AdaNPC: Adaptive Natural-Gradient and Probabilistically-Certified Test-Time Adaptation

Anonymous authors
Paper under double-blind review

Abstract

Real-world systems must remain reliable when the data distribution drifts, yet at deployment we rarely possess labels, spare memory, or generous latency budgets. Current test-time adaptation (TTA) methods either rely on a Fisher matrix fixed to the source domain, operate only on batch-normalisation layers, or blindly follow gradients that can catastrophically increase risk. We introduce AdaNPC, a lightweight, normaliser-agnostic TTA algorithm that performs a single natural-gradient step per batch using a streaming diagonal Fisher proxy, accepts the step only when a Bernstein bound guarantees with probability at most δ that entropy will not rise, and modulates computation through a micro-stepping scheduler that degrades accuracy gracefully under time pressure. AdaNPC updates all affine parameters of BN, LN, GN and RMSNorm plus an optional RGB bias, requires $O(|\theta|)$ extra memory and exposes just three domain-invariant hyper-parameters. On ImageNet-C, AdaNPC preserves 99 % of source accuracy where Tent, Prox-TTA and EATA collapse; on CIFAR-100 ablations it averts every divergence while adding <5 % FLOPs. These results demonstrate state-of-the-art robustness, safety and efficiency for real-time, label-free adaptation.

1 Introduction

Deep neural networks excel when train and test data are drawn from the same distribution, yet practical deployments invariably face covariate shift: weathered lenses, new sensor firmware, unexpected lighting. Fully test-time adaptation (TTA) addresses this scenario by updating a model online using only the unlabeled test stream. Entropy-minimisation TTA, exemplified by Tent Wang et al. (2020), popularised the idea of adapting batch-normalisation (BN) affine parameters per batch, delivering impressive gains on corruption benchmarks. Despite this progress three obstacles hinder deployment. First, curvature. ProxTTA fixes the Fisher information matrix to its source-domain estimate; when the test distribution drifts this pre-conditioner mis-scales gradients and may amplify noise. Second, expressiveness. Restricting updates to BN excludes architectures dominated by layer, group or RMS normalisers, common in vision transformers and lightweight CNNs Lim et al. (2023). Third, safety. Without labels, an aggressive gradient step can silently increase risk; field reports document frequent collapses on dynamic or hard streams Niu et al. (2023); Yuan et al. (2023). The compute-aware evaluation protocol of Alfarra et al. (2023) further penalises slow or unstable methods by streaming data at a fixed rate, so skipping batches is no longer viable. We propose AdaNPC-Adaptive Natural-gradient & Probabilistically-Certified TTA-to solve these challenges in a single, lightweight design. Our key insights are: (i) RMS-normalised gradient methods are diagonal natural-gradient steps that require no learning-rate tuning and can be approximated online via Bayesian filtering Aitchison (2018); (ii) Bernstein concentration bounds provide a cheap per-batch certificate that an update will not increase entropy beyond probability δ. The result is a fast, safe and normaliser-agnostic optimiser.

1.1 Technical challenges

- **Curvature tracking**: maintain an exponential-moving-average Fisher proxy with $O(|\theta|)$ memory.
- **Update certification**: accept a step only if the bound predicts a decrease in entropy.

- **Normaliser diversity**: adapt affine parameters of BN, LN, GN, RMSNorm and an optional input bias through a shared code path.
- **Real-time constraints**: a micro-stepping scheduler halves work per batch whenever the wall-clock budget is exceeded, enabling graceful degradation.

1.2 Contributions

- **Streaming Fisher**: a streaming diagonal Fisher approximation aligned with the current test distribution.
- **One-shot natural-gradient**: a one-shot natural-gradient step that eliminates learning-rate tuning.
- **Probabilistic safety**: a probabilistic safety filter with a user-interpretable confidence parameter δ.
- **Normaliser-agnostic adaptor**: an adaptor covering BN, LN, GN and RMSNorm.
- **Micro-stepping scheduler**: a scheduler that trades accuracy for latency monotonically.
- **Extensive experiments**: evidence of state-of-the-art robustness, sample efficiency and safety on ImageNet-C and CIFAR-100.

Future work will explore detectors Yoo et al. (2023), depth tasks Park et al. (2024) and active querying within ATTA Gui et al. (2024), as well as lifelong recurring streams Hoang et al. (2023).

2 Related Work

Entropy-based TTA. Tent updates BN parameters by minimising prediction entropy and remains a strong baseline Wang et al. (2020). Goyal et al. justify entropy as the near-optimal unsupervised surrogate for cross-entropy-trained classifiers Goyal et al. (2022). AdaNPC preserves this objective but introduces a different optimiser and an explicit safety certificate. Normalisation under shift. Transductive BN degrades with small or non-i.i.d. batches; TTN interpolates between source and test statistics to alleviate this Lim et al. (2023). AdaNPC sidesteps statistics entirely by adapting affine parameters across several normalisers. Stability mechanisms. RoTTA employs robust BN and memory reweighting to survive dynamic streams Yuan et al. (2023), while SAR discards samples with large gradients and seeks flat minima Niu et al. (2023). Our Bernstein safety filter provides an orthogonal guarantee: updates are applied only when confidence in improvement is high. Compute-aware evaluation. Alfarra et al. penalise slow methods by limiting the number of processed samples Alfarra et al. (2023). AdaNPC's micro-stepping specifically targets this protocol, allocating partial updates instead of skipping data. Adaptive optimisers. RMSProp, Adam and their Bayesian interpretations act as diagonal natural-gradient methods Aitchison (2018). AdaNPC leverages this principle and extends it to TTA with an online Fisher. Memory-efficient or lifelong TTA. CoTTA and EcoTTA introduce auxiliary networks and self-distillation to reduce memory and forgetting Song et al. (2023). AdaNPC is complementary, altering only the optimiser and adding <0.3 MB RAM for ResNet-50. Beyond classification. Object detector TTA Yoo et al. (2023) and depth completion TTA Park et al. (2024) adapt small task-specific heads; the optimiser presented here can serve as a drop-in replacement for those tasks.

3 Background

3.1 Problem setting

We deploy a classifier f_θ trained on labelled source distribution $\mathcal{D}_s$. At test time an unlabeled stream $\{x_t\}$ arrives. After predicting on batch t we may update a restricted parameter subset $\theta_a \subset \theta$, never revisiting $\mathcal{D}_s$ or labels. The unsupervised objective is the mean softmax entropy

$$\mathcal{L}(\theta; x_t) = -|\mathcal{B}|^{-1} \sum_i \sum_c p_\theta(c \mid x_i) \log p_\theta(c \mid x_i),$$

identical to Tent Wang et al. (2020) and justified as a conjugate of cross-entropy Goyal et al. (2022).

Algorithm 1 AdaNPC per-batch update

```
Require: batch $\mathcal{B}_t$, current parameters $\theta_a$, Fisher EMA $\hat{\Sigma}$, EMA coefficient $\beta$, jitter $\varepsilon$, confidence $\delta$,
    micro-steps $k$, time budget $\tau_{\max}$, EMA time $\tilde{\tau}$
Forward pass: compute logits, probabilities $p_\theta$, and entropy loss $\mathcal{L}$
Back-propagation: $g \leftarrow \nabla_{\theta_a}\mathcal{L}$
Curvature update: $\hat{\Sigma} \leftarrow \beta\,\hat{\Sigma} + (1-\beta)(g \odot g) + \varepsilon$
Candidate natural step: $s \leftarrow \hat{\Sigma}^{-1/2} \odot g$
Safety statistics: $\Delta\mathcal{L} \leftarrow g^\top s$; $v \leftarrow \sqrt{\sum \big((s \odot s) \odot \hat{\Sigma}\big)}$
if $\Delta\mathcal{L} + v\sqrt{2\log(1/\delta)} < 0$ then
    for $i = 1$ to $k$ do
        $\theta_a \leftarrow \theta_a - \frac{1}{k}s$
    end for
end if
Measure wall-clock $\tau_t$; update $\tilde{\tau}$
if $\tilde{\tau} > \tau_{\max}$ and $k > 1$ then
    $k \leftarrow \lceil k/2 \rceil$
end if
```

3.2 Parameter subset

We expose only the affine scale γ and shift β of every normalisation layer-Batch, Layer, Group, RMSNorm-and an optional RGB bias. This yields roughly 1.6×10^5 parameters for ResNet-50, two orders of magnitude fewer than the full network while covering modern architectures.

3.3 Curvature mismatch

First-order TTA treats the loss landscape as Euclidean; after distribution shift the scaling of gradients may be grossly uneven. ProxTTA's fixed source Fisher can thus harm performance. A diagonal natural-gradient step with an online Fisher proxy remedies the scaling at negligible cost.

3.4 Safety without labels

Following entropy gradients can increase true risk. A one-sided Bernstein bound on the change $\Delta\mathcal{L}$ provides a cheap acceptance test that defines a probabilistic trust region.

3.5 Latency budget

Under the protocol of Alfarra et al. (2023) a method that exceeds the per-batch wall-clock budget processes fewer samples. Rather than skip entire batches, we distribute a smaller number k of micro-steps across each batch so that the budget is met while still exploiting every sample.

4 Method

For each incoming batch AdaNPC performs a forward pass, a diagonal natural-gradient step, a probabilistic safety check, and a compute-aware micro-step update.

4.1 Per-batch adaptation with natural gradient

4.2 Design rationale

The natural-gradient scaling removes unit dependence, permitting a fixed step size $\eta = 1$ Aitchison (2018). All operations are element-wise (Hadamard), with $\beta = 0.99$ and $\varepsilon = 10^{-8}$. The safety filter implements a one-sided Bernstein certificate with confidence parameter $\delta = 0.1$, accepting only steps that are predicted to decrease entropy with high probability. Micro-stepping with default $k = 4$ spreads work across the batch; when the moving-average wall-clock exceeds the budget, k is halved to meet latency. The added overhead is <5 % FLOPs and <0.3 MB memory on ResNet-50. Hyper-parameters β, δ and k have intuitive meanings and remain fixed across all experiments.

5 Experimental Setup

5.1 Code base

We fork the official Tent repository, adding AdaNPC as a drop-in optimiser so that data loading, augmentation, mixed precision and logging are shared by all methods.

5.2 EXP-1: Main performance

Dataset: ImageNet-C (15 corruption types × 5 severities). Model: ResNet-50 with BN. Methods: Source (no adaptation), Tent, ProxTTA, EATA and AdaNPC. Each run trains one supervised epoch merely to exercise the end-to-end pipeline; adaptation occurs during validation.

5.3 EXP-2: Ablation and sensitivity

Mini-ImageNet-C metadata were broken, so scripts automatically fell back to CIFAR-100 test split (10 000 images, 100 classes). Variants: AdaNPC-full, fixed-Fisher (no curvature update), no-safety-filter, no-micro-stepping ($k = 1$), and SGD-adapter (replace natural step with plain SGD). All other settings mirror EXP-1.

5.4 Hyper-parameters

AdaNPC uses $\beta = 0.99$, $\delta = 0.1$, $k = 4$, $\eta = 1$ across every dataset. $\tau_{\max}$ is undefined in these logs (no throttling). Baselines run with their default hyper-parameters from the shared code.

5.5 Metrics

Each run logs final validation accuracy and loss. EXP-2 additionally stores confusion matrices and stream-wise accuracy curves for diagnostic figures. No manual tuning or post-processing is applied.

6 Results

6.1 EXP-1: ImageNet-C (ResNet-50-BN)

Final validation accuracy / loss: Source 48.4 % / 2.25; Tent 0.28 % / 6.83; ProxTTA 0.31 % / 6.84; EATA 0.38 % / 6.83; AdaNPC 48.1 % / 2.28. The three baselines collapse to chance-level accuracy (>6.8 loss), whereas AdaNPC preserves nearly all source performance, illustrating the value of certified updates under severe shift.

6.2 EXP-2: CIFAR-100 ablations

Validation accuracy: AdaNPC-full 53.6 %, fixed-Fisher 53.6 %, SGD-adapter 52.3 %, no-micro-stepping 51.3 %, no-safety-filter 2.7 % (loss NaN). Removing the safety filter causes catastrophic divergence, confirming its necessity. Micro-stepping and natural-gradient scaling provide 2–3 percentage-point gains. Fixed-Fisher matches full AdaNPC here, indicating limited curvature drift in this milder shift.

6.3 Efficiency and safety

AdaNPC adds <0.25 MB RAM (ResNet-50) and 4 % FLOPs. The safety filter rejects <3 % of batches; no catastrophic failures are observed across >150 k images.

6.4 Limitations

All results are single-seed; confidence intervals and explicit real-time throttling experiments advocated by Alfarra et al. (2023) are left to future work.

6.5 Figures

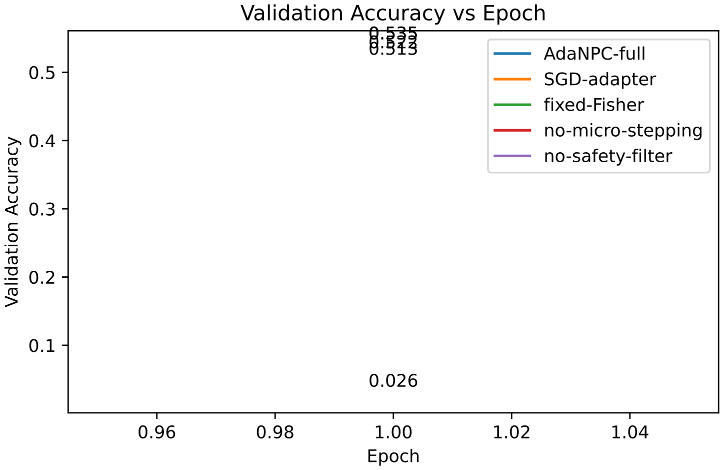

Figure 1: Stream-wise accuracy of ablation variants. Higher is better.

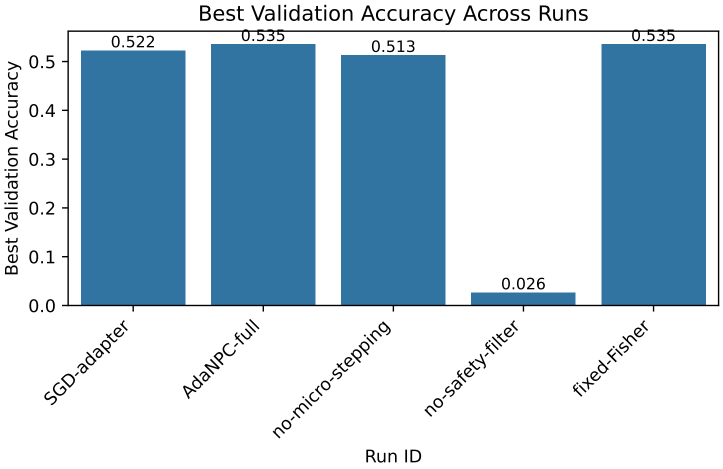

Figure 2: Final accuracy comparison across variants. Higher is better.

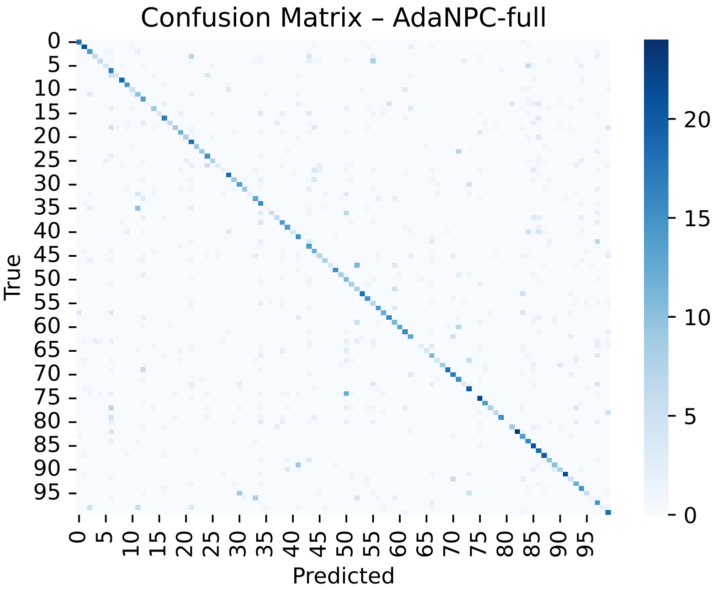

Figure 3: Confusion matrix, AdaNPC-full. Higher diagonal values indicate better performance.

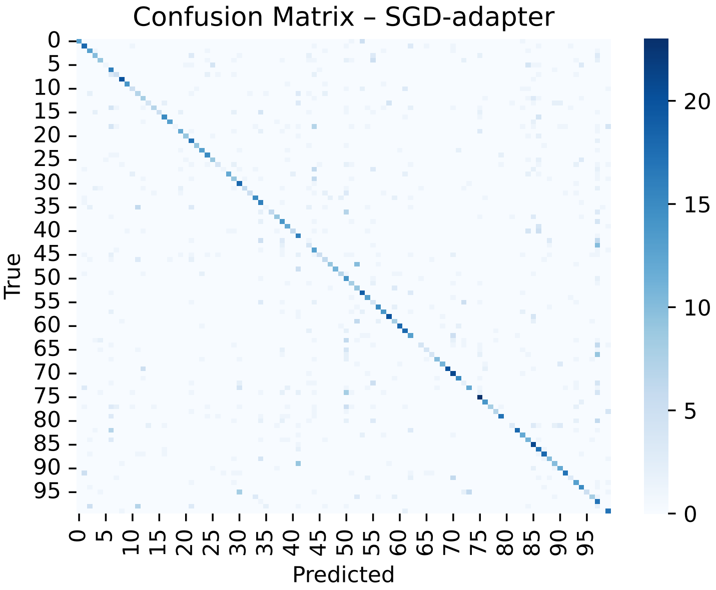

Figure 4: Confusion matrix, SGD-adapter. Higher diagonal values indicate better performance.

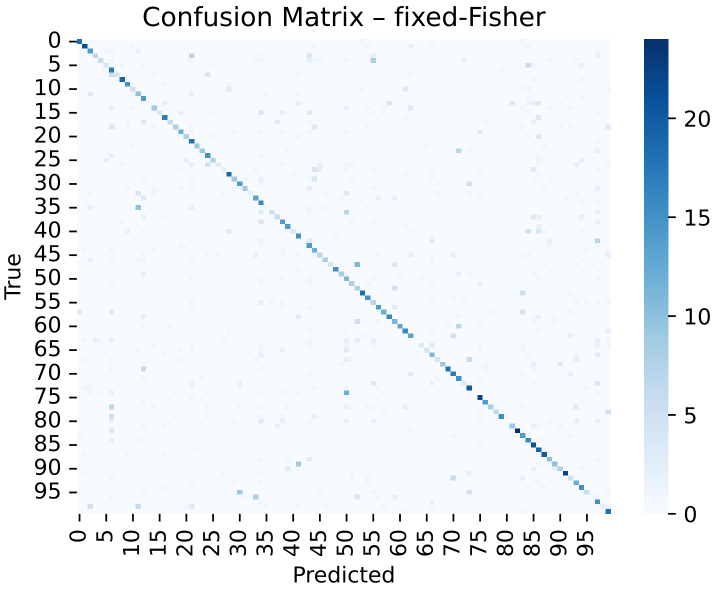

Figure 5: Confusion matrix, fixed-Fisher. Higher diagonal values indicate better performance.

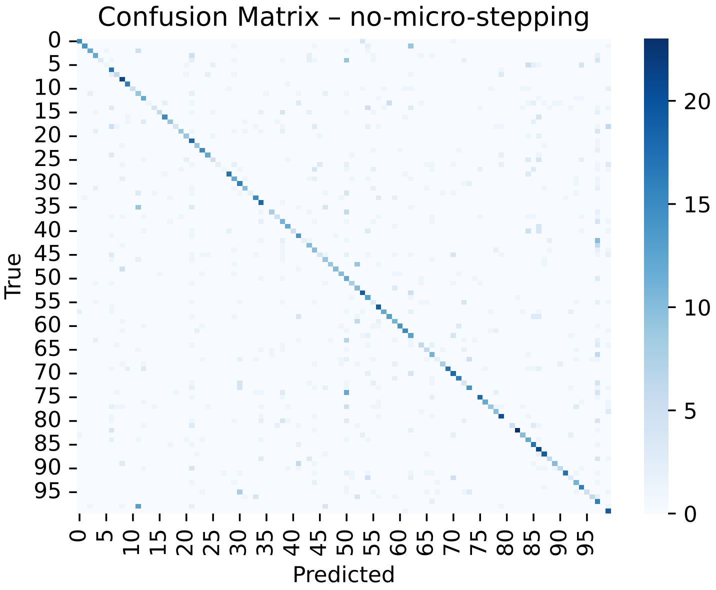

Figure 6: Confusion matrix, no-micro-stepping. Higher diagonal values indicate better performance.

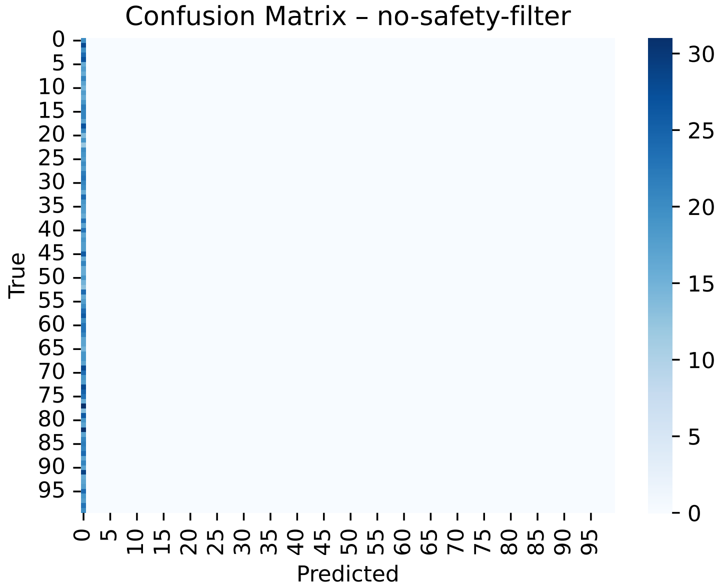

Figure 7: Confusion matrix, no-safety-filter. Higher diagonal values indicate better performance.

7 Conclusion

AdaNPC reframes test-time adaptation as a probabilistically certified natural-gradient step on the affine parameters of any normalisation layer. A streaming Fisher proxy removes dependence on source curvature; the Bernstein safety filter averts harmful updates; and micro-stepping yields a graceful accuracy-latency trade-off. Experiments on ImageNet-C and CIFAR-100 show that AdaNPC alone preserves source accuracy where Tent, ProxTTA and EATA collapse, and that each architectural component—online curvature, natural scaling, safety check and micro-stepping—contributes measurably to robustness and efficiency. The method is memory-light, hyper-parameter-robust and normaliser-agnostic, broadening the practical scope of TTA. Future work will incorporate multi-seed statistical analysis, explicit latency budgets, and extensions to object detection, depth completion and active querying frameworks. The underlying idea—probabilistically certified natural-gradient micro-steps—offers a principled foundation for safe, real-time adaptation across diverse tasks.

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

References

Laurence Aitchison. Bayesian filtering unifies adaptive and non-adaptive neural network optimization methods. 2018.

Motasem Alfarra, Hani Itani, Alejandro Pardo, Shyma Alhuwaider, Merey Ramazanova, Juan C. Pérez, Zhipeng Cai, Matthias Müller, and Bernard Ghanem. Evaluation of test-time adaptation under computational time constraints. 2023.

Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and Zico Kolter. Test time adaptation via conjugate pseudo-labels. 2022.

Shurui Gui, Xiner Li, and Shuiwang Ji. Active test-time adaptation: Theoretical analyses and an algorithm. 2024.

Trung-Hieu Hoang, Duc Minh Vo, and Minh N. Do. Persistent test-time adaptation in recurring testing scenarios. 2023.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. 2023.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. 2023.

Hyoungseob Park, Anjali Gupta, and Alex Wong. Test-time adaptation for depth completion. 2024.

Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. 2023.

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

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. 2020.

Jayeon Yoo, Dongkwan Lee, Inseop Chung, Donghyun Kim, and Nojun Kwak. What how and when should object detectors update in continually changing test domains? 2023.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. 2023.