Mobile face anti-spoofing through privileged multi-teacher distillation under SWAP constraints*

Ostap Stets[1,*,†], Ihor Konovalenko[1,†], Andrii Bomba[2,†] and Oleh Shmanko[3,†]

[1] *Ternopil Ivan Puluj National Technical University, Ruska str., 56, 46001, Ternopil, Ukraine*
[2] *National University of Water and Environmental Engineering, 11 Soborna St., Rivne, 33028, Ukraine*
[3] *I. Horbachevsky Ternopil National Medical University, Maidan Voli St., 1, Ternopil, 46002, Ukraine*

Abstract

Face anti-spoofing (FAS) on mobile devices requires models that are not only accurate and robust across domains but also optimized under strict SWAP (Speed, Weight, Accuracy, Power) constraints. Current approaches often face a trade-off: strong generalization relies on additional modalities such as depth or rPPG signals, but mobile deployment can only afford lightweight RGB-only models. In this paper, we propose a privileged-information knowledge distillation (PI-KD) framework that enables multi-teacher supervision during training while keeping the deployment student efficient and mobile-friendly. Specifically, we outline how temporal teachers (rPPG-based) and geometric teachers (depth-based) can transfer complementary supervisory signals into a compact RGB-only student model. We discuss expected advantages for cross-dataset generalization, robustness to unseen attacks, and SWAP trade-offs, and we propose evaluation protocols that future empirical work should adopt. Our contribution lies in presenting a methodology that bridges privileged multimodal supervision with practical on-device constraints, opening a pathway toward more reliable and efficient mobile face anti-spoofing systems.

Keywords

face anti-spoofing (FAS), presentation attack detection (PAD), knowledge distillation (KD), CEUR-WS [1]

1. Introduction

Face anti-spoofing (FAS) is a critical safeguard in mobile biometric authentication, where decisions must be both reliable and resource-efficient. Production systems must optimize SWAP metrics: Speed (latency), Weight (model size/memory), Accuracy (PAD metrics such as APCER/BPCER/ACER/RIAPAR), and Power (energy consumption per inference). The ISO/IEC 30107-3 test methodology formalizes PAD error metrics and evaluation protocols, anchoring how operating points are compared in practice [1]. Deep learning methods have substantially advanced FAS, including pixel-wise supervision and domain-generalization strategies. Yet, models that generalize best, often rely on sensing or capacity that mobile devices cannot afford at inference (e.g., depth, rPPG, or heavy backbones). This tension in robustness versus deployability persists across recent surveys reviewing cross-dataset performance and multi-modal setups [2].

A principled way to bridge this gap is Learning Using Privileged Information (LUPI): during training, a "Teacher" may access auxiliary information not available at test time, while the deployed "Student" obeys the standard runtime contract (e.g., RGB-only). LUPI formalizes how such additional signals can shape decision boundaries and accelerate learning without altering the inference interface [3]. Within FAS, two privileged cues are especially complementary: depth (3D structure/recapture artifacts) and rPPG (physiological rhythms). Liu et al. showed that pixel-wise depth estimation and sequence-level rPPG supervision guide a network toward more

*AdvAIT-2025: 2nd International Workshop on Advanced Applied Information Technologies: AI & DSS, December 05, 2025, Khmelnytskyi, Ukraine, Zilina, Slovakia
[1,*] Corresponding author.
[†] These authors contributed equally.
ostap.stets@gmail.com (O. Stets); aicxxan@gmail.com (I. Konovalenko); abomba@ukr.net (A. Bomba); chmankoov@tdmu.edu.ua (O. Shmanko)
0009-0007-9147-4728 (O. Stets); 0000-0002-2529-9980 (I. Konovalenko); 0000-0001-5528-4192 (A. Bomba); 0000-0001-8029-9433 (O. Shmanko)

discriminative. generalizable liveness cues, improving intra-dataset and cross-dataset robustness – exactly the properties desired before compressing knowledge into a mobile-sized student [4].

To realize a deployable solution, we propose to adopt knowledge distillation as the compression mechanism – transferring the behavior of one or more teachers into a compact student that preserves accuracy while meeting mobile constraints [5]. In our privileged multi-teacher KD design, an rPPG-aware temporal teacher and a depth/landmark-aware geometric teacher supervise an RGB-only student; the student alone runs on-device, enabling better SWAP trade-offs without changing the inference inputs. This paper is a method proposal and discussion: we specify the PI-KD design choices and a SWAP-aware evaluation blueprint for future empirical validation.

2. Problem formulation

This work proposes a privileged-information, multi-teacher distillation scheme that trains on multi-modal supervision (available only during training) while deploying a single RGB-only student optimized under SWAP constraints. We formalize the task, models, objective, and constraints below.

2.1. Task and data

Let $x \in X_{RGB}$ denote an RGB video clip (or short sequence of frames) captured by a mobile camera, and $y \in \{0,1\}$ the PAD label (0: attack, 1: bona fide). During training (only), we assume access to privileged signals tied to the same clip:

- z^{rPPG} – temporal/physiological cues (e.g., rPPG traces or features)
- z^{geom} – geometric cues (e.g., depth/landmarks)

Training data thus form $D_{train} = (x_i, y_i, z_i^{rPPG}, z_i^{geom}, s_i)_{i=1}^{N}$, where s_i indicates the source domain (sensor, PAI type, capture condition). At deployment, only x is available, meaning that the target domain may be unseen. Performance is evaluated with PAD metrics (APCER, BPCER, ACER) as formalized in ISO/IEC 30107-3 [1].

2.2. Models

- Student $f_\theta : X_{RGB} \rightarrow [0,1]$ outputs a liveness score $p_\theta(y=1|x)$. The backbone is mobile-efficient (e.g., MobileNetV3/EfficientFormer/MobileViT) to enable SWAP-constrained deployment
- Temporal teacher T_ω is trained with/for rPPG-aware supervision on (x, z^{rPPG}), producing logits $t_\omega(x)$ or intermediate features $\phi_\omega(x)$
- A geometric teacher G_ψ is trained with/for depth/landmark supervision on (x, z^{geom}), producing logits $g_\psi(x)$ or features $\phi_\psi(x)$

This setting follows Learning Using Privileged Information (LUPI) – auxiliary signals guide learning but are absent at test time [3] – and uses knowledge distillation (KD) to transfer teacher behavior into the deployable student [5]. Prior PAD work shows that depth and rPPG provide complementary liveness cues that improve generalization [4].

2.3. Training objective

The student is trained with standard classification on hard labels with binary cross-entropy to establish a strong baseline decision surface. Let $\sigma(\cdot/T)$ denote softmax at temperature T. Then, classification loss can be described using formula 1:

$$L_{cls} = \frac{1}{N}\sum_{i=1}^{N} BCE(y_i, f_\theta(x_i)) \quad (1)$$

where x_i is the RGB input sample; $y_i \in \{1,0\}$ (1 = bona fide, 0 = attack); $f_\theta(x_i)$ is the student's predicted probability for bona fide; N is the number of training samples; BCE is Binary Cross Entropy.

We align the student's softened outputs with both temporal (rPPG) and geometric (depth/landmark) teachers using KL divergence; the temperature $T>0$ and weights α, β control the strength of each signal.

$$L_{logit} = \frac{1}{N}\sum_{i=1}^{N} (\alpha\, KL(\sigma(t_\omega(x_i)/T) \| \sigma(f_\theta(x_i)/T)) + \beta\, KL(\sigma(g_\psi(x_i)/T) \| \sigma(f_\theta(x_i)/T))) \quad (2)$$

where $t_\omega(x_i)$ and $g_\psi(x_i)$ are logits from the temporal (rPPG) and geometric (depth/landmark) teachers; C is the number of classes (for PAD, $C=2$); $\alpha, \beta \geq 0$ are distillation weights; $T>0$ is the temperature.

To further transfer inductive biases, we could match intermediate representations: student features may be projected and pulled toward each teacher's features with squared L_2 penalties weighted by γ, δ:

$$L_{feat} = \frac{1}{N}\sum_{i=1}^{N} (\gamma \left\| P_\theta \phi_\theta(x_i) - P_\omega \phi_\omega(x_i) \right\|_2^2 + \delta \left\| P_\theta \phi_\theta(x_i) - P_\psi \phi_\psi(x_i) \right\|_2^2) \quad (3)$$

where $\phi_\theta, \phi_\omega, \phi_\psi$ are student/teacher feature vectors; $P_\theta, P_\omega, P_\psi$ are linear projections to a shared feature space; $\|u\|_2^2$ denotes the squared L_2 norm; $\gamma, \delta \geq 0$ are feature-distillation weights.

The final objective sums the classification term, dual logit-distillation, and (optionally) feature-distillation; only the student is deployed at inference.

$$min_\theta\, L = L_{cls} + L_{logit} + L_{feat}\ with\ hyperparameters: \alpha, \beta, \gamma, \delta, T \quad (4)$$

This formulation supports domain-generalization by (1) leveraging complementary privileged cues during training [4], and (2) allowing standard regularizers (e.g., style/augmentation schedules) without altering the runtime model.

2.4. SWAP constraints and optimization goal

Let the measured deployment characteristics on a target device be:

- $L(\theta)$ – latency (ms/inference)
- $W(\theta)$ – model size or peak memory (MB)
- $E(\theta)$ – energy (mJ/inference)
- $Err(\theta, \tau)$ – PAD error at operating point τ (e.g., ACER computed per [1]).

Then we get the optimization view:

$$min_{\theta,\tau}\, Err(\theta,\tau) + \lambda_L L(\theta)/L_{max} + \lambda_W W(\theta)/W_{max} + \lambda_E E(\theta)/E_{max} \quad (5)$$

where $L_{max}, W_{max}, E_{max}$ encode the deployment constraints, which are device and app-specific; the operating point τ is the decision threshold which should be selected according to [1, 6] to meet application requirements (e.g., fixed BPCER); KD terms $\alpha, \beta, \gamma, \delta$ act as training-time levers that can shift the Pareto frontier toward lower error without increasing runtime cost, consistent with the LUPI/KD paradigm [3,5]

2.5. Evaluation targets (for future empirical work)

Although this paper is a method proposal, a future study validating the approach should report:

- Accuracy: APCER/BPCER/ACER or other metrics at specified τ per [1, 6], including cross-dataset and unseen-attack splits.
- Speed & Weight: on-device latency and model size of the student.
- Power: energy per inference of the student (mJ/inference).
- Ablations: effect of each teacher (set $\alpha=0$ or $\beta=0$); effect of feature-level vs. logit-level distillation.
- Pareto analysis: Accuracy over Latency and Accuracy over Energy curves to illustrate SWAP trade-offs.

3. Comparative analysis of known solutions

3.1. RGB-only classifiers vs. auxiliary or privileged supervision

Early deep FAS systems commonly treat spoof detection as binary RGB classification, which is attractive for mobile deployment but fragile under cross-dataset shift. Surveys consistently report that purely RGB, label-only training tends to overfit to capture artifacts and backgrounds, degrading robustness when the camera, illumination, or PAI species changes [2].

A major step forward was auxiliary/privileged supervision: Liu et al. [4] showed that training with pixel-wise depth and sequence-level rPPG targets guides the network toward liveness cues tied to 3D structure and physiological rhythms, boosting both intra-dataset and cross-dataset accuracy. The catch is that these signals are not available on consumer phones at inference time.

3.2. Multi-modal at training vs. deployment reality (LUPI)

The Learning Using Privileged Information (LUPI) paradigm formalizes exactly this situation: richer signals may be used during training to shape the student's decision boundary, while the deployed interface remains RGB-only [3]. LUPI provides the theoretical justification for separating training-time supervision from test-time inputs and for using teachers to accelerate and regularize learning.

3.3. Knowledge distillation for FAS

Knowledge distillation (KD) is the practical mechanism to compress teacher behavior into a small student with little or no runtime overhead. Classic KD uses temperature-softened targets to transfer "dark knowledge", improving the accuracy of compact models. Recent FAS work applies KD explicitly to efficiency: a head-aware Transformer KD compresses to ~5 MB while remaining competitive, combining logit-level and feature-level terms. However, existing KD for FAS typically relies on single-teacher RGB supervision, leaving rPPG and depth advantages on the table [4, 7].

3.4. Mobile backbones and SWAP considerations

For on-device use, students should be instantiated with mobile-efficient backbones. MobileNetV3 [8] is hardware-aware and widely used in production; EfficientFormer [10] demonstrates

transformer-level accuracy at MobileNet-class latency on iPhone; MobileViT [9] is a light hybrid that brings global context with modest cost. These neural network architectures form strong, fair baselines for SWAP-aware comparisons (latency/size/energy vs. accuracy).

3.5. Metrics and operating points

When comparing systems, ISO/IEC 30107-3 requires reporting APCER, BPCER, ACER, and RIAPAR at defined operating points [1]. Many efficiency papers under-report operating-point specifics and almost never report energy per inference, complicating appropriate SWAP comparisons across models and devices. A standards-aligned, SWAP-aware protocol is therefore essential for credible claims.

4. Suggested improvements

1. Dual-teacher, privileged KD rather than single-teacher RGB KD. Use an rPPG-aware temporal teacher and a depth and landmark-aware geometric teacher to transfer complementary liveness signals into an RGB-only student. This leverages the robustness gains of auxiliary supervision (depth/rPPG) while keeping mobile inference unchanged [3, 4].
2. SWAP-aware student selection and reporting. Ground the study in one of mobile-focused architecture based students (MobileNetV3, EfficientFormer, MobileViT) [8, 9, 10], and report accuracy and device-measured latency, size, energy alongside ISO 30107-3 metrics [1] at clearly stated operating points, closing the common reporting gap.
3. Logit and feature KD with minimal overhead. Follow evidence from FAS-specific KD by combining temperature-softened logit matching with one light feature-matching head (single mid or high-level tap), then ablate to show the smallest configuration that still moves the Pareto frontier [7].
4. LUPI-consistent training schedule. Warm up on hard labels, enable KD with moderate temperature, optionally add a small feature loss; this aligns with LUPI's goal of accelerating learning while preserving a simple RGB runtime interface [5].
5. Comparative baselines that reflect deployment reality. Compare against: RGB-only students trained with labels; RGB single-teacher KD (as in head-aware KD) [7]; and the proposed dual-teacher privileged KD – reported under identical, ISO-aligned protocols [1].

Net effect: relative to RGB-only or single-teacher KD baselines, the proposed PI-KD aims to shift the SWAP Pareto front, achieving lower accuracy metrics at fixed latency, size, and energy by distilling temporal (rPPG) and geometric (depth) knowledge into a single deployable student. This directly reconciles the robustness of auxiliary supervision with the constraints of mobile inference.

5. Results & discussion

This paper proposed a privileged-information, dual-teacher knowledge distillation (PI-KD) framework for mobile face anti-spoofing that explicitly targets SWAP constraints – Speed, Weight, Accuracy, and Power. The key idea is to learn from temporal (rPPG-aware) and geometric (depth and landmark-aware) teachers during training, while deploying only a compact RGB-only student on the device. We detailed the training objective (hard-label BCE and two logit-KD terms, with optional feature-KD), articulated a SWAP-aware optimization view, and outlined a standards-aligned evaluation blueprint (ISO/IEC 30107-3 operating points; cross-dataset and unseen-attack protocols; on-device latency, size, and energy with Pareto analysis).

If proved out experimentally, PI-KD should shift the Pareto frontier, delivering lower accuracy metrics at fixed latency, size, and energy, i.e., improved robustness without increasing deployment

cost. By decoupling rich training-time supervision from test-time simplicity, the approach reconciles laboratory gains with real-world mobile constraints.

There are some limitations, as the approach assumes access to privileged signals (or credible surrogates) for teacher training, and introduces teacher-pretraining cost. Domain mismatch between teachers and target environments, as well as privacy considerations for physiological and depth data, must be managed.

Immediate next steps include empirical validation on OULU-NPU, SiW-M, CASIA-FASD, and CelebA-Spoof datasets under unified SWAP reporting; SWAP-U extensions that integrate calibrated uncertainty/abstention for operating-point control; test-time domain handling for training-free normalization or light adapters compatible with on-device budgets; data-centric boosts (e.g., diffusion-generated hard negatives) to widen coverage of unseen PAIs; fairness audits across demographics and cosmetics; and standardized energy-measurement protocols across Android/iOS toolchains. Together, these directions aim to turn the proposed blueprint into a deployable, reproducible, SWAP-centric path for reliable mobile PAD.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] ISO/IEC JTC1 SC37 Biometrics. ISO/IEC 30107-3. Information Technology – Biometric presentation attack detection – Part 3: Testing and Reporting. International Organization for Standardization, 2023.

[2] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, G. Zhao: Deep Learning for Face Anti-Spoofing: A Survey. URL: https://doi.org/10.48550/arXiv.2106.14948.

[3] V. Vapnik, R. Izmailov: Learning Using Privileged Information: Similarity Control and Knowledge Transfer, Journal of Machine Learning Research 16, Columbia University, 2015. URL: https://jmlr.org/papers/volume16/vapnik15b/vapnik15b.pdf.

[4] Y. Liu, A. Jourabloo, X. Liu: Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision, Department of Computer Science and Engineering Michigan State University, East Lansing MI 48824. URL: https://openaccess.thecvf.com/content_cvpr_2018/papers/Liu_Learning_Deep_Models_CVPR_2018_paper.pdf.

[5] G. Hinton, O. Vinyals, J. Dean: Distilling the Knowledge in a Neural Network. URL: https://doi.org/10.48550/arXiv.1503.02531.

[6] O. Stets, I. Konovalenko, T. Gancarczyk, A. Mykytyshyn: Face anti-spoofing systems optimal threshold selection criteria, CITI'2024: 2nd International Workshop on Computer Information Technologies in Industry 4.0, June 12–14, 2024, Ternopil, Ukraine. URL: https://ceur-ws.org/Vol-3742/short2.pdf.

[7] J. Zhang, Y. Zhang, F. Shao, X. Ma, S. Feng, Y. Wu, D. Zhou: Efficient face anti-spoofing via head-aware transformer based knowledge distillation with 5 MB model parameters, Applied Soft Computing, Volume 166, November 2024. URL: https://doi.org/10.1016/j.asoc.2024.112237.

[8] A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, H. Adam: Searching for MobileNetV3. URL: https://arxiv.org/abs/1905.02244.

[9] S.Mehta, M. Rastegari: MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. URL: https://arxiv.org/abs/2110.02178.

[10] Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, J. Ren: EfficientFormer: Vision Transformers at MobileNet Speed. URL: https://doi.org/10.48550/arXiv.2206.01191.