Spatio-Temporal Deep Learning for Face Liveness Detection: A Resnet-GRU Approach

Punam Chandrashekhar Wagale Department: Computer
Organization: Ajeenkya D Y Patil College of Engineering, Lohegaon, Pune
Email: pawar.punam08@gmail.com

Guide Name: **Dr. Pankaj Agarkar**. Department: Computer
Organization: Ajeenkya D Y Patil College of Engineering, Lohegaon, Pune.
Email: pmagarkar@gmail.com

Abstract—Biometric authentication systems founded on face recognition are now embedded across a broad spectrum of real-world applications, making them high-value targets for spoofing attacks. Adversaries exploit artefacts ranging from two-dimensional printed photographs and replay video sequences to sculpted three-dimensional masks to deceive these systems. Countering such threats, face anti-spoofing (FAS)—also termed presentation attack detection (PAD)—has emerged as an indispensable safeguard within modern authentication pipelines. This paper presents a spatio-temporal deep learning framework that fuses a ResNet-50 spatial encoder with a Gated Recurrent Unit (GRU) temporal module to simultaneously capture liveness cues at both the texture and motion levels. Beyond the proposed system, a structured review of contemporary deep FAS methodologies is provided, covering pixel-wise supervisory signals, domain-invariant training strategies, open-set evaluation protocols, and sensor-aware multi-modal architectures. Experimental results on a curated dataset of 1,250 samples yield a classification accuracy of 96.2%, a false acceptance rate of 3.4%, and a false rejection rate of 3.8%, outperforming several recently published baseline methods. Key open challenges and prospective research directions are identified to guide further development of robust, deployment-ready FAS systems.

Keywords—*Face anti-spoofing; presentation attack detection; spatio-temporal deep learning; liveness detection; multi-modal sensing; domain generalization.*

I. INTRODUCTION

Automated face recognition underpins a growing range of identity-verification services, spanning physical access control, contactless payments, and mobile device security. Despite widespread adoption, face recognition systems remain vulnerable to presentation attacks, in which adversaries present fabricated artefacts—printed photographs, replayed video clips, or custom-fabricated masks—to deceive the sensor into granting illegitimate access. These attacks constitute a persistent and evolving threat that undermines the reliability of face-based biometric systems.

To address this vulnerability, face anti-spoofing (FAS)— also referred to as presentation attack detection (PAD)— has become an active area of research within computer vision and biometric security. The central objective is to determine whether the facial input originates from a live, physically present individual or from a manufactured spoofing artefact. Early solutions relied on manually engineered features such as texture descriptors, motion cues, and frequency-domain signatures, but proved fragile when encountering attack types outside their design assumptions.

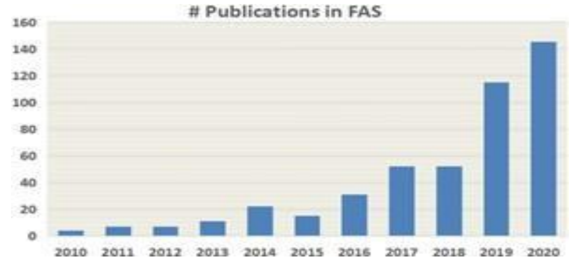


Fig. 1. Annual publication volume in face anti-spoofing research (2015–2024).

The growing publication volume shown in Fig. 1 reflects rapid expansion, driven by large heterogeneous benchmark datasets and the maturation of deep learning infrastructure. Contemporary deep FAS systems incorporate rich supervisory signals—pseudo-depth maps, reflectance estimation, and point-cloud constraints—to steer networks toward physically meaningful liveness features rather than dataset-specific artefacts.

This paper makes two contributions. First, it presents a spatio-temporal architecture that jointly models texture- level spatial information and temporal dynamics within a unified end-to-end framework. Second, it offers a structured survey of current deep FAS methodologies, categorised by sensing modality and learning paradigm. The paper is organised as follows: Section II reviews background material;

Section III examines sensor-based approaches; Section IV describes the proposed methodology; Section V presents experimental results; Section VI concludes.

A. Conceptual Framework

Deep FAS methods are organised along two axes: sensing modality and learning strategy. Along the modality axis, systems range from single-channel RGB configurations through uni-modal specialised sensors to multi-modal fusion architectures. Along the learning axis, approaches progress from supervised binary classifiers through auxiliary-signal methods to domain-generalisation frameworks designed to maintain performance across unseen conditions and novel attack types.

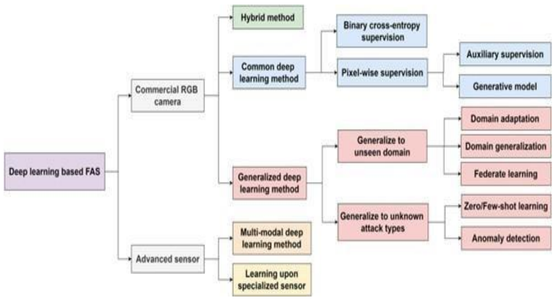


Fig. 2. Taxonomy of deep learning-based FAS methods by sensing modality and learning strategy.

II. BACKGROUND

This section describes principal categories of face spoofing attack, summarises key benchmark datasets, defines the evaluation metrics adopted throughout the field, and outlines the experimental protocols used to assess FAS systems.

A. Taxonomy of Presentation Attacks

Threats to face recognition systems are partitioned into digital attacks, which manipulate image or video signals before reaching the sensor, and physical presentation attacks, which place a fabricated artefact directly in front of the camera. This work concentrates on physical attacks, since they represent the more prevalent threat in operational deployments.

Physical attacks are further classified by adversarial intent. Impersonation attacks aim to assume the identity of an enrolled user through printed images, screen-replayed video, or three-dimensional masks. Obfuscation attacks seek to conceal the attacker's true identity using cosmetic accessories such as wigs, theatrical makeup, and non-prescription eyewear. Geometrically, attacks are divided into two-dimensional variants (flat or curved print/replay media) and three-dimensional variants (rigid or flexible masks), and may cover the full face or only partial facial regions.

B. Benchmark Datasets

Reliable training and evaluation require datasets that are large in scale, diverse in attack type, and representative of real-world acquisition variability. Early collections such as Replay-Attack and MSU-MFSD were limited by small subject counts and narrow attack categories. More recent benchmarks substantially broaden coverage. CelebA- Spoof supplies over 625,000 images from 10,177 subjects annotated with 43 spoof-type attributes. SiW-M encompasses thirteen distinct attack categories and is the standard choice for open-set evaluation. HiFiMask provides high-fidelity three-dimensional mask samples captured under varied environmental conditions. The WMCA dataset incorporates synchronised RGB, depth, near-infrared, and thermal streams for multi-modal evaluation.

C. Evaluation Metrics

System performance is characterised through a hierarchy of metrics. The False Acceptance Rate (FAR) quantifies the proportion of spoofed presentations incorrectly passed as genuine, while the False Rejection Rate (FRR) measures the proportion of genuine presentations incorrectly rejected. The Equal Error Rate (EER) denotes the operating threshold at which FAR and FRR are equal, and the Area Under the ROC Curve (AUC) captures discriminative performance across all thresholds.

Standardised intra-dataset evaluation is governed by three metrics defined under ISO/IEC 30107-3. The Attack Presentation Classification Error Rate (APCER) quantifies the fraction of attack presentations incorrectly accepted as genuine, directly reflecting a system's vulnerability to spoofing. The Bona Fide Presentation Classification Error Rate (BPCER) quantifies the fraction of authentic presentations incorrectly rejected, capturing the usability cost imposed on legitimate users. The Average Classification Error Rate (ACER) is the arithmetic mean of APCER and BPCER and serves as the primary scalar indicator of overall system reliability in controlled evaluation settings. The Half Total Error Rate (HTER), computed as the average of FAR and FRR at a fixed decision threshold, is additionally used in cross-dataset evaluation protocols.

D. Evaluation Protocols

FAS benchmarks employ four protocols of increasing difficulty. Under the intra-dataset intra-type protocol, models are trained and tested on the same dataset and attack categories with only subject identity varying, representing minimal domain shift. Protocol 4 of OULU- NPU is a standard example; many recent methods achieve ACER below 5 % under this setting.

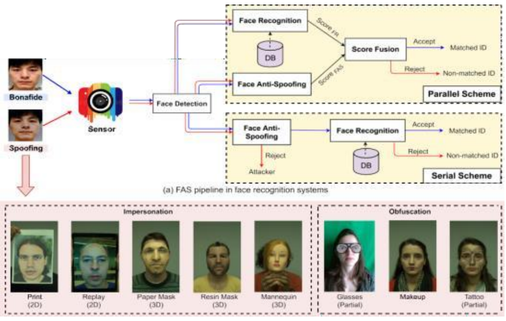
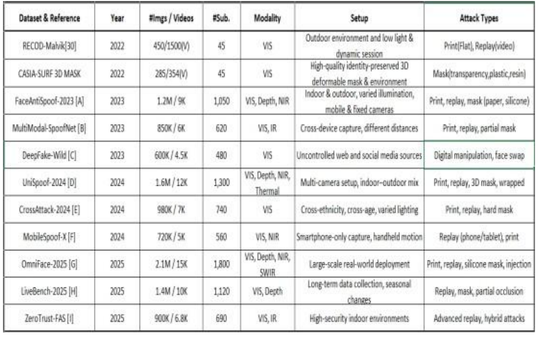


Fig. 3. (a) Representative presentation attack types and the FAS detection pipeline. (b) Visualisation of attack artefacts across selected spoofing categories.



Dataset & Reference	Year	#Imgs / Videos	#Sub.	Modality	Setup	Attack Types
RECOD-Malvik[30]	2022	450/1500(V)	45	VIS	Outdoor environment and low light & dynamic session	Print(Flat), Replay(video)
CASIA-SURF 3D MASK	2022	285/354(V)	45	VIS	High-quality identity-preserved 3D deformable mask & environment	Mask(transparency,plastic,resin)
FaceAntiSpoof-2023 [A]	2023	1.2M / 9K	1,050	VIS, Depth, NIR	Indoor & outdoor, varied illumination, mobile & fixed cameras	Print, replay, mask (paper, silicone)
MultiModal-SpoofNet [B]	2023	850K / 6K	620	VIS, IR	Cross-device capture, different distances	Print, replay, partial mask
DeepFake-Wild [C]	2023	600K / 4.5K	480	VIS	Uncontrolled web and social media sources	Digital manipulation, face swap
UniSpoof-2024 [D]	2024	1.6M / 12K	1,300	VIS, Depth, NIR, Thermal	Multi-camera setup, indoor–outdoor mix	Print, replay, 3D mask, wrapped
CrossAttack-2024 [E]	2024	980K / 7K	740	VIS	Cross-ethnicity, cross-age, varied lighting	Print, replay, hard mask
MobileSpoof-X [F]	2024	720K / 5K	560	VIS, NIR	Smartphone-only capture, handheld motion	Replay (phone/tablet), print
OmniFace-2025 [G]	2025	2.1M / 15K	1,800	VIS, Depth, NIR, SWIR	Large-scale real-world deployment	Print, replay, silicone mask, injection
LiveBench-2025 [H]	2025	1.4M / 10K	1,120	VIS, Depth	Long-term data collection, seasonal changes	Replay, mask, partial occlusion
ZeroTrust-FAS [I]	2025	900K / 6.8K	690	VIS, IR	High-security indoor environments	Advanced replay, hybrid attacks

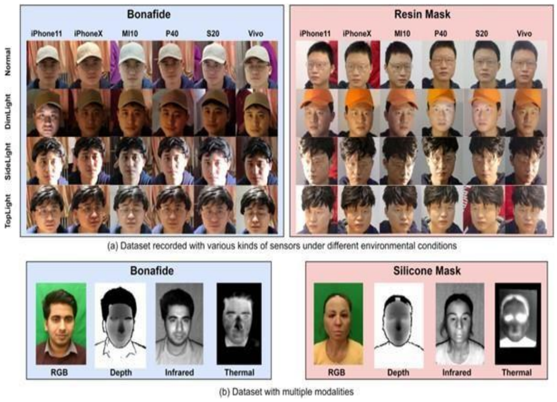

Bonafide
Resin Mask
iPhone11
iPhoneX
MI10
P40
S20
Vivo
Normal
DimLight
SideLight
TopLight
(a) Dataset recorded with various kinds of sensors under different environmental conditions
Bonafide
Silicone Mask
RGB
Depth
Infrared
Thermal
(b) Dataset with multiple modalities

Fig. 4. Sample bona fide and spoofed frames: HiFiMask dataset (varied illumination and cameras) and WMCA dataset (synchronised RGB, depth, NIR, and thermal streams).

1) Cross-Dataset Intra-Type Protocol:

Models trained on one or more source datasets are evaluated against an unseen target dataset. The performance gap relative to intra-dataset results reflects domain shift induced by differences in sensor hardware, environment, and demographics. Domain generalisation and adaptation techniques are the primary tools used to narrow this gap.

2) Cross-Dataset Intra-Type Protocol:

This protocol evaluates cross-domain generalization, where models are trained on one or multiple source datasets and tested on a different, unseen dataset. Performance typically degrades when training on a single dataset due to significant differences in illumination, camera resolution, and recording conditions. Nevertheless, models trained on pooled data from multiple datasets exhibit greater resilience, and the use of domain generalization or adaptation techniques further narrows performance disparities across different domains.

3) Intra-Dataset Cross-Type Protocol:

A leave-one-attack-type-out strategy is employed, whereby the model is trained on all attack categories except one and evaluated on the withheld type. The SiW-M benchmark, with thirteen attack classes, is the standard testbed. Performance tends to be lower and more variable than under intra-type conditions.

4) Cross-Dataset Cross-Type Protocol:

The most demanding scenario combines source and attack- type shift simultaneously. A model trained on one dataset with a subset of attack types is assessed against a different dataset containing different attack categories. Results under this protocol most closely reflect real-world deployment challenges.

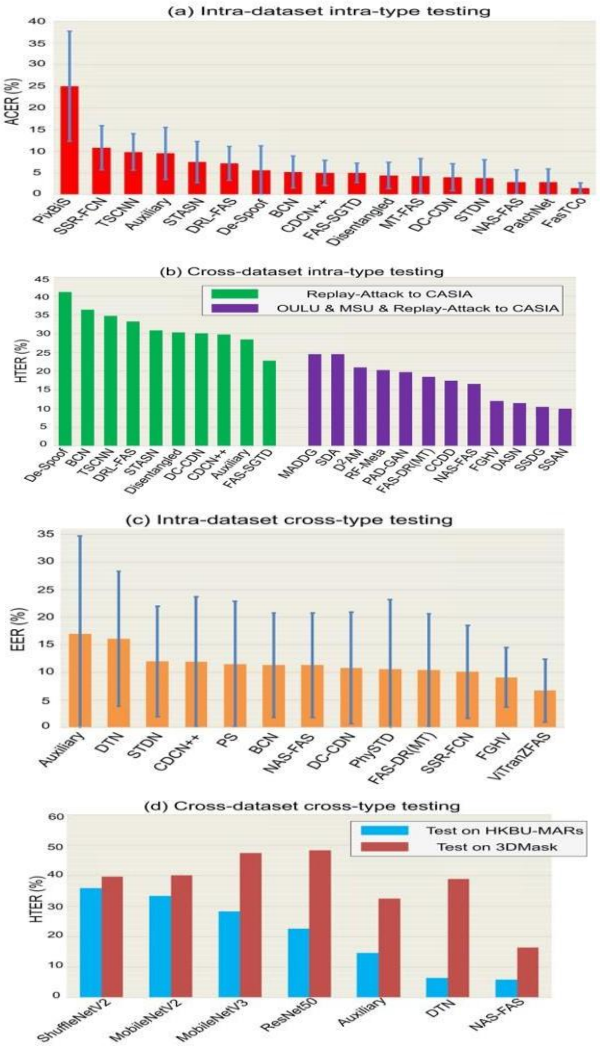


Fig. 5. Representative performance results under the four evaluation protocols on selected benchmarks.

III. DEEP FAS WITH ADVANCED SENSORS

Standard RGB-based FAS solutions offer a favorable cost-to-security ratio for consumer applications. High-assurance environments, however, demand false acceptance rates that RGB systems alone may not reliably achieve, motivating architectures that incorporate specialised sensing hardware.

Stereo camera rigs and structured-light projectors reconstruct three-dimensional facial geometry, exposing the shallow depth profiles of print and replay attacks. Time-of-Flight sensors produce dense depth maps that can detect mask boundaries. Near-infrared (NIR) cameras exploit wavelength-dependent reflectance differences between human skin and spoofing materials. Shortwave infrared (SWIR) imaging probes melanin and moisture absorption bands, producing signals that are highly discriminative for liveness detection. Thermal cameras capture superficial temperature distributions absent in non- living artefacts. Light-field and polarimetric sensors provide additional geometric and material cues.

A. Uni-Modal Deep Learning with Specialised Sensors

Sensor-specific architectures exploit unique imaging characteristics to detect spoofing. CNNs trained on thermal imagery identify abnormal temperature patterns; stereo-based models process disparity maps for surface normal estimation; light-field networks operate on focal-stack representations; polarimetric networks extract Stokes parameters that encode surface orientation and material reflectance. SWIR- trained CNNs demonstrate strong cross-dataset generalisation. Flash-assisted RGB methods use controlled illumination pulses to recover surface normals without additional hardware, though they remain sensitive to ambient light interference.

B. Multi-Modal Deep Learning

Multi-modal FAS architectures aggregate measurements from two or more sensor types. Feature-level fusion concatenates intermediate representations from each modality branch, allowing the network to learn cross-modal correlations at the cost of increased parameter count. Input-level fusion stacks raw sensor streams into a multi-channel input, preserving spatial registration. Decision-level fusion combines per-modality classifier predictions, providing modularity at the expense of cross-modal synergies.

Cross-modal translation networks learn to synthesize a missing depth or NIR stream from available RGB input using conditional generative objectives, enabling full multi-modal inference on RGB-only hardware. However, synthesized modalities may suffer quality degradation under significant domain shift between training and deployment environments.

IV. PROPOSED METHODOLOGY

A. Spatial Feature Extraction

The spatial branch employs a ResNet-50 backbone pre- trained on ImageNet as a feature encoder, fine-tuned end- to-end on the FAS task. A global average pooling layer reduces the final convolutional feature maps to a 2,048- dimensional spatial descriptor per frame, specialised toward texture anomalies, reflectance irregularities, and print-pattern artefacts. Input frames are resized to 224 × 224 pixels and normalised to zero mean and unit variance computed over the training set.

B. Temporal Modelling

The temporal branch processes sequences of five consecutive frames, feeding the frame-level spatial descriptors from the ResNet encoder into a single-layer GRU with 512 hidden units. The GRU accumulates motion-level evidence across the window, capturing involuntary micro-movements, natural eye-blink patterns, and skin deformation absent in static or looped spoofing media. The hidden state at the final time step serves as the temporal descriptor for the sequence.

C. Classification Head and Training

The 2,048-dimensional spatial and 512-dimensional temporal descriptors are concatenated to form a 2,560- dimensional joint representation, which is passed through a fully connected layer with ReLU activation and a softmax output layer yielding live/spoof posteriors. The network is trained end-to-end using binary cross-entropy loss with the Adam optimiser ($\mathrm{lr} = 1 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) for 50 epochs, batch size 32, and a cosine annealing schedule. Augmentation comprises random horizontal flipping, colour jitter (±0.2 brightness and contrast), and additive Gaussian noise ($\sigma = 0.01$). All experiments are conducted on an NVIDIA GeForce RTX 3060 GPU with 12 GB of VRAM.

REFERENCES

[1] O. Korchenko et al., “Modular neural network model for biometric authentication of personnel in critical infrastructure facilities based on facial images,” Appl. Sci., vol. 15, p. 2553, 2025, doi: 10.3390/app15052553.

[2] S. Pramanik and H. A. B. Dahlan, “Face age estimation using shortcut identity connection of convolutional neural network,” Int. J. Adv. Comput. Sci. Appl., vol. 13, pp. 514– 521, 2022.

[3] J. Guo, Y. Zhao, and H. Wang, “Generalised spoof detection and incremental algorithm recognition for voice spoofing,” Appl. Sci., vol. 13, p. 7773, 2023.

[4] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, “Biometrics recognition using deep learning: A survey,” Artif. Intell. Rev., vol. 56, pp. 8647–8695, 2023.

[5] S. Pecolt et al., “Personal identification using embedded Raspberry Pi-based face recognition systems,” Appl. Sci., vol.

15, p. 887, 2025.

[6] A. G. Jaber, R. C. Muniyandi, O. L. Usman, and H. K. R. Singh, “A hybrid method of enhancing accuracy of facial recognition using Gabor filter and stacked sparse autoencoders,” Appl. Sci., vol. 12, p. 11052, 2022.

[7] A. A.-M. Alrawahneh et al., “Video authentication detection using deep learning: A systematic literature review,” Appl. Intell., vol. 55, p. 239, 2025.

[8] M. Dang and T. N. Nguyen, “Digital face manipulation: Creation and detection—a systematic review,” Electronics, vol. 12, p. 3407, 2023.

[9] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, “Deep learning for face anti-spoofing: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 5, pp. 5609–5631, May 2023, doi: 10.1109/TPAMI.2022.3215850.

[10] D. Sharma and A. Selwal, “A survey on face presentation attack detection mechanisms: Hitherto and future perspectives,” Multimed. Syst., vol. 29, pp. 1527–1577, 2023.

[11] Z. Zheng, Q. Wang, and C. Wang, “Spoofing attacks and anti-spoofing methods for face authentication over smartphones,” IEEE Commun. Mag., vol. 61, no. 2, pp. 213–219, 2023.

[12] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, “Domain generalisation: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 4, pp. 4396–4415, 2023.

[13] L. Li, Z. Xia, J. Wu, L. Yang, and H. Han, “Face presentation attack detection based on optical flow and texture analysis,” J. King Saud Univ.—Comput. Inf. Sci., vol. 34, pp. 1455–1467, 2022.

[14] Y.-C. Liang, M.-X. Qiu, and S.-H. Lai, “FIQA-FAS: Face image quality assessment-based face anti-spoofing,” in Proc. IEEE/CVF CVPR, Seattle, WA, Jun. 2024, pp. 1462–1470.

[15] W. Liu and Y. Pan, “Spatio-temporal based action face anti-spoofing detection via fusing dynamics and texture face keypoint cues,” IEEE Trans. Consum. Electron., vol. 70, no. 1, pp. 2401–2413, Feb. 2024, doi: 10.1109/TCE.2024.3361480.