

**The
Alan Turing
Institute**

**Robustness of AI-Based Face
Presentation Attack Detection in
Identity Proofing**

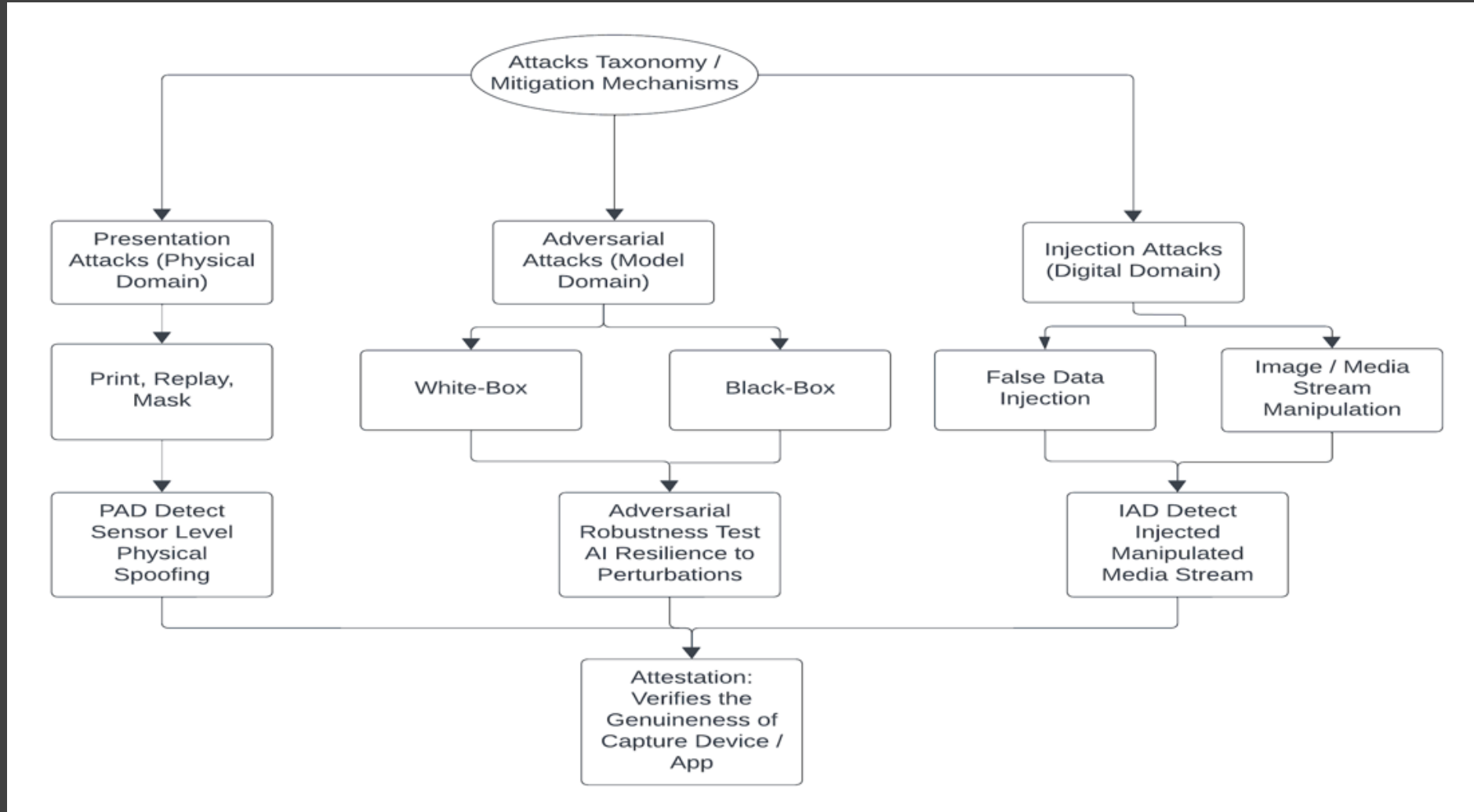
Dr Idris Zakariyya



Background and Motivation

- Face biometrics protect critical systems in finance and border control.
- AI-driven manipulation is rapidly expanding the attack surface.
- Morphing attacks manipulated images matching multiple identities.
- Injection Attack (IA) manipulated images / media streams.
- Presentation Attack (PA) remain the most common threats to biometric system.
- Existing Presentation Attack Detection (PAD) / Injection Attack Detection (IAD) techniques lacks robustness to digital adversarial threats.

Simplified Threat View for Face Biometrics



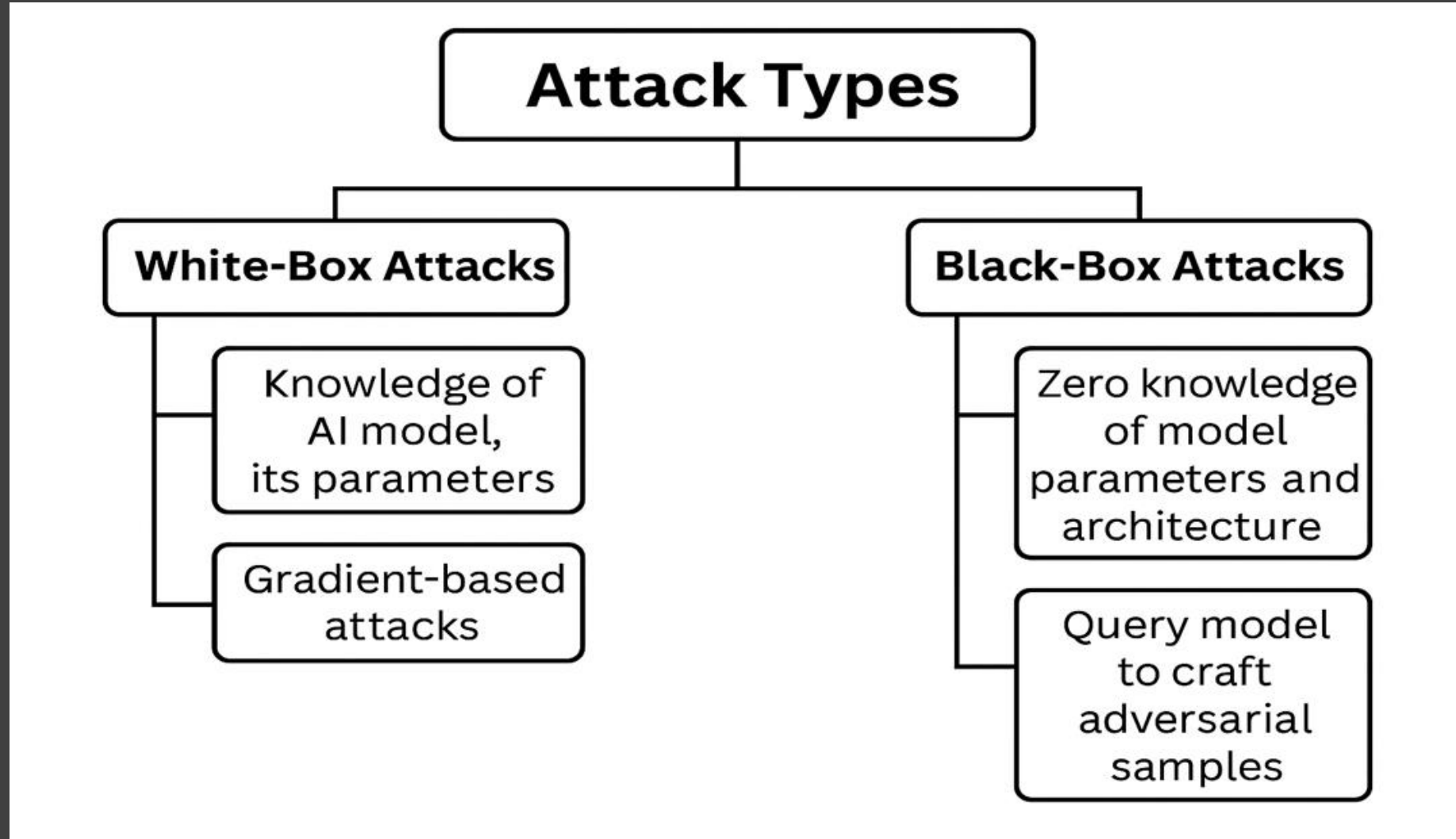
Note: Attack categories can overlap. For example, a deepfake may be presented to a camera, injected via a virtual camera, or used during enrolment.

PAD Performance and Robustness Benchmarking

- Effectiveness in detecting physical spoofing, measured by APCER and BPCER (ISO/IEC 30107).
- Model resilience to white-box and black-box adversarial attacks.



Adversarial Attacks



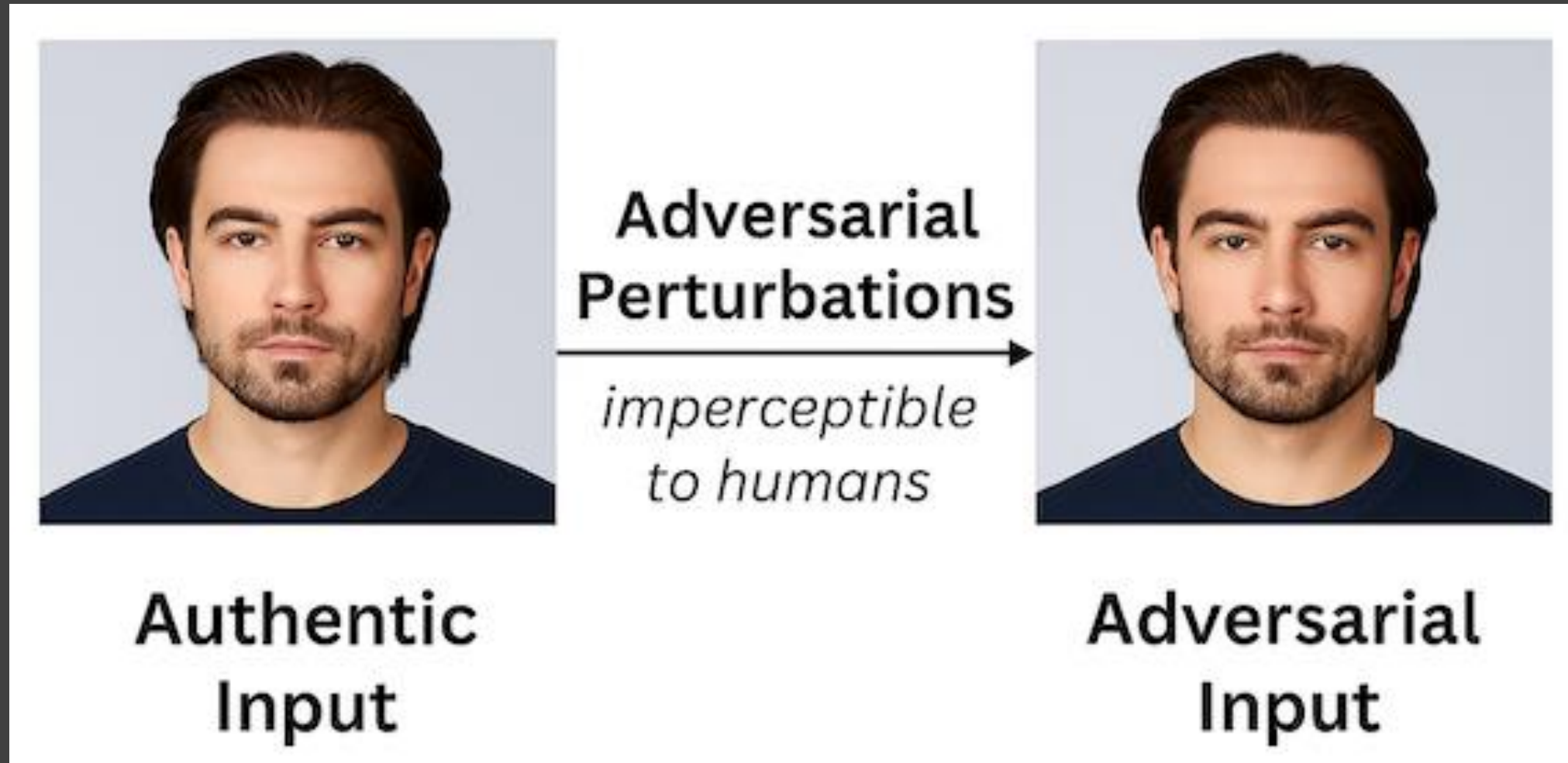
FGSM Adversarial Attack Generation Procedure

FGSM adversarial sample is generated by adding small perturbation to the original input in Equation 1.

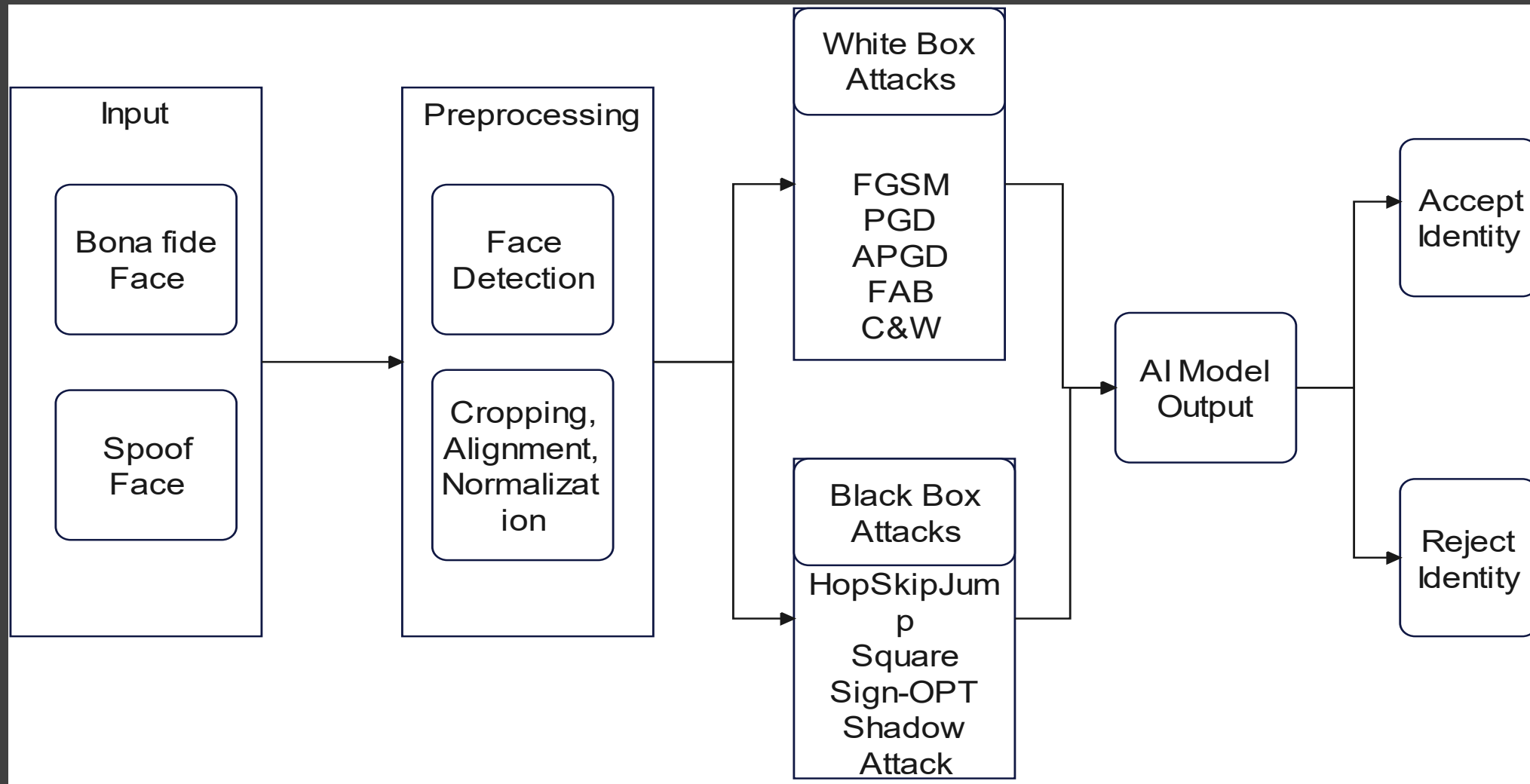
$$\mathbf{x}_{adv} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(\theta, \mathbf{x}, y)) \quad (1)$$

- \mathbf{x} is the (authentic) input image (e.g., face image), \mathbf{x}_{adv} is the adversarial input image.
- ϵ is the perturbation magnitude (controls noise strength)
- θ represents the model parameters (white-box access)

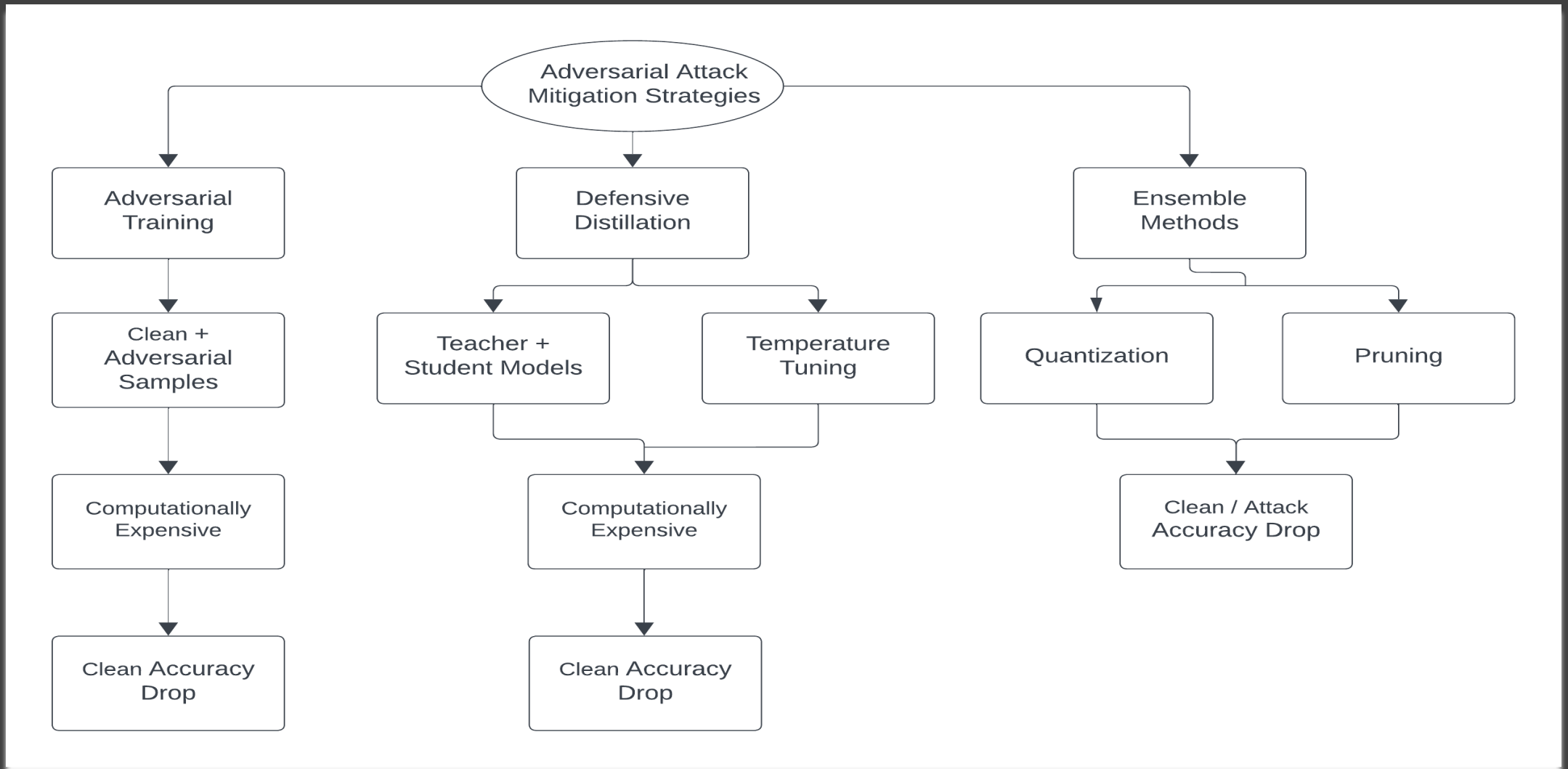
FGSM Generated Image



Face PAD Robustness Evaluation Framework



Adversarial Attack Mitigation Strategies



Online Hard Example Mining (OHEM) for Defense

- Integrated with PAD Training
- Optimize AI model robustness by focusing training on the hardest samples.
- Samples are dynamically weighted per batch based on decision difficulty.
- It is an adaptive technique that ensures model learns effectively and efficiently.

Evaluation

- Investigated two models ResNet50 and ViT-B16x4 foundational model.
- Employed CASIA-FASD: 1,200 attack (spoof) images and 455 bona fide images.
- Evaluated both models for PAD and adversarial robustness.
- Generate adversarial samples using FGSM ($\epsilon = 8/255$) and perform a Shadow attack with 10 optimization steps.

Face PAD Performance Results

Table 1: PAD Performance of ResNet50 and ViT-B16x4 Model against CASIA-FASD

Model	#Params	AUC	APCER	BPCER	ACER
ResNet50-FT	23.5M	99.99	0.00	20.95	10.48
ResNet50-Sel	14.9M	99.25	0.97	0.95	0.96
ResNet50-LP	2.1K	87.85	22.73	19.05	20.89
ResNet50-LPOHEM	2.1K	88.82	23.70	15.24	19.47
ViT-B16x4-FT	85.8M	99.99	0.00	4.76	2.38
ViT-B16x4-FTOHEM	85.8M	99.62	0.65	7.62	4.14
ViT-B16x4-Sel	21.3M	99.99	0.00	4.76	2.38
ViT-B16x4-LP	769	95.63	12.66	7.62	10.14

Face PAD Robustness Results

Table 2: Adversarial Robustness Performance of ResNet50 and ViT-B16x4 Model against CASIA-FASD

Procedure	Model	#Params	AUC	APCER	BPCER	ACER
FGSM Attack	ResNet50-FT	23.5M	50.15	47.40	46.67	47.04
	ResNet50-Sel	14.9M	50.15	47.40	46.67	47.04
	ResNet50-LP	2.1K	50.73	49.35	50.48	49.92
	ResNet50-LPOHEM	2.1K	53.75	45.45	45.71	45.58
	ViT-B16x4-FT	85.8M	17.43	71.75	72.38	72.07
	ViT-B16x4-FTOHEM	85.8M	47.82	38.96	38.09	38.52
Shadow Attack	ResNet50-FT	23.5M	63.27	39.29	38.09	38.69
	ResNet50-Sel	14.9M	63.27	39.29	38.09	38.69
	ResNet50-LP	2.1K	65.71	38.96	38.09	38.53
	ResNet50-LPOHEM	2.1K	66.55	36.68	38.09	37.39
	ViT-B16x4-FT	85.8M	90.21	19.80	19.04	19.42
	ViT-B16x4-FTOHEM	85.8M	96.75	9.74	10.47	10.11

Conclusion

- Investigated two models ResNet50 and ViT-B16x4 foundation model under PAD and adversarial attacks.
- For PAD, selective finetuning provides better BPCER and APCER for ResNet50 model.
- On CASIA-FASD, OHEM improved empirical robustness against the tested FGSM and Shadow attack settings.
- OHEM with ViT-B16x4 detect shadow attack with 96.75% AUC.
- These results are an initial step toward evaluating robustness for real-world deployment.

Limitation

- Results are based on CASIA-FASD and two attack settings.
- The study does not yet evaluate morphing, virtual-camera injection, document fraud, template attacks, demographic effects, or full identity-proofing workflows.
- Further cross-dataset and operational testing is needed.

Questions