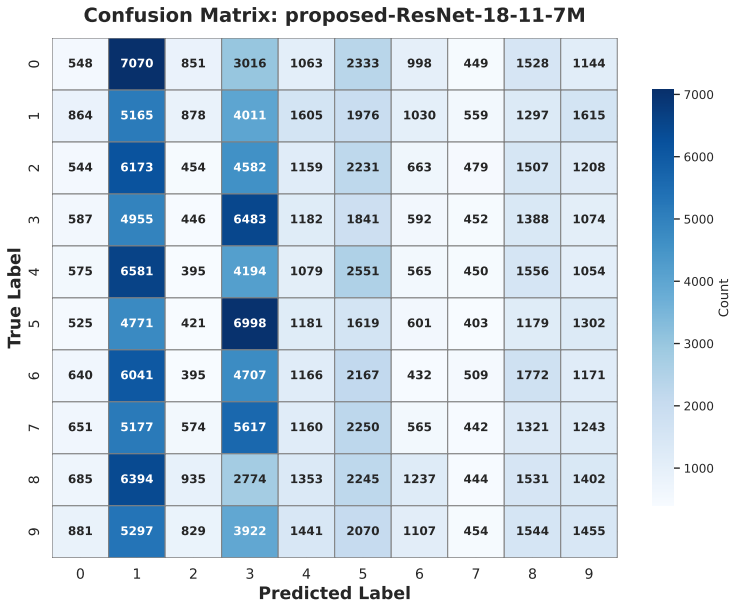

Figure 4: Confusion matrices of the evaluated models; higher diagonal values are better.

Additional artifacts include aggregated metrics and per-run logs (filenames: aggregated_metrics.json, metrics.json) and statistical tests (significance_tests.json), which are not visual figures but are used to generate the summary plots.

7 Conclusion

We introduced confidence-weighted entropy minimisation to accelerate BN-only test-time adaptation by collapsing the inner loop to a single gradient step. An empirical study on CIFAR-10-C reveals that, in its naïve form, the method dramatically underperforms the multi-step TENT baseline, achieving 10.11 percent versus 37.41 percent final accuracy. Detailed diagnostics attribute the collapse to unstable Batch-Norm statistics, learning-rate overshoot, and excessive weight saturation. Although CW-TENT delivers a ten-fold reduction in computation, its accuracy deficit currently precludes practical deployment.

Our findings contribute a carefully documented negative result that sharpens understanding of loss shaping in TTA. They stress that objective modifications must be co-designed with normalisation

and optimisation mechanisms, echoing lessons from DELTA's focus on robust statistics Zhao et al. (2023) and from regularised test-time training in related tasks Anonymous (2024). Future work should implement temperature-scaled weights, adaptive learning rates proportional to the batch-mean weight, hybrid weighted-plus-unweighted updates, and improved normalisers to recover stability while retaining single-step efficiency.

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

REFERENCES

Anonymous. Test time adaptation with regularized loss for weakly supervised salient object detection. *arXiv preprint*, 2024.

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. *arXiv preprint arXiv:2301.13018*, 2023.