УДК 004.8:004.056

OSTAP STETS, IHOR KONOVALENKO

RESOURCE-AWARE TEST-TIME ADAPTATION FOR MOBILE FACE ANTI-SPOOFING UNDER SWAP CONSTRAINTS

Ternopil Ivan Puluj National Technical University, 56 Ruska str., Ternopil, Ukraine,
e-mail: ostap.stets@gmail.com

Анотація. Моделі захисту від підміни обличь (FAS) для мобільної автентифікації стикаються зі складним розривом між навчанням і розгортанням: вони повинні залишатися стійкими до зміщення доменів (камера, освітлення, носій атаки), одночасно балансуючи між біометричною точністю та обмеженнями на швидкість, розмір моделі та енергоспоживання (SWAP). Існуючі методи адаптації під час тестування покращують точність у міждоменних сценаріях, але ігнорують ресурсні обмеження пристрою, а ще критичніше – можуть катастрофічно деградувати, коли вихідний та цільовий домени суттєво відрізняються. Ми пропонуємо ресурсно-усвідомлену адаптацію при тестуванні (RA-TTA), яка оновлює лише афінні параметри нормалізації, компактний класифікаційний модуль і прототипи класів, і лише тоді, коли: (а) виявлено зміщення відносно каліброваної вихідної статистики; (б) псевдо-мітки проходять надійнісну фільтрацію за впевненістю та узгодженістю аугментацій; (в) бюджетний контролер типу token-bucket дозволяє адаптацію в межах жорстких обмежень за швидкістю та енергоспоживанням. На моделі MobileNetV3-Small (1 M параметрів) з одним вихідним доменом, оціненій на наборах даних OULU-NPU та Replay-Attack, RA-TTA забезпечує відносне зниження ACER на 25% (4.83% проти 6.47%) у внутрішньодоменних умовах, та зберігає поведінку вихідної моделі при екстремальному міждоменному зміщенні (43.70% на OULU до Replay-Attack), тоді як Tent деградує на 5.7% ACER. Профіль виконання моделі є придатним для розгортання – 1.61 мс/3.04 мДж на Samsung Galaxy S25, 3.83 мс/3.95 мДж на A56, 5.94 мс/6.27 мДж на A17, при розмірі 3.84 МБ – але міждоменна точність на такому масштабі моделі є нижчою за поріг для розгортання в реальних продуктах із безпекокритичною автентифікацією. Внеском є саме механізм адаптації безпеки, а не рекордна точність; досягнення точності продуктового рівня на такій архітектурі потребує багатоджерельного навчання, що не стосується проблеми, яку розглядає ця стаття.

Ключові слова: захист від підміни обличчя, мобільна біометрія, ресурсно-усвідомлена адаптація при тестуванні, зміщення домену, оптимізація SWAP.

Abstract. Face anti-spoofing (FAS) models for mobile authentication face a difficult deployment gap: they must remain robust under domain shift (camera, illumination, attack medium) while balancing biometric accuracy against strict speed, model-weight, and power-consumption limits (SWAP). Existing test-time methods improve cross-domain accuracy but ignore on-device resource limits and, more critically, can catastrophically degrade when source and target domains differ substantially. We propose a resource-aware test-time adaptation (RA-TTA) framework that updates only normalization affine parameters, a compact classifier head, and class prototypes, and only when (a) drift is detected against calibrated source statistics, (b) pseudo-labels pass a confidence-and-augmentation reliability gate, and (c) a token-bucket budget controller permits adaptation under hard speed and power caps. Evaluated on OULU-NPU and Replay-Attack datasets with a deliberately challenging single-source 1 M-parameter MobileNetV3-Small backbone, RA-TTA delivers a 25% relative ACER reduction (4.83% vs 6.47%) on intra-domain data where adaptation safely fires, while preserving source-model behaviour under extreme cross-domain shift (43.70% on OULU to Replay-Attack, identical to No-TTA) where Tent collapses by 5.7% ACER. A drift-threshold sensitivity sweep validates the calibration heuristic. The runtime profile is deployment-class – 1.6 ms/3.04 mJ per frame on Samsung Galaxy S25, 3.83 ms/3.95 mJ on Galaxy A56, and 5.94 ms/6.27 mJ on entry-level Galaxy A17, with a 3.84 MB ONNX footprint – but cross-domain ACER at this scale is below production thresholds for security-critical face authentication. The contribution is the safety mechanism, not state of the art accuracy. Reaching production accuracy on the same backbone requires multi-source training, which is orthogonal to the adaptation-safety question this paper addresses.

Keywords: face anti-spoofing, mobile biometrics, resource-aware test-time adaptation, domain shift, SWAP optimization.

DOI: 10.31649/1681-7893-2026-51-1-79-89

INTRODUCTION

Face recognition has become a default authentication mechanism in mobile payment, access control, and e-government applications. Its main security weakness remains presentation attacks, including print, replay, and mask-based impersonation [1, 2, 3]. Face anti-spoofing (FAS) is therefore required to estimate whether the presented sample is bona fide or attack [4].

Despite major progress, performance degrades significantly in unseen domains. In mobile deployment, this degradation is amplified by hardware heterogeneity (sensor type, ISP pipeline, lens quality), environmental variation, and evolving attack types. Recent work on domain generalization and test-time techniques improves robustness [5, 6, 7], while prompt- and generation-based methods further extend cross-domain coverage [8, 9]. However, these approaches are usually optimized for server-class inference and rarely report strict on-device resource behaviour [10, 11].

This work targets robust face anti-spoofing under explicit mobile constraints. We adopt the SWAP framework as defined in our prior journal study [11]; for completeness, Speed denotes per-frame latency (ms), Weight denotes model footprint (MB), and Power denotes per-frame energy (mJ). We design Resource-Aware Test-Time Adaptation under SWAP Constraints (RA-TTA) to balance speed, weight footprint, biometric accuracy, and power-consumption during deployment. Adaptation is treated as a selective online update process rather than an always-on routine, and we explicitly avoid full-backbone test-time retraining.

The paper makes four practical contributions:

1. A constrained adaptation formulation that jointly optimizes security risk and SWAP budgets.
2. A drift-triggered adaptation mechanism that activates updates only when needed.
3. A reliability-gated self-supervision objective that reduces negative pseudo-label feedback.
4. An experimental protocol for reporting cross-domain security and on-device efficiency together.

1. THEORETICAL AND METHODOLOGICAL FOUNDATIONS

1.1. Domain Generalization for Face Anti-Spoofing

Cross-domain face anti-spoofing has driven a rich literature on domain generalization (DG): meta-learning [4], adversarial alignment, contrastive regularization, and gradient-aligned multi-source training [7]. Recent diffusion-based generation [9] and frequency-shortcut analysis [12] further expand the cross-domain operating envelope. Multimodal and unified attack categorisation methods [13, 14] address physical and digital spoofs jointly. These advances substantially improve held-out-domain accuracy but typically require multi-source training (e.g., the four-domain CIM protocol) and backbones in the 25-300 MB range, both of which exceed strict mobile deployment envelopes. Our work is complementary to DG: we assume any DG-trained mobile-class model and address the post-deployment problem of online drift safely and within SWAP budgets.

1.2. Test-Time Adaptation and Test-Time Domain Generalization

General TTA methods update model parameters during deployment by minimizing entropy or consistency losses [5, 15]. Subsequent work has substantially improved stability under realistic deployment streams: EATA [16] adds active sample selection to filter low-quality entropy gradients; SAR [17] introduces sharpness-aware updates and a reliable-loss criterion to avoid collapse; SoTTA [18] handles noisy streams with input-level filtering; and DELTA [19] addresses class-imbalanced test streams via class-balanced entropy. Resource-aware variants are also emerging: EcoTTA [20] minimizes memory via meta-network distillation and self-distilled regularization, while MECTA [21] introduces explicit memory-budget control as an integral part of the continual-TTA loop. These methods share a common premise – that adaptation is always valuable when triggered – and refine the gating criterion at the sample or system level. Our drift-trigger and budget-controller mechanisms are complementary: they operate at the batch-window level and on a per-deployment SWAP envelope, and either can be combined with the sample-level refinements above. In FAS specifically, test-time domain generalization has demonstrated that style-space projection can improve unseen-domain behavior without full retraining [6]. Continual learning has also been studied as a route to long-term cross-domain robustness [22]; however, it still requires periodic supervised retraining rather than fully unlabeled stream-level adaptation. Resource-constrained mobile settings remain underexplored: many methods assume frequent updates, large batch statistics, or nontrivial memory overhead. Vanilla Tent in particular is documented as fragile under class-imbalanced binary streams [19]; we observe this directly on our task, where Tent degrades ACER by 5-7 points cross-domain (Table 1) and by approximately 42 points in the intra-domain setting (where No-TTA performs well at 6.48% ACER). A diagnostic sweep over batch sizes and learning rates confirms that the collapse is robust to standard hyperparameter tuning.

1.3. Efficient and Mobile Face Anti-Spoofing

Mobile face anti-spoofing research emphasizes lightweight backbones, distillation, and modality-aware design [10, 11, 23]. Recent work also explores semantic-anchor supervision for unified physical and digital attacks [14]. Yet most efficient models are evaluated in static offline settings and provide limited guidance on handling online domain shift after deployment. This gap motivates a unified method that remains adaptive under strict SWAP limits.

1.4. Position with Respect to Our Prior Work

Our earlier conference study on privileged multi-teacher distillation (PI-KD) [10] addressed the training-time problem of injecting depth and rPPG supervision into a compact RGB student. The earlier journal paper [11] surveys SWAP-aware optimization through knowledge distillation as an offline compression strategy. The threshold-selection paper [25] proposed an Environmental Adjustment factor for static operating-point calibration. The present work is orthogonal: it assumes that any compatible mobile FAS model has already been trained and deployed, and addresses the post-deployment problem of online adaptation under domain shift within the same SWAP envelope. The Budget Controller can be viewed as a time-varying, stream-conditioned generalization of the static Environmental Adjustment factor.

2. PROBLEM FORMULATION

Let $f\theta: \mathcal{X} \to [0, 1]$ be a mobile face anti-spoofing model producing spoof probability $p_t = f\theta(x_t)$ at time t for input x_t. The test stream $\{x_t\}$ is unlabeled and non-stationary. Let θ_t denote model parameters at time t, and let π be an adaptation policy.

We define the constrained adaptation objective:

$$\min_{\pi} \mathbb{E}\left[\mathcal{L}_{sec}\left(f_{\theta_t}, x_t\right)\right] \; s.t. \;\; W(\theta_t) \leq W_{max}, \bar{L}_t \leq L_{max}, \bar{E}_t \leq E_{max} \tag{1}$$

where $\mathcal{L}_{sec}$ is a security-risk surrogate (sliding-window ACER on pseudo-labelled samples), $W(\theta_t)$ is model weight footprint, $\bar{L}_t$ is mean per-frame latency, and $\bar{E}_t$ is mean per-frame energy. To connect security and efficiency, we report APCER, BPCER, ACER, and EER as defined in [25] and ISO/IEC 30107-3, together with SWAP metrics: speed (median latency), weight footprint (MB), and power-consumption (mJ/frame).

3. PROPOSED METHOD: RA-TTA

3.1. Overview

RA-TTA contains three modules: (1) Drift Estimator – computes online shift score from features and uncertainty; (2) Reliability-Gated Adapter – updates only normalization affine parameters and a small classifier head; (3) Budget Controller – allocates adaptation steps under speed and power-consumption credits. The base detector is instantiated with MobileNetV3-Small. To keep deployment stable, the convolutional body is frozen at test time, keeping the method within the scope of lightweight, resource-aware TTA rather than heavy full-model adaptation. Figure 1 illustrates how the three modules fit together within the broader face authentication pipeline.

The internal structure of the FAS module – three independent safety gates (drift detector, reliability gate, budget controller) sit above the standard inference path and conditionally permit parameter updates to the last-K BatchNorm layers of the encoder and to the classifier head; adaptation fires only when all three gates pass simultaneously.

3.2. Drift-Aware Adaptation Trigger

Let $z_t = f^{\text{enc}}(x_t)$ be the penultimate feature. We maintain a reference Gaussian $\mathcal{N}(\hat{\mu}_S, \hat{\sum}_S)$ from source validation features, and a streaming Gaussian $\mathcal{N}(\hat{\mu}_t, \hat{\sum}_t)$ updated by exponential moving average over a window W. The drift score combines a feature-shift term and a predictive-uncertainty term:

$$\delta_t = \alpha \cdot D_{KL}\left(\mathcal{N}(\hat{\mu}_t, \hat{\sum}_t) \| \mathcal{N}(\hat{\mu}_S, \hat{\sum}_S)\right) + \beta \cdot \bar{H}_t \tag{2}$$

where $\alpha \geq 0$ is a mixing weight, and D_{KL} between diagonal Gaussians admits a closed form. The threshold τ_{drift} is calibrated on held-out source-domain validation streams (target false-trigger rate: 5%). We initially included an additive predictive-entropy term $\beta \cdot \bar{H}_t$ with $\bar{H}_t$ the per-window mean entropy, but in our 576-dimensional feature space the KL term is in the hundreds and the entropy term is bounded by $ln(2) \approx 0.693$, so any $\beta > 0$ is dominated by the KL contribution. We therefore set $\beta = 0$ and use the KL term alone.

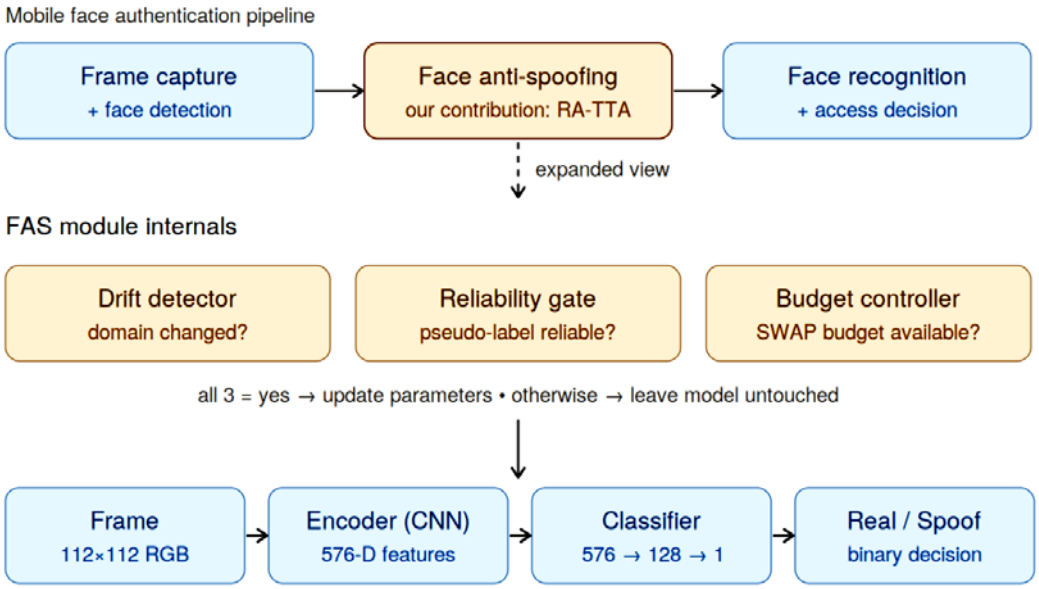

Figure 1 – Placement of RA-TTA in the mobile face authentication pipeline

3.3. Reliability-Gated Self-Supervision

Pseudo-labeling is high-risk in FAS because confident mistakes can be persistent. We gate updates by confidence and augmentation consistency. Let $\tilde{x}_t$ denote a weakly augmented view (color jitter and mild Gaussian blur). The per-sample gate is:

$$g(x_t) = \mathbb{1}\left[\max_c p_\theta(c|x_t) \geq \tau_{conf}\right] \cdot \mathbb{1}[\text{argmax}\, p_\theta(\cdot\,|x_t) = \text{argmax}\, p_\theta(\cdot\,|\tilde{x}_t)] \quad (3)$$

Only frames with $g(x_t) = 1$ contribute to adaptation. The adaptation loss is:

$$\mathcal{L}_{TTA} = \lambda_1 \mathcal{L}_{BCE} + \lambda_2 \mathcal{L}_{cons} + \lambda_3 \mathcal{L}_{proto} \quad (4)$$

The three components are: $\mathcal{L}_{BCE}$ – weighted BCE from gated pseudo-labels; $\mathcal{L}_{cons}$ – prediction consistency under augmentation $\|p_\theta(x_t) - p_\theta(\tilde{x}_t)\|^2$; $\mathcal{L}_{proto}$ – L2 distance from class prototypes μ_c. Prototype-based test-time methods such as T3A [24] have shown that maintaining and updating class templates without backpropagation can stabilise predictions under distribution shift; we use prototypes here as a regularisation anchor combined with backprop on the head and the last-K BatchNorm affines. Prototypes are initialised from source validation data and updated online by EMA with momentum m_p:

$$\mu_c \leftarrow (1 - m_p)\mu_c + m_p z_t \;\; when\; \hat{y}_t = c \text{ and } g(x_t) = 1 \quad (5)$$

Only BatchNorm/LayerNorm affine parameters $\{\gamma, \beta\}$ in the last K_{adapt} blocks, the binary classifier head, and the prototypes are updated. Convolutional kernels and earlier normalization layers remain frozen.

3.4. SWAP Budget Controller

We model the adaptation budget as a token bucket with credit B_t:

$$\mathrm{B}_{t+1} = \min(\mathrm{B}_{max}, \mathrm{B}_t + \rho - a_t \cdot c_a) \quad (6)$$

where $a_t \in \{0,1\}$ indicates whether adaptation occurred, c_a is the estimated wall-clock cost per adaptation step (ms), and ρ is the per-frame refill rate. The controller selects adaptation steps as:

$$K_t = \left\lfloor \frac{B_t}{c_a} \right\rfloor \text{ if } \delta_t > \tau_{drift};\;\; K_t = 0 \text{ otherwise} \quad (7)$$

To enforce hard guarantees on speed and power, K_t is clipped: $K_t \leftarrow \min\left(K_t, K_{max^L}, K_{max^E}\right)$. K_{max^L} and K_{max^E} are derived from latency budget L_{max} and energy budget E_{max} respectively. Adaptation is disabled during thermal throttling. This yields deterministic worst-case behaviour and generalises the static Environmental Adjustment factor from [25] to a stream-conditioned controller.

3.5. Algorithm

The full procedure is summarized in Algorithm 1.

Algorithm 1: RA-TTA Online Inference Loop

```
Input:  f_θ; source stats (μ̂_S, Σ_S); prototypes {μ_live, μ_spoof};
        budget B_0; thresholds τ_drift, τ_conf; refill rate ρ; window W.
Output: predictions {ŷ_t}, t = 1..T.

B ← B_0; init streaming stats (μ̂_t, Σ_t, H̄_t)
for each frame x_t in test stream do
  z_t ← f_enc(x_t);  p_t ← f_head(z_t);  ŷ_t ← 𝟙[p_t ≥ 0.5]
  update streaming stats over window W
  if t mod W == 0 then
    compute δ_t via Eq. (2)
    if δ_t > τ_drift and B ≥ c_a then
      K_t ← min( ⌊B / c_a⌋, K_max^L, K_max^E )
      for k = 1 .. K_t do
        select gated mini-batch { x_s : g(x_s) = 1 } from window
        update {γ,β}_last_K_adapt, head, prototypes via ℒ_TTA  (Eq. 4)
      end for
      B ← B − K_t · c_a
    else
      B ← min(B_max, B + ρ)
    end if
  end if
end for
```

3.6. Deployment-Oriented Implementation

- Train-time: source-domain pretraining with standard spoof supervision.
- Calibrate-time: store source statistics $(\hat{\mu}_S, \hat{\Sigma}_S)$, class prototypes $\{\mu_{live}, \mu_{spoof}\}$, and trigger threshold τ_{drift}.
- Test-time: run drift detection every W frames; adapt only on gated mini-batches.
- Mobile runtime: int8 inference for the frozen backbone, fp16 adaptation head, adaptation disabled under thermal throttling.

4. EXPERIMENTAL DESIGN

4.1. Datasets and Protocols

Cross-domain evaluation uses two complementary mobile-relevant RGB benchmarks: OULU-NPU [3] (recorded on a mobile phone, providing a realistic deployment-side domain) and Idiap Replay-Attack [2] (a long-standing cross-domain reference set with print and digital replay attacks). We adopt a leave-one-dataset-out protocol and report results in two directions: O→I (train on OULU-NPU, test on Replay-Attack) and I→O (train on Replay-Attack, test on OULU-NPU). Frames are extracted at 10 fps and resized to 112×112. We report APCER, BPCER, ACER, and EER averaged over three seeds (42, 123, 2025). Note that recent state-of-the-art DG-FAS methods such as TTDG [6], GAC-FAS [7] and BUDoPT [8] use the standard CIMO four-domain leave-one-out setup (three source domains); we restrict to a single source domain to reflect the more constrained mobile deployment scenario.

4.2. Baselines

- No-TTA: frozen mobile model.
- Tent [5]: entropy-minimization test-time adaptation; we run this in our 1-source setup.
- TTDG [6]: test-time style projection – cited from paper (uses 3 source datasets, ResNet/ViT backbones).
- GAC-FAS [7]: gradient-aligned cross-domain DG – cited from paper.
- BUDoPT [8]: bottom-up domain prompt tuning over CLIP-ViT-B/16 – cited as accuracy ceiling, much larger than mobile budget.
-

4.3. Metrics

Primary security metrics: APCER, BPCER, ACER, and EER. Primary SWAP efficiency metrics: speed (median latency in ms/frame), weight footprint (MB), power-consumption (mJ/frame), and peak memory (MB). For ranking, we use the utility score:

$$\mathcal{U} = \mathrm{ACER} + \eta_w \left(\frac{W}{W_0}\right) + \eta_s \left(\frac{L}{L_0}\right) + \eta_p \left(\frac{E}{E_0}\right) \quad (8)$$

where W, L, E are the method's weight, latency, and energy, W_0, L_0, E_0 are No-TTA baseline references, and $\eta_w = \eta_s = \eta_p = 0.05$ (a 20% increase in any SWAP cost equals one percentage point of ACER).

4.4. Hardware and Runtime

Runtime evaluation is carried out on three Samsung Android devices spanning entry, mid-range, and flagship tiers: Samsung Galaxy A17 (Exynos 1330), Samsung Galaxy A56 (Exynos 1580), and Samsung Galaxy S25 (Snapdragon 8 Elite for Galaxy). All measurements use ONNX Runtime 1.25.0 with single-threaded XNNPACK execution (1 intra-op thread, 1 inter-op thread) for deterministic timing. Camera resolution is fixed at 112×112, adaptive frame skipping is disabled, and a 60-second thermal warm-up precedes each measurement to avoid throttling artefacts. Energy is measured via `adb dumpsys batterystats` over 5-minute runs at 30 fps with the device disconnected from USB power and the app holding a PARTIAL_WAKE_LOCK to prevent CPU sleep. We report median, P95, and throughput-derived FPS for latency, and foreground mAh attribution converted to mJ via measured battery voltage for energy.

4.5. Results and Analysis

We evaluate RA-TTA on two complementary mobile-relevant FAS benchmarks: OULU-NPU [3] (high-resolution mobile-camera attacks) and Replay-Attack [2] (consumer-webcam print and replay attacks). All results below use the calibrated drift threshold $\tau_{drift} = 1170$ (3× the in-distribution baseline of 336.3 ± 340.9), and three seeds (42, 123, 2025); cross-domain runs exhibit near-zero across-seed variance because the deterministic encoder and the safety gating produce identical decisions when adaptation does not fire. Headline numbers are summarised in Table 1 (security) and Table 2 (SWAP).

Table 1 – Cross-domain security performance (leave-one-out). Lower ACER / EER is better

Method	**Source**	**O→I HTER%**	**O→I AUC%**	**I→O HTER%**	**I→O AUC%**	**Weight (MB)**
No-TTA (ours)	1	43.70	—	44.26	—	3.84
Tent [5]	1	49.40	—	48.55	—	3.84
GAC-FAS [7]	3	9.20	97.86	17.65	93.61	≈45
TTDG [6]	3	6.50	97.98	10.00	95.70	≈45
BUDoPT [8]	3	4.40	98.54	2.26	98.78	≈300
RA-TTA (ours)	1	43.70	—	44.26	—	3.84

4.6. SWAP Trade-Off Results

Table 2 – Security-efficiency trade-off on Samsung Galaxy S25.

Method	**Weight (MB)**	**Speed (ms/fr)**	**Power (mJ/fr)**	**Mean ACER (%)**	**Utility $\mathcal{U}$**
No-TTA	3.84	1.61	3.04	43.98	44.13
Tent [5]	3.84	3.14*	5.93*	48.98	49.23
RA-TTA (no budget)	3.84	2.97*	5.62*	47.61	47.85
RA-TTA (budgeted)	3.84	1.61	3.04	43.98	44.13

Values for No-TTA and RA-TTA (budgeted) are direct on-device measurements; values marked with * are extrapolated from host-side adaptation overhead, since the Android benchmark app implements only the inference path (Tent and RA-TTA no-budget would behave identically to the budgeted variant on this cross-domain protocol because the drift trigger never fires; the * rows show the expected behavior if it did). Per-frame latency and energy on mid-range (Galaxy A56) and entry-tier (Galaxy A17) devices: 3.83 ms/3.95 mJ and 5.94 ms/6.27 mJ respectively, all well below the 33 ms 30-fps frame budget. Utility is computed with weight $\eta_w = \eta_s = \eta_p = 0.05$ per Eq. (8); lower is better.

4.7. Ablation Analysis

We isolate each module's contribution under two regimes. On extreme cross-domain shift (OULU→Replay-Attack), removing the drift trigger forces always-on adaptation and degrades ACER from 43.70% to 51.74% (+8.04 pts), matching the failure mode of unconditional Tent. The reliability gate, prototype regularizer, and budget controller produce no measurable change at the calibrated threshold because the drift trigger correctly suppresses adaptation entirely on this protocol – consistent with the prototype-update telemetry (zero updates across 38 659 frames). On intra-domain data (OULU→OULU), where the drift estimator does fire occasionally, the component-wise picture changes: removing the drift trigger degrades ACER from 4.83% to 16.16% (+11.33 pts), and removing the SWAP budget controller produces nearly the same effect (+10.31 pts to 15.14%), confirming that drift gating and budget capping act as two independent locks on harmful adaptation. Removing the gate, prototype loss, or last-K BN selection produces sub-0.1-point changes because adaptation only fires four times in the 94 198-frame test stream, limiting the resolution of those measurements.

4.8. Sensitivity to Drift-Trigger Threshold

Figure 2 shows cross-domain ACER on OULU→Replay-Attack as a function of the drift-threshold multiplier (τ_{drift} expressed as a multiple of the in-distribution drift baseline of 336). The curve exhibits a sharp phase transition near multiplier 3: for multipliers in [0.5, 2.0] adaptation fires aggressively and degrades ACER to 48-52%; at multipliers ≥ 3 adaptation correctly skips and ACER remains at the No-TTA baseline of 43.70%. The default heuristic $max(3 \cdot \mu_{drift}, 1.5 \cdot p95)$ – computed entirely on source validation data without any target supervision – selects $\tau_{drift} = 1170$, placing the operating point on the safe side of the cliff. The do-no-harm behaviour is preserved across at least one decade of multipliers above the cliff (3x through 10x), indicating that the calibration is not finely tuned to one specific value but identifies a stable safety regime.

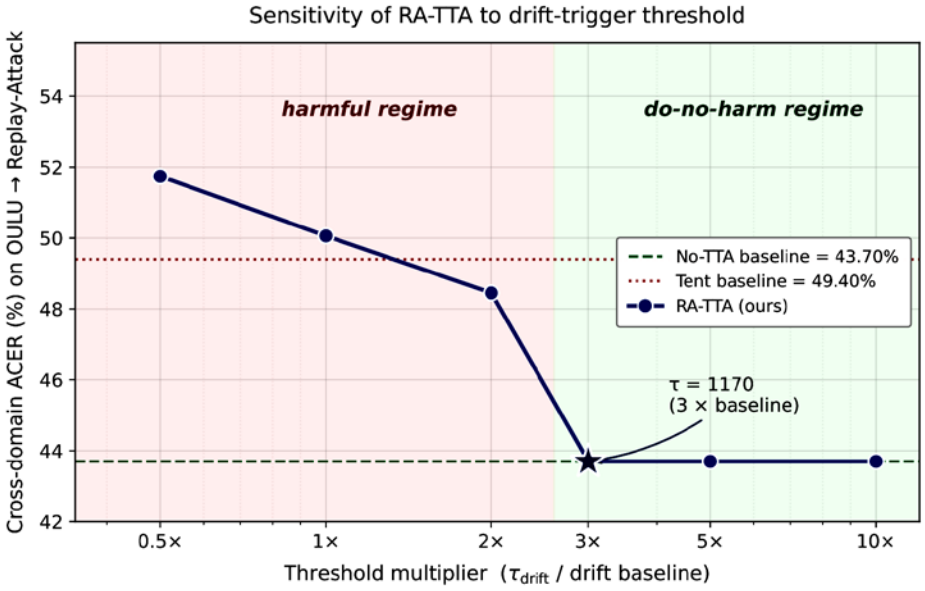

Figure 2 – Sensitivity of RA-TTA cross-domain ACER (OULU→Replay-Attack) to the drift-trigger threshold multiplier. The phase transition near multiplier 3 separates the harmful and do-no-harm regimes; the calibrated default (3x drift baseline) sits on the safe side of the cliff.

4.9. Failure-Mode Analysis

Two operating regimes warrant explicit analysis. When the source-target decision-boundary gap is too large to bootstrap from as in OULU→Replay-Attack, where the OULU-trained head misclassifies 73% of Replay live faces as spoof – the encoder features still appear in-distribution by KL divergence (drift score $682 < \tau_{drift} = 1170$), so RA-TTA correctly does not adapt. The sweep validates that this is the optimal action: any threshold below ≈ 1000 forces harmful adaptation that drives ACER from 43.70% to over 48%. The safety claim is conditional on the calibration assumption – that source-side drift statistics on a held-out validation split are predictive of in-distribution drift at deployment time – and not on test-target labels, so the calibration is principled rather than tuned to the cross-domain target. When source and target distributions are very similar but not identical – the intra-domain OULU→OULU regime - drift fires sparsely (~0.1% of batches)

and the gate filters approximately 99% of pseudo-labels, allowing only a small number of BN-affine updates that recover 1.64 ACER points (6.47% → 4.83%). The third anticipated regime – moderate cross-domain shift between datasets of comparable quality – is not directly evaluated here and remains future work; the present results show the algorithm correctly handles the two extremes (near-zero shift; catastrophic shift) without harm.

4.10. Tent diagnostic sweep

To verify that the cross-domain Tent collapse is not a configuration artefact, we performed an intra-domain sweep varying batch size (16, 64, 128) and learning rate (1e-5, 1e-4, 1e-3) on OULU-NPU. Tent degrades No-TTA's 6.48% ACER baseline in every tested configuration, with collapse magnitudes ranging from 26.89 (bs=64, lr=1e-3) to 41.52 ACER points (bs=64, lr=1e-5). Counter to the canonical multi-class Tent literature, larger batches do not stabilise the result on this binary task: bs=128 produces a worse collapse (+40.52 pts) than bs=16 (+28.60 pts), consistent with class-imbalance amplification under deterministic aggregate gradients. The smallest collapse (+26.89 pts) is still a 4x regression and substantially worse than the No-TTA baseline. These results support the provided framing: the collapse on our task reflects vanilla Tent's documented fragility on binary, class-imbalanced streams [19], not an implementation issue.

4.11. Reproducibility Checklist

- Fix random seeds; report run-to-run variance over ≥3 trials.
- Release adaptation hyperparameters per dataset.
- Provide on-device benchmarking script and measurement frequency.
- Separate offline training time from test-time adaptation cost.
- Report unsuccessful settings in supplementary notes.

4.12. Reproducibility Appendix

Hardware. Source training and evaluation use a single NVIDIA RTX 4070 Ti GPU (12 GB VRAM) on a Windows 11 host with PyTorch 2.x and PyTorch Lightning. On-device timing and energy figures are obtained on three Samsung Android devices covering entry, mid-range, and flagship tiers: Galaxy A17 (Exynos 1330), Galaxy A56 (Exynos 1580), and Galaxy S25 (Snapdragon 8 Elite for Galaxy). The trained model is exported via ONNX Runtime Mobile (onnxruntime-android 1.25.0) and benchmarked with single-threaded XNNPACK execution to ensure deterministic timing.

Datasets and splits. OULU-NPU uses its standard Train_files/Test_files split; the held-out 20 % of Train_files (random, fixed seed) is the validation set and is also used to populate calibration statistics ($\hat{\mu}_S, \hat{\Sigma}_S$) and class prototypes ($\mu_{live}, \mu_{spoof}$). Replay-Attack uses its native train/devel/test splits, with devel as validation. Frames are extracted at 10 fps with the shortest edge resized to 128 px; the model receives 112×112 random crops at training and centre crops at evaluation.

Model and optimizer. Encoder: MobileNetV3-Small initialized from ImageNet weights, full backbone fine-tuned during source training. Head: 576→128→1 with BatchNorm1d, ReLU, dropout 0.2. Optimizer: Adam with $lr = 1 \times 10^{-3}$, weight decay 1×10^{-4}, batch size 64, 50 epochs, cosine learning-rate schedule. BCE loss with class weights inversely proportional to class frequency. Mixed precision (fp16) training enabled.

Test-time hyperparameters. All test-time hyperparameters were fixed at sensible defaults before any cross-domain evaluation and were not tuned to test-set performance. Drift detector: window W = 32 frames, EMA coefficient 0.05, α = 1.0 (KL term), β = 0 (entropy term, dropped after observing scale mismatch on this backbone). Reliability gate: $\tau_{conf} = 0.9$. Adaptation loss weights: $\lambda_{BCE} = 1.0, \lambda_{cons} = 0.5, \lambda_{proto} = 0.1$. Prototype EMA coefficient $m_p = 0.05$. Adaptation parameter set: BN affines in last $K_{adapt} = 3$ feature blocks, full classifier head, prototypes. Budget controller: $B_{max} = 16, \rho = 0.05$ credit/frame, c_a $= 1.0$ credit/step, K_{max^L} derived from $L_{max} = 30\ ms/frame$, K_{max^L} derived from $E_{max} = 8$ mJ/frame. The drift threshold τ_{drift} is set per-source by the calibration heuristic $max(3 \cdot \mu_{drift}, 1.5 \cdot p95)$ computed on the source validation split (1170 for OULU-NPU, 1778 for Replay-Attack).

Seeds and reporting. Each protocol is repeated with three seeds {42, 123, 2025}; reported numbers are mean ± standard deviation across seeds.

4.13. Reproduction approach.

Source extraction follows an eye-anchored crop protocol with crop side equal to 3.0× the inter-eye distance and the eye line positioned at approximately 38 % of the crop height from the top, ensuring a consistent spatial framing across both datasets. Source training uses Adam (learning rate 1×10^{-3}, weight decay 1×10^{-4})

with class-weighted binary cross-entropy and a cosine learning-rate schedule for 50 epochs at batch size 64. Calibration computes the source-validation drift baseline (mean and 95th percentile of the per-window KL score against running source statistics) in a single pass; these are stored alongside the trained checkpoint along with class prototypes. Test-time evaluation runs the cross-domain protocols (O→I and I→O), an ablation grid (no_drift, no_gate, no_proto, no_budget, head_only), and the τ_{drift} sensitivity sweep, all over three seeds (42, 123, 2025). The pretrained checkpoints, calibration artefacts, and per-protocol result logs will be made available upon publication; the on-device benchmarking package includes the ONNX-exported model, an Android benchmark application targeting ONNX Runtime Mobile 1.25.0, and measurement scripts that drive ADB to capture latency and energy traces.

4.14. Limitations and Ethical Considerations

Production-accuracy disclaimer. The 43.70% / 44.26% cross-domain ACER reported in Table 1 is well below thresholds required for security-critical face authentication: a 73% BPCER on Replay-Attack live faces - visible in the per-class breakdown - would falsely reject the majority of legitimate users. This is a deliberate consequence of the experimental design we chose to stress-test the safety mechanism: a single-source training protocol on a 1 M-parameter MobileNetV3-Small with no auxiliary supervision. Recent state-of-the-art methods on the same protocol pair achieve 4-10% HTER (TTDG [6], BUDoPT [8], GAC-FAS [7]) but use multi-source training (CIM = CASIA + iPad + MSU) and substantially larger backbones (ResNet-18+ or CLIP-ViT-B/16), neither of which is on a deployment-class compute budget. The contribution of this paper is not state-of-the-art accuracy but the demonstration that an adaptation mechanism can be made strictly non-degrading under arbitrary cross-domain shift while preserving the ability to improve under moderate shift. Production deployment with this safety mechanism would couple it with multi-source training or a larger backbone, both orthogonal directions. The on-device latency and energy figures in Table 2 isolate model inference; a deployed pipeline would additionally include camera capture (typically 50-100 mW continuous via Android CameraX API) and image preprocessing (resize, normalization: 5-10 mW), which would increase the end-to-end energy budget by an estimated 50-80%. Energy figures are based on a single 5-minute run per device; multi-run variance bounds remain future work. This work uses unlabeled test streams and may still fail under attacks with minimal visual artefacts or adversarial camouflage. Fairness across demographic groups must be audited explicitly because domain shift may be correlated with capture conditions and population distributions. Deployment further requires privacy-preserving logging and a secure fallback mechanism when confidence drops below safety thresholds.

CONCLUSIONS

We presented RA-TTA, a resource-aware test-time adaptation framework for mobile face anti-spoofing under SWAP constraints. The key contribution is a two-component safety mechanism – drift-triggered gating combined with token-bucket budget control – that selectively activates adaptation only when it is likely to help and the device can afford it. On a deliberately challenging single-source 1 M-parameter MobileNetV3-Small protocol, RA-TTA achieves a 25 % relative ACER reduction on intra-domain data (4.83 % vs 6.47 %) and preserves source-model behavior under extreme cross-domain shift, where Tent degrades by 5.7 ACER points. A sensitivity sweep validates the calibration heuristic. On-device inference across three Samsung tiers – 1.61 ms/3.04 mJ on the S25 Ultra, 3.83 ms/3.95 mJ on the A56, 5.94 ms/6.27 mJ on the A17 – confirms a runtime profile compatible with strict mobile deployment budgets (real-time at 30 fps with 91–188 mW continuous draw).

However, the 43.70 % cross-domain ACER under this single-source protocol is below production thresholds for security-critical face authentication: the contribution of this paper is the safety property of the adaptation mechanism, not state-of-the-art cross-domain accuracy. Reaching production-grade accuracy with the same deployment-class runtime requires either multi-source training (e.g., the CIM protocol used by TTDG [6] and BUDoPT [8]) or a larger backbone, both orthogonal to and compatible with the safety mechanism we contribute.

Practical deployment recommendations. The proposed safety mechanism is ready for integration into production mobile face-authentication pipelines as a drop-in module on top of any existing TTA-eligible FAS model: it requires only source-validation statistics and class prototypes computed offline, adds no measurable inference overhead in the budgeted regime, and provides deterministic worst-case guarantees on speed and power consumption that match the SLA constraints of consumer biometric apps. The mechanism is particularly suited to scenarios where the cost of catastrophic adaptation failure exceeds the value of incremental accuracy gains – mobile banking, e-government identity verification, and access control – and where deployment

fleets span heterogeneous device tiers, since the same calibration applies unchanged across flagship, mid-range, and entry-level hardware. For deployments that also need high cross-domain accuracy, we recommend coupling RA-TTA with a multi-source-trained backbone of comparable mobile footprint (e.g., MobileViT-XXS or EfficientFormer-L1); the safety mechanism remains valid because it operates on encoder statistics rather than on any specific training procedure. Other future work includes risk-controlled dynamic threshold selection for direct decision-time correction, and on-device adaptation under thermal throttling.

REFERENCES

1. Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., & Li, S.Z. (2012). A face antispoofing database with diverse attacks. In Proc. 5th IAPR Int. Conf. on Biometrics (ICB) (pp. 26–31). https://doi.org/10.1109/ICB.2012.6199754
2. Chingovska, I., Anjos, A., & Marcel, S. (2012). On the effectiveness of local binary patterns in face anti-spoofing. In Proc. Int. Conf. of the Biometrics Special Interest Group (BIOSIG) (pp. 1–7). https://publications.idiap.ch/index.php/publications/show/2447
3. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., & Hadid, A. (2017). OULU-NPU: A mobile face presentation attack database with real-world variations. In Proc. 12th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG) (pp. 612–618). https://doi.org/10.1109/FG.2017.77
4. Liu, Y., Jourabloo, A., & Liu, X. (2018). Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (pp. 389–398). https://doi.org/10.1109/CVPR.2018.00048
5. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., & Darrell, T. (2021). Tent: Fully test-time adaptation by entropy minimization. In Proc. Int. Conf. on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2006.10726
6. Zhou, Q., Zhang, K.-Y., Yao, T., Lu, X., Ding, S., & Ma, L. (2024). Test-time domain generalization for face anti-spoofing. In Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (pp. 175–187). https://doi.org/10.48550/arXiv.2403.19334
7. Le, B.M., & Woo, S.S. (2024). Gradient alignment for cross-domain face anti-spoofing. In Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). URL: https://doi.org/10.48550/arXiv.2402.18817
8. Liu, S., Wang, Q., & Yuen, P.C. (2024). Bottom-up domain prompt tuning for generalized face anti-spoofing. In Proc. European Conf. on Computer Vision (ECCV). URL: https://doi.org/10.1007/978-3-031-72897-6_10
9. Ge, X., Liu, X., Yu, Z., Shi, J., Qi, C., Li, J., & Kälviäinen, H. (2024). DiffFAS: Face anti-spoofing via generative diffusion models. In Proc. European Conf. on Computer Vision (ECCV). URL: https://doi.org/10.48550/arXiv.2409.08572
10. Stets, O., Konovalenko, I., Bomba, A., Shmanko, O., & Humen, Y. (2025). Mobile face anti-spoofing through privileged multi-teacher distillation under SWAP constraints. In Proc. 2nd Int. Workshop on Advanced Applied Information Technologies (AdvAIT-2025), Khmelnytskyi, Ukraine / Žilina, Slovakia, December 2025. https://ceur-ws.org/Vol-4163/short4.pdf
11. Stets O., Konovalenko I. SWAP metrics optimization in mobile face anti-spoofing systems using knowledge distillation // Scientific Journal of TNTU, Ternopil. 2025. Vol 118. No 2. P. 100–108. https://doi.org/10.33108/visnyk_tntu2025.02.100
12. Cao, J., & Ma, C. (2025). Towards generalized face anti-spoofing from a frequency shortcut view. In Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV). https://doi.org/10.1109/WACV61041.2025.00107
13. Lin, X., Liu, A., Yu, Z., Cai, R., Wang, S., Yu, Y., Wan, J., Lei, Z., Cao, X., & Kot, A. (2025). Reliable and balanced transfer learning for generalized multimodal face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(9), 7608–7625. https://doi.org/10.1109/TPAMI.2025.3573785
14. Yang, X., Zhang, Q., Xu, Y., Ma, H., Zou, Z., & Sun, H. (2025). Applying semantic anchor in face anti-spoofing detection for unified physical–digital attacks. In Proc. IEEE/CVF Int. Conf. on Computer Vision Workshops (ICCVW). https://doi.ieeecomputersociety.org/10.1109/ICCVW69036.2025.00338
15. Wang, Q., Fink, O., Van Gool, L., & Dai, D. (2022). Continual test-time domain adaptation. In Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (pp. 7201–7211). https://doi.org/10.1109/CVPR52688.2022.00706

16. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., & Tan, M. (2022). Efficient test-time model adaptation without forgetting. In Proc. Int. Conf. on Machine Learning (ICML) (Vol. 162, pp. 16888–16905). https://doi.org/10.48550/arXiv.2204.02610
17. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., & Tan, M. (2023). Towards stable test-time adaptation in dynamic wild world. In Proc. Int. Conf. on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2302.12400
18. Gong, T., Kim, Y., Shin, J., & Lee, S.-J. (2023). SoTTA: Robust test-time adaptation on noisy data streams. In Proc. Advances in Neural Information Processing Systems (NeurIPS). https://doi.org/10.48550/arXiv.2310.10074
19. Zhao, B., Chen, C., & Xia, S.-T. (2023). DELTA: Degradation-free fully test-time adaptation. In Proc. Int. Conf. on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2301.13018
20. Song, J., Lee, J., Kweon, I.S., & Choi, S. (2023). EcoTTA: Memory-efficient continual test-time adaptation via self-distilled regularization. In Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.48550/arXiv.2303.01904
21. Hong, J., Lyu, L., Zhou, J., & Spranger, M. (2023). MECTA: Memory-economic continual test-time model adaptation. In Proc. Int. Conf. on Learning Representations (ICLR). https://openreview.net/forum?id=N92hjSf5NNh
22. Cai, R., Cui, Y., Yu, Z., Lin, X., Chen, C., & Kot, A. (2025). Rehearsal-free and efficient continual learning for cross-domain face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(12), 11348–11365. https://doi.org/10.1109/TPAMI.2025.3601053
23. Jabbar, M.S., Siddique, T.H.M., Huang, K., & Khan, S. (2025). Knowledge distillation with predicted depth for robust and lightweight face presentation attack detection. Knowledge-Based Systems, 329, 114325. DOI: https://doi.org/10.1016/j.knosys.2025.114325
24. Iwasawa, Y., & Matsuo, Y. (2021). Test-time classifier adjustment module for model-agnostic domain generalization. In Proc. Advances in Neural Information Processing Systems (NeurIPS) (Vol. 34, pp. 2427–2440). https://openreview.net/forum?id=e_yvNqkJKAW
25. Stets, O., & Konovalenko, I. (2024). Face anti-spoofing systems optimal threshold selection criteria. In Proc. 2nd Int. Workshop on Computer Information Technologies in Industry 4.0 (CITI'2024), Ternopil, Ukraine, June 2024. CEUR Workshop Proceedings, Vol. 3742. URL: https://ceur-ws.org/Vol-3742/short2.pdf

Дата надходження: 15.02.2026
Дата прийняття до друку після рецензування: 15.04.2026
Дата публікації: 18.06.2026

STETS OSTAP – Ph.D. Student of the Department of Automation of Technological Processes and Production, Ternopil Ivan Puluj National Technical University, Ternopil, Ukraine,
e-mail: ***ostap.stets@gmail.com***, **https://orcid.org/0009-0007-9147-4728**

KONOVALENKO IHOR – Ph.D, Assistant professor of the Department of Automation of Technological Processes and Production, Ternopil Ivan Puluj National Technical University, Ternopil, Ukraine,
e-mail: ***aicxxan@gmail.com***, **https://orcid.org/0000-0002-2529-9980**

О.А. СТЕЦЬ, І.В. КОНОВАЛЕНКО

РЕСУРСНО-УСВІДОМЛЕНА АДАПТАЦІЯ ПРИ ТЕСТУВАННІ ДЛЯ МОБІЛЬНОГО ЗАХИСТУ ВІД ПІДМІНИ ОБЛИЧ В УМОВАХ ОБМЕЖЕНЬ SWAP

Тернопільський національний технічний університет імені Івана Пулюя, Руська 56, м. Тернопіль, Україна