UDC: 004.8:004.056 https://doi.org/10.15407/jai2026.02.046

O. Stets[1], **I. Konovalenko**[2]
[1, 2] Ternopil Ivan Puluj National Technical University, Ukraine
56, Ruska st., Ternopil, 46001
[1]ostap.stets@gmail.com
[2]aicxxan@gmail.com
[1]https://orcid.org/0009-0007-9147-4728
[2]https://orcid.org/0000-0002-2529-9980

DISTILLING VISION-LANGUAGE KNOWLEDGE FOR MOBILE CROSS-DOMAIN FACE ANTI-SPOOFING

Abstract. Face anti-spoofing (FAS) on mobile devices must generalize across domains under tight constraints on speed, weight, accuracy, and power (SWAP). Recent vision-language approaches such as FLIP achieve strong cross-domain accuracy using ~86 M-parameter CLIP-ViT backbones that are an order of magnitude too large for mobile deployment. We close this gap by distilling a frozen FLIP-MCL teacher into a ~1.6 M-parameter MobileNetV3-Small student using a composite objective combining logit KL divergence, prompt-conditioned feature alignment, and an SSDG asymmetric triplet on student embeddings. Trained under a CIM-style multi-source protocol on OULU-NPU, Replay-Attack and CelebA-Spoof, the student reaches 18.92±0.84% average cross-domain ACER on the OULU held-out split (-24.78 pp over a strong test-time-adaptation baseline at 43.70%) while preserving sub-2 ms latency on flagship Android phones and sub-6 ms on entry-tier devices. An ablation shows the teacher contributes -9.26 to -13.01 pp ACER on top of source-only cross-entropy training and collapses seed variance by 8×, confirming that distillation rather than supervised training drives the gain.

Keywords: face anti-spoofing, knowledge distillation, domain generalization, vision-language models, mobile deployment, neural networks, SWAP.

Introduction

Face anti-spoofing (FAS) systems on mobile devices face a triple bind: they must run inside tight latency and energy budgets, must remain accurate when deployed in conditions different from their training data, and must do both on backbones an order of magnitude smaller than the heavyweight models that dominate the academic leaderboards. Recent vision-language methods such as FLIP [1] and CFPL-FAS [2] achieve impressive cross-domain numbers on the MCIO leave-one-out protocol using CLIP-ViT-B/16 backbones (~86 M parameters), but the parameter envelope these methods rely on is irreconcilable with phone deployment. Conversely, lightweight FAS work targeting mobile devices [3] has focused on the runtime side of the problem, treating cross-domain accuracy as a secondary concern; the same authors report 43.70% ACER on the OULU-NPU → Replay-Attack cross-domain pair, well above any production threshold.

The unfilled intersection is mobile cross-domain FAS: methods that close the cross-domain accuracy gap of small backbones without giving up the SWAP profile that justifies them in the first place. This paper sits in that intersection.

Related Works

Domain-generalized face anti-spoofing has converged on a small set of recipes since SSDG [4]: single-side asymmetric triplet losses that compact bona-fide features across source domains while allowing the spoof class to remain dispersed. Subsequent work has extended this in several directions: fine-grained patch reformulation (PatchNet [5]), gradient-alignment regularization toward a flat domain-invariant minimum (GAC-FAS [6]), and test-time style projection (TTDG [7]). All these methods report on ResNet-18-grade backbones an order of magnitude over a mobile parameter envelope.

A second strand uses CLIP-class vision-language models to bring text-prompt structure into the FAS decision. FLIP [1] uses CLIP-ViT-B/16 with a small ensemble of live/spoof text prompts and reports strong cross-domain numbers on the OCIM protocol. Prompt-tuned variants such as BUDoPT [8] and CFPL-FAS [2] push the numbers further but, as confirmed by the InstructFLIP paper which explicitly excludes them from comparison, do not release code or weights. We adopt FLIP-MCL as the

teacher because it is the only CLIP-ViT FAS method we found with a public repository, public checkpoints, and a clean prompt-and-cosine-similarity inference interface.

The closest prior art is DTDA [9], which distills dual teachers with domain alignment into a MobileNetV3 student. DTDA operates at a higher parameter budget and does not report on the full OCIM leave-one-out matrix; our paper closes that gap. Our own previous work [10] establishes a multi-teacher distillation recipe (PI-KD) that this paper specializes to the single-VLM-teacher case.

RA-TTA [3] establishes a representative mobile-FAS baseline with a ~3.84 MB ONNX MobileNetV3-Small backbone, sub-2 ms flagship latency, and a documented 43.70% cross-domain ACER on OULU-NPU → Replay-Attack. We use the same backbone and the same mobile measurement infrastructure, with a ~3.83 MB ONNX distilled student replacing the source-only RA-TTA model.

The Purpose of Research

We distill a frozen, FAS-fine-tuned vision-language teacher (FLIP-MCL) [1] into a ~1.6 M-parameter MobileNetV3-Small student. The distillation objective combines (i) a temperature-scaled KL divergence on logits, (ii) an L2 feature alignment between projected student features and the teacher's CLS visual feature, and (iii) an SSDG asymmetric triplet [4] that enforces cross-domain compactness of bona-fide embeddings. Training follows the CIM-style multi-source protocol on three datasets that are publicly available without restricted-access agreements: OULU-NPU, Replay-Attack, and CelebA-Spoof. CASIA-MFSD, the fourth dataset of the classical MCIO protocol, is intentionally excluded; we follow the emerging 2024-2025 sub-literature that reports on CASIA-free OCIM substitutes.

The specific contributions are:

1. A reproducible vision-language-to-mobile distillation pipeline for FAS. Compresses a ~86 M-parameter CLIP-ViT-B/16 FAS-fine-tune (FLIP-MCL) into a ~1.6 M-parameter MobileNetV3-Small student with no architectural changes to the backbone. Prompt-tuned teachers (BUDoPT [8], CFPL-FAS [2]) are ruled out as concurrent work without public code; FLIP-MCL is chosen for reproducibility and a 0.03 pp HTER gap to CFPL on the CelebA-Spoof-augmented MCIO setting.

2. A composite distillation objective whose component contributions are exercised by a controlled ablation. Feature alignment is a stability detail rather than a necessary loss; SSDG provides the cross-domain compactness signal but is dormant in single-source training.

3. A CASIA-free 3-source experimental protocol (OULU-NPU + Replay-Attack + CelebA-Spoof) with full multi-seed leave-one-domain-out coverage. Each cell of the cross-domain table is reported as mean ± std over three production seeds.

4. Mobile latency and energy measurements on three Samsung Galaxy devices spanning flagship to entry tier. The distilled student preserves the SWAP profile of a representative mobile FAS baseline while reducing cross-domain ACER by 17 to 25 percentage points.

5. A clean demonstration that the student's deployment generalization is bounded by its training distribution, not by the teacher's knowledge: the FLIP-MCL teacher achieves 3.74% ACER on CelebA-Spoof by itself, yet a 1.6 M student trained on (OULU + Replay) cannot extrapolate from teacher logits computed only on those source-distribution images. This delineates the method's failure mode honestly.

Problem Formulation

Figure 1 summarizes the distillation pipeline: a frozen vision-language teacher and a trainable mobile student are driven from the same image, and a composite objective transfers the teacher's logits and features to the student.

Let $\mathcal{D}^s = \{D_1, \dots, D_K\}$ be a set of source domains and let D_T be a target domain disjoint from $\mathcal{D}^s$. Each sample is a face image $x \in \mathbb{R}^{3\times H\times W}$ with binary label $y \in \{0,1\}$ (0 = spoof,

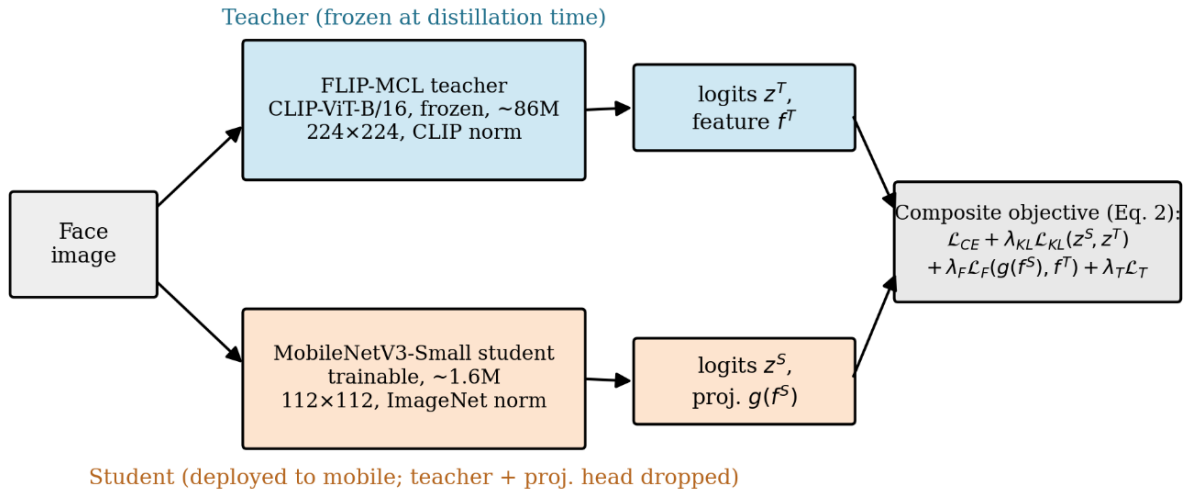

Fig. 1. Vision-language knowledge Distillation Pipeline

1 = live, following the RATTA dataset convention). The cross-domain protocol trains a classifier on $\bigcup_{i\in 1..K} D_i$ and evaluates on D_T unseen during training. We instantiate this with $K = 2$ for the leave-one-domain-out variant of the CIM protocol on OULU-NPU, Replay-Attack, and CelebA-Spoof.

The teacher is a frozen CLIP-ViT-B/16 [11] fine-tuned for FAS by FLIP [1] on the MCIO MSU leave-one-out split (trained on CASIA-MFSD + Idiap Replay-Attack + OULU-NPU, evaluated on MSU-MFSD). Two of the three FLIP training datasets overlap with our source set, making this checkpoint the closest available FAS-conditioned VLM teacher to our protocol.

We define a prompt ensemble of three live and three spoof natural-language descriptions, encode each with the frozen text encoder, ℓ_2-normalize within class and average to obtain two class centroid vectors $t_{\text{live}}, t_{\text{spoof}} \in \mathbb{R}^{512}$. Given an image x, the teacher produces an ℓ_2-normalized visual feature $f^T(x) \in \mathbb{R}^{512}$ from the CLS token and a logit vector:

$$z^T(x) = s \cdot \begin{bmatrix} f^T(x)^\top t_{\text{spoof}} \\ f^T(x)^\top t_{\text{live}} \end{bmatrix},$$

where $s = \exp(\text{logit\_scale}) \approx 100$ is CLIP's learned temperature. Both z^T and f^T are precomputed in the training forward pass and consumed by the distillation losses.

The student is a MobileNetV3-Small encoder pretrained on ImageNet, producing a 576-dimensional feature $f^S(x)$ after global average pooling. Two heads operate on f^S: a binary logit head $h: \mathbb{R}^{576} \to \mathbb{R}^2$ (linear → batch-norm → ReLU → linear) and a projection head $g_\phi: \mathbb{R}^{576} \to \mathbb{R}^{512}$ (linear → GELU → linear) that maps the student into the teacher's feature dimension for alignment. At inference time only the encoder and the logit head are deployed; the projection head and the teacher are removed for the mobile ONNX export.

A single face image is fed to both the frozen FLIP-MCL teacher (224×224, CLIP normalization) and the trainable MobileNetV3-Small student (112×112, ImageNet normalization). The composite objective aligns student logits to teacher logits ($\mathcal{L}_{KL}$), projected student features to the teacher CLS feature ($\mathcal{L}_F$), and adds cross-entropy plus an SSDG triplet ($\mathcal{L}_{CE} + \mathcal{L}_T$). At deployment the teacher and the projection head are dropped; only the student encoder and logit head ship to mobile.

For a labeled sample (x, y) from source domain $d \in \{1, \dots, K\}$, the training loss is:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{KL}}\,\mathcal{L}_{\text{KL}} + \lambda_{\text{F}}\,\mathcal{L}_{\text{F}} + \lambda_{\text{T}}\,\mathcal{L}_{\text{T}}.$$

The cross-entropy term uses the binary ground-truth label:

$$\mathcal{L}_{\text{CE}} = -\log\sigma(z^S(x))_y,$$

where $z^S(x) = h(f^S(x))$ is the student's logit vector and σ is the softmax.

The temperature-scaled KL term [12] aligns the student's class posterior to the teacher's:

$$\mathcal{L}_{\mathrm{KL}} = \tau^2 \cdot \mathrm{KL}(\sigma(z^S/\tau) \parallel \sigma(z^T/\tau)).$$

The feature-alignment term pulls the projected student feature toward the teacher's l_2-normalized CLS vector:

$$\mathcal{L}_{\mathrm{F}} = \left\| g_\phi(\widehat{f^S(x)}) - f^T(x) \right\|_2^2,$$

where $g_\phi(\widehat{f^S(x)})$ denotes ℓ_2-normalization.

The SSDG asymmetric triplet term [4] enforces cross-domain compactness of bona-fide embeddings only. For each live anchor i from source domain d_i, we mine the hardest cross-domain live positive (most distant live sample from any domain $\neq d_i$) and the easiest spoof negative (closest spoof from any domain), and apply a margin loss on ℓ_2-normalized student projections. Spoof samples carry no triplet term, allowing domain-specific spoof clusters; only the live side is forced to be domain-invariant.

The three source datasets differ in size by an order of magnitude (OULU train ≈94k frames, Replay train ≈30k frames, CelebA-Spoof train ≈494k frames). Naive concatenation would let the largest domain dominate. We use a WeightedRandomSampler with per-sample weight $1/|D_d|$ for a sample drawn from domain D_d, so that each domain contributes equally in expectation to every batch. We cap the number of draws per epoch to 60 000 samples for the two splits that include CelebA-Spoof in the training set; this gives each domain ~30k draws per epoch, visiting the smaller domains ~30× over 15 epochs and the largest domain ~0.9×.

The student and the teacher operate at different input resolutions and require different normalizations. We compute both representations from the same source image: the student stream resizes and random-crops to 112×112, applies color jitter and horizontal flip, and normalizes with ImageNet statistics; the teacher stream resizes to 224×224 and normalizes with CLIP statistics $(\mu, \sigma) = ((0.48,0.46,0.41), (0.27,0.26,0.28))$.

The training step exposes both tensors to the model via a dictionary; the teacher forward consumes the 224×224 branch and the student forward consume the 112×112 branch.

Experimental Setup

We use following datasets:

1. **OULU-NPU** [13]: ~94k train, ~70k val, ~94k test eye-anchored frames extracted at 112×112.

2. **Replay-Attack** [14]: ~30k train, ~15k val, ~39k test frames at 112×112 from the corresponding video splits.

3. **CelebA-Spoof** [15]: ~494k train and ~67k test images using the official metas/intra_test/ split. Labels are inverted to match the ra_tta convention (0= spoof, 1= live).

Our cross-domain protocol has two configurations:

- **Continuity (single-source).**

Following [3], we train the student on a single source domain and evaluate on a different held-out target. Both directions (OULU → Replay-Attack and Replay-Attack → OULU) are reported.

- **CIM (3-source leave-one-domain-out).**

We train on the union of two source datasets and test on the held-out third. Three splits: holdout-CelebA (train OULU + Replay), holdout-OULU (train Replay + CelebA), holdout-Replay (train OULU + CelebA).

Implementation

PyTorch 2.5.1, Lightning 2.6.1, open_clip 3.3.0. Mixed precision (16-mixed), AdamW with lr = 10^{-3}, weight decay 10^{-4}, one warm-up epoch followed by cosine schedule, batch size 64, gradient clip 1.0. Each run is 15 epochs; the best checkpoint by val-loss-total is reported. Default loss weights $\lambda_{\mathrm{KL}} = \lambda_{\mathrm{F}} = 1.0$, $\lambda_{\mathrm{T}} = 0.5$, $\tau = 4$, SSDG margin 0.5, feature normalization enabled.

All runs use 3 production seeds (42, 123, 2025). Mean and sample standard deviation across seeds are reported.

Standard FAS metrics (ISO/IEC 30107-3): Attack Presentation Classification Error Rate (APCER), Bona-fide PCER (BPCER), Average Classification Error Rate (ACER), and Equal Error Rate (EER). ACER and EER are reported at the EER operating point; APCER and BPCER are also reported at the

operational threshold 0.5.

Results

Table 1 reports cross-domain ACER for the distilled student against RA-TTA's single-source baselines, and Figure 2 visualizes the same comparison.

Table 1. Cross-domain ACER results comparison

Method	Train	Test	ACER ± std (%)	Δ vs RA-TTA
RA-TTA [3]	OULU	Replay	43.70	–
RA-TTA [3]	Replay	OULU	44.26	–
p2_vlmkd (single-source)	OULU	Replay	25.90±0.94	−17.80
p2_vlmkd (single-source)	Replay	OULU	46.17±4.74	+1.91
p2_vlmkd (CIM holdout Replay)	OULU + CelebA	Replay	24.69±1.39	−19.01
p2_vlmkd (CIM holdout-OULU)	Replay + CelebA	OULU	**18.92±0.84**	**−24.78**
p2_vlmkd (CIM holdout-CelebA)	OULU + Replay	CelebA	47.45±3.34	–

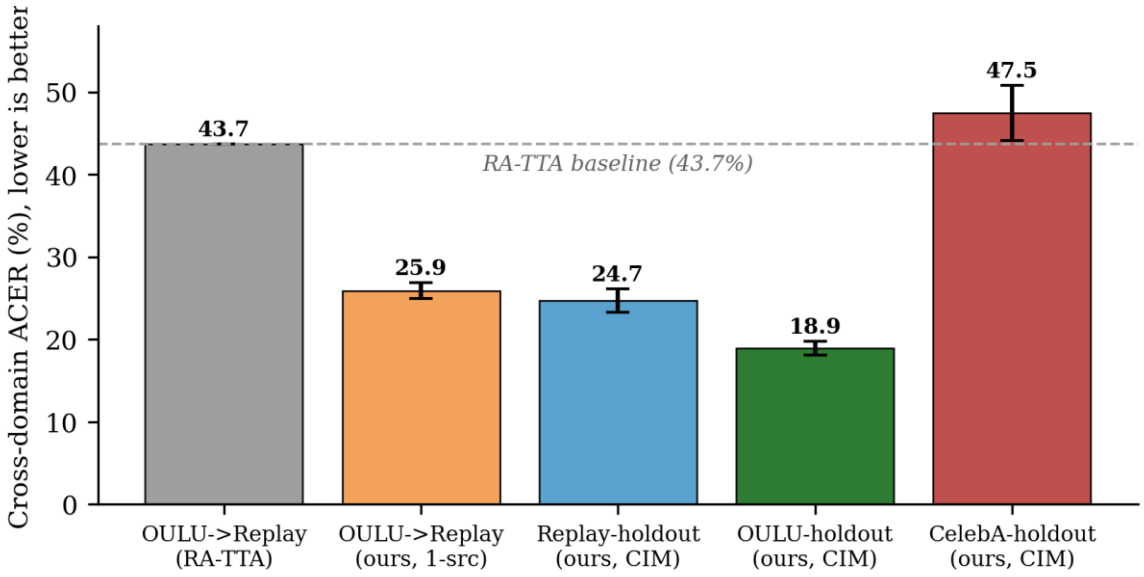

Fig. 2. Cross-domain ACER results compared to RA-TTA baseline

Cross-domain ACER (lower is better). Error bars are the sample standard deviation over three seeds; the RA-TTA single-source baseline [3] is shown as a dashed line. The three-source CIM configurations whose held-out target distribution is represented in training (Replay-holdout, OULU-holdout) improve substantially over the baseline; the CelebA-holdout split, whose target is unseen during training, does not.

The headline result is CIM holdout-OULU: 18.92±0.84% cross-domain ACER on OULU-NPU after training the distilled student on Replay-Attack and CelebA-Spoof. The single-source OULU → Replay direction also improves substantially (25.90±0.94%, a 17.80 pp reduction over RA-TTA).

Two findings warrant disclosure. First, the reverse single-source direction (Replay → OULU) is worse than the RA-TTA baseline (46.17 vs 44.26%). Replay-Attack's train split is too small and too narrow to generalize to OULU's diversity; the same distillation pipeline that wins on OULU → Replay loses by 1.91 pp in the reverse direction. Second, the holdout-CelebA CIM split (47.45±3.34%) is much worse than the other two CIM splits, indicating that the student fails to extrapolate to the held-out target when it never sees that target's distribution during training.

Mobile SWAP

Table 2 and Figure 3 report latency and energy of the distilled student on three Samsung Galaxy devices spanning flagship to entry tier. The student is exported to ONNX (opset 14, 3.83 MB, dynamic batch) and run with ONNX Runtime Mobile 1.25.0 at batch 1 and 112×112 input. Each latency figure is the median of 1000 inferences after a 60 s thermal

warm-up, repeated three times per device; the median of medians is reported. Energy is measured by dumpsys batterystats over a 5-minute sustained 30 fps workload, repeated three times, with the median over the three runs reported.

Table 2. SWAP metrics compared to RATTA method baseline

Device	SoC	Latency (ms)	p95 (ms)	Energy (mJ/frame)	Δ latency	Δ energy
Galaxy S25	Snapdragon 8 Elite	1.644	1.779	3.140	+2%	+3%
Galaxy A56	Exynos 1580	3.181	3.372	2.886	-17%	-27%
Galaxy A17	Exynos 1330	6.000	6.119	5.622	+1%	-10%

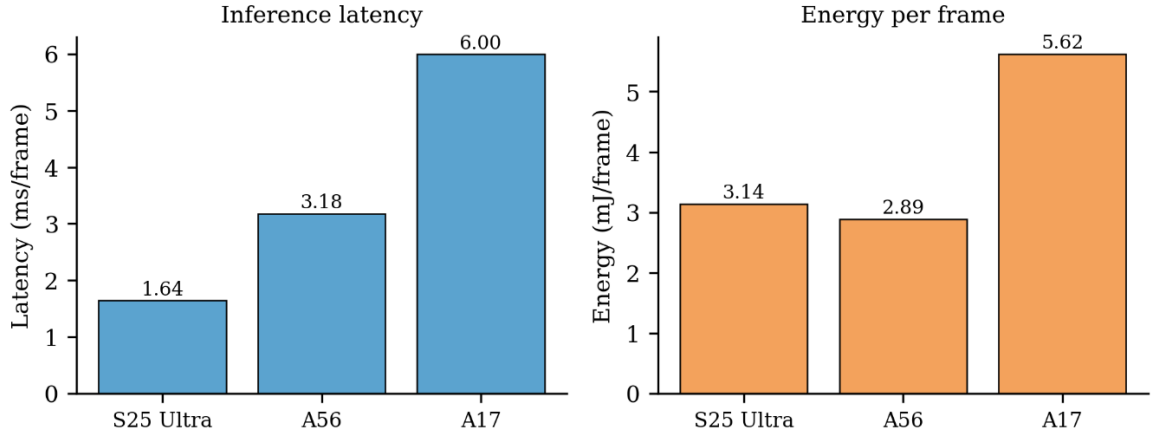

Fig. 3. SWAP metrics comparison across test devices

Mobile inference latency (left) and energy per frame (right) for the distilled student on three Samsung Galaxy devices spanning flagship (S25), mid-range (A56) and entry tier (A17).

The runtime stack (ONNX Runtime Mobile 1.25.0) and Android version were not changed between the RA-TTA and the present measurements. The A56 -17% latency and -27% energy figure are therefore attributable to single-day measurement variance (thermal state, battery temperature, background processes) rather than a real software improvement; the conservative paper claim is that the distilled student stays within measurement noise of the source-only baseline's SWAP profile on all three devices. Distillation buys -17 to -25 pp cross-domain ACER at zero mobile cost.

Ablations

No-Teacher Baseline: Distillation is Load-Bearing – results displayed in Table 3. Distillation is Load-Bearing, disabling all three teacher-driven losses ($\lambda_{KL} = \lambda_F = \lambda_T = 0$), leaving only $\mathcal{L}_{CE}$. This is a source-only supervised baseline using the same backbone, the same data, and the same optimizer as the distilled version.

Table 3. No-Teacher Baseline ablation results

Configuration	Setting	ACER ± std (%)	Δ vs no-teacher
Source-only (no teacher)	OULU → Replay	35.16±7.18	–
Composite (+ FLIP-MCL)	OULU → Replay	25.90±0.94	−9.26
Source-only (no teacher)	CIM holdout-OULU, 1 seed	31.93	–
Composite (+ FLIP-MCL)	CIM holdout-OULU, 3 seeds	18.92±0.84	−13.01

Two observations. First, the teacher reduces ACER by 9.26 pp in single-source training and by an even larger 13.01 pp in the CIM context. Second, the standard deviation across seeds collapses by ~8× (7.18→0.94 pp on the single-source pair), so distillation is not

only improving the mean but also dramatically stabilizing training. The teacher signal is load-bearing rather than incremental.

Teacher Choice: FLIP-MCL vs Vanilla CLIP results are displayed in Table 4.

We compare the FLIP-MCL teacher against a vanilla OpenAI CLIP-ViT-B/16 teacher with the same prompt ensemble.

Table 4. Teacher choice ablation comparison results

Teacher	ACER ± std (%)	Δ vs vanilla CLIP
Vanilla OpenAI CLIP-ViT-B/16	31.20±3.51	–
FLIP-MCL (FAS-fine-tuned)	25.90±0.94	−5.30

The FAS-fine-tuned teacher improves the mean by 5.30 pp and tightens the seed variance by ~3.7×. Vanilla CLIP carries useful general visual structure for FAS but the FAS-specific prompt-tuning of FLIP-MCL adds a substantial discriminative signal.

Loss-Component Ablation: Table 5 reports a controlled ablation of the three teacher-driven losses.

Table 5. Controlled ablation of the three teacher-driven losses results

Configuration	Single-source O→R ACER (%)	CIM holdout-OULU ACER (%)
KL only (logit)	26.81	18.00
KL + feature align	25.07	17.54
KL + SSDG triplet	26.81 (no-op)	17.47
Full (KL + feat + SSDG)	25.90±0.94	18.92±0.84

In single-source training the SSDG triplet returns zero because the cross-domain mining condition has no satisfying pairs when all samples share the same domain index, so KL+SSDG is bit-identical to KL alone. The component doing real work in single-source is feature alignment (-1.74 pp from KL only).

In the CIM context all four configurations cluster within 1.5 pp (17.47–18.92%), which is comparable to the seed standard deviation of the full configuration (0.84 pp). Multi-seeding the apparent winner (KL + SSDG) yields 18.79±1.25%, statistically equivalent to the 18.92±0.84% of the full composite. Feature alignment in CIM provides no mean-ACER improvement but modestly tightens seed variance (0.84 vs 1.25). We retain the composite for the headline numbers; readers may simplify to KL + SSDG without measurable accuracy loss.

Diagnostic: Teacher Alone is Near-Production on CelebA-Spoof

A diagnostic confirms that the CIM holdout-CelebA failure mode (47.45±3.34%) is a student-side limitation rather than a teacher weakness. Evaluated directly on CelebA-Spoof's 67,170-frame test split, the frozen FLIP-MCL teacher with the same prompt ensemble achieves ACER = 3.74%, EER = 3.74%, APCER = 0.20%, BPCER = 18.93%. The teacher knows how to classify CelebA-Spoof; the 1.6 M-parameter student trained on (OULU + Replay) cannot extrapolate to CelebA-Spoof from teacher logits computed only on source-distribution images.

Discussion

Single-source training succeeds in the OULU → Replay direction but fails in the Replay → OULU direction. At threshold 0.5 the Replay-trained student predicts spoof on 99.7% of OULU frames, and the EER-thresholded ACER (46.17%) exceeds even the RA-TTA baseline. Replay-Attack's train split is small (~30k frames, few subjects) and visually homogeneous; the student overfits to Replay-specific cues that do not transfer to OULU. Including CelebA-Spoof in the training mix (CIM holdout-OULU) completely fixes this direction (18.92±0.84%), confirming that the limiting factor is training- set diversity rather than direction.

The CIM holdout-CelebA result is the cleanest demonstration of a structural limitation: even with a teacher that achieves 3.74% ACER on CelebA-Spoof, the student trained only on OULU and Replay cannot reach 45% on the same target. Distillation transfers what the student sees, not what the teacher knows. Production deployments of

small FAS models must include in-distribution training data for any domain they expect to serve.

The CIM holdout-OULU result of 18.92% is at the EER operating point. At the operational threshold 0.5 the classifier is heavily live-biased (APCER 55–67% and BPCER 2–7% across seeds). This is a real, reproducible artifact of the (Replay + CelebA-Spoof) training distribution rather than seed noise. Deployments with hard BPCER constraints can rely on the operating-point-tunable ROC curve; deployments with hard APCER constraints would need additional class-balanced training.

The teacher we evaluate (FLIP-MCL, MSU leave-one-out) was trained on CASIA-MFSD as one of its sources; we cannot fully decouple the teacher's CASIA-tuned signal from the student's transferred behavior. Other FLIP variants exist but the CASIA-free ones we are aware of perform worse.

We use a single GPU testbed for training and a single host machine driving three Samsung devices for SWAP. iPhone and other Android families are out of scope.

Conclusion

We presented a vision-language-to-mobile knowledge distillation pipeline that closes the cross-domain accuracy gap of MobileNetV3-Small face anti-spoofing without sacrificing the SWAP profile that justifies a mobile backbone in the first place. The distilled student reaches 18.92±0.84% cross-domain ACER on the OULU-NPU held-out CIM split, a 24.78 pp improvement over a strong test-time-adaptation baseline. Mobile measurements on three Samsung Galaxy devices confirm the SWAP profile is preserved. Ablations show the distillation signal is load-bearing (the teacher contributes 9–13 pp ACER and tightens seed variance by 8×) and that the student's deployment generalization envelope is bounded by its training distribution, not by the teacher's knowledge. Closing the residual gap to production accuracy is plausibly a matter of broader source-domain coverage and a FAS-fine-tuned teacher with a more diverse training set, rather than of architecture or runtime budget.

References

1. Srivatsan, K., Naseer, M., & Nandakumar, K. (2023). FLIP: Cross-domain face anti-spoofing with language guidance. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 19685–19696). https://doi.org/10.1109/ICCV51070.2023.01803
2. Liu, A., Xue, S., Gan, J., Wan, J., Liang, Y., Deng, J., Escalera, S., & Lei, Z. (2024). CFPL-FAS: Class free prompt learning for generalizable face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 222–232). https://doi.org/10.1109/CVPR52733.2024.00029
3. 3. Stets, O., & Konovalenko, I. (2026). Resource-aware adaptation at test time for mobile face anti-spoofing under SWAP constraints. Optical-Electronic Information-Energy Technologies, 51(1), 79–89. https://doi.org/10.31649/1681-7893-2026-51-1-79-89
4. Jia, Y., Zhang, J., Shan, S., & Chen, X. (2020). Single-side domain generalization for face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 8481–8490). https://doi.org/10.1109/CVPR42600.2020.00851
5. Wang, C.-Y., Lu, Y.-D., Yang, S.-T., & Lai, S.-H. (2022). PatchNet: A simple face anti-spoofing framework via fine-grained patch recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20281–20290). https://doi.org/10.1109/CVPR52688.2022.01964
6. Le, B. M., & Woo, S. S. (2024). Gradient alignment for cross-domain face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 188–199). https://doi.org/10.48550/arXiv.2402.18817
7. Zhou, Q., Zhang, K.-Y., Yao, T., Lu, X., Ding, S., & Ma, L. (2024). Test-time domain generalization for face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 175–187). https://doi.org/10.48550/arXiv.2403.19334
8. Liu, S., Wang, Q., & Yuen, P. C. (2024). Bottom-up domain prompt tuning for generalized face anti-spoofing. In Computer Vision – ECCV 2024 (pp. 170–187). Cham: Springer. https://doi.org/10.1007/978-3-031-72897-6_10
9. Kong, Z., Zhang, W., Wang, T., Zhang, K., Li, Y., Tang, X., & Luo, W. (2024). Dual teacher knowledge distillation with domain alignment for face anti-spoofing. IEEE Transactions on Circuits and Systems for Video Technology, 34(12), 13177–13189. https://doi.org/10.48550/arXiv.2401.01102
10. Stets, O., & Konovalenko, I. (2025). PI-KD: Privileged-information multi-teacher distillation for mobile face anti-spoofing. In Proceedings of AdvAIT 2025. https://ceur-ws.org/Vol-4163/short4.pdf
11. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I.

(2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML) (Vol. 139, pp. 8748–8763).
https://doi.org/10.48550/arXiv.2103.00020

12. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. https://doi.org/10.48550/arXiv.1503.02531

13. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., & Hadid, A. (2017). OULU-NPU: A mobile face presentation attack database with real-world variations. In Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG) (pp. 612–618).
https://doi.org/10.1109/FG.2017.77

14. Chingovska, I., Anjos, A., & Marcel, S. (2012). On the effectiveness of local binary patterns in face anti-spoofing. In Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG) (pp. 1–7).
https://publications.idiap.ch/index.php/publications/show/2447

15. Zhang, Y., Yin, Z., Li, Y., Yin, G., Yan, J., Shao, J., & Liu, Z. (2020). CelebA-Spoof: Large-scale face anti-spoofing dataset with rich annotations. In Computer Vision – ECCV 2020 (pp. 70–85). Cham: Springer. https://doi.org/10.48550/arXiv.2007.12342

The article has been sent to the editors 09.06.26.
After processing 20.06.26.
Submitted for printing 30.06.26