Optical Flow and Synthetic Depth Augmentation for Face Presentation Attack Detection

Muhammad Shahid Jabbar[a] and Shujaat Khan[a,b]

[a]SDAIA-KFUPM Joint Research Center for Artificial Intelligence, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia
[b]Department of Computer Engineering, College of Computing and Mathematics, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia

ABSTRACT

Face presentation attack detection (FacePAD) is crucial for preserving the integrity of modern vision-based facial authentication systems against spoofing attempts, including print, replay, and 3D mask attacks. In this work, we propose a triple-stream multimodal framework that jointly integrates synthetic depth maps and optical flow dynamics with original RGB video frames to achieve enhanced liveness detection and improved robustness. The proposed architecture comprises three coordinated branches: (i) a depth stream that models photometric and geometric facial structure using synthetic depth maps estimated directly from RGB frames, (ii) a flow stream that captures motion irregularities across temporal sequences through optical flow analysis, and (iii) an RGB stream that encodes fine-grained spatial and textural cues. Representations from these complementary modalities are fused to exploit depth–motion–appearance correlations, thereby strengthening discrimination between live and spoofed faces under diverse conditions. Importantly, both depth and flow representations are synthesized from monocular video input, requiring no additional sensing hardware or depth cameras. This design enables the model to perceive both static surface characteristics and dynamic behavioral patterns that are typically absent in spoof media, leading to stronger generalization to unseen attack types. Experimental evaluations on three benchmark datasets—Replay-Attack, Replay-Mobile, and ROSE-Youtu—demonstrate state-of-the-art performance, achieving 0.0% HTER on RA and RM, and a superior 0.454% HTER on RY, surpassing recent FacePAD approaches. Overall, this multimodal strategy provides a unified, scalable, and energy-efficient pathway toward robust and interpretable biometric authentication.

Keywords: Face presentation attack detection, multimodal fusion, optical flow, synthetic depth-map, face anti-spoofing, embedded AI, multimedia security, computer vision

1. INTRODUCTION

Face recognition systems are widely deployed in biometric and interactive AI applications, including smartphone authentication,[1] mobile payments,[2] access control,[3] and surveillance.[4] Recent advances in deep learning, particularly multi-branch and fusion-based architectures, have substantially improved face representation quality by integrating global structural information with fine-grained texture cues.[5,6] Despite these advances, face recognition systems remain vulnerable to presentation attacks, such as printed photos, replayed videos, and 2D/3D mask attacks, which can severely compromise system reliability. Consequently, face presentation attack detection (FacePAD) or face anti-spoofing has become a critical component of secure biometric systems.

Early FacePAD approaches relied on handcrafted texture descriptors, including LBP, SIFT, and SURF, combined with classical classifiers such as SVM or LDA. While computationally efficient, these methods are highly sensitive to illumination changes, camera characteristics, and attack quality, resulting in limited robustness against sophisticated spoofing media. Convolutional neural network (CNN)-based methods later become dominant, learning discriminative features directly from data in an end-to-end manner.[7–9] However, image-based

Further author information: (Send correspondence to S.K.)
M.S.J.: E-mail: muhammad.jabbar@kfupm.edu.sa, S.K.: E-mail: shujaat.khan@kfupm.edu.sa

Real-time Processing of Image, Depth, and Video Information 2026, edited by Gian Domenico Licciardo, Viktor J. Schneider, Matthias F. Carlsohn, Proc. of SPIE Vol. 14087, 1408705 · 2026 Published by SPIE
0277-786X · doi: 10.1117/12.3099891

models often overfit to dataset-specific artifacts, such as background patterns or display borders, leading to poor generalization across unseen attack types and acquisition conditions.

Recent FacePAD research has explored auxiliary supervision signals beyond static appearance cues to improve robustness. Physiological signals, such as remote photoplethysmography (rPPG), provide strong liveness features but are computationally expensive and sensitive to noise, limiting their suitability for real-time deployment.[10] Depth-map based approaches have also gained increasing attention, as geometric cues effectively capture the structural differences between bonafide faces and spoof media, which often exhibit planar or rigid characteristics.[11] While depth supervision improves generalization, many existing methods rely on dedicated depth sensors or heavy estimation pipelines, restricting their applicability.

In parallel, temporal information from video frames has emerged as a powerful cue for FacePAD. Genuine faces exhibit non-rigid and temporally consistent motion patterns that are difficult to replicate using spoofing media. Optical flow provides an effective representation of such dynamics by modeling pixel-level motion between consecutive video frames. Prior works have demonstrated that motion irregularities and temporal inconsistencies can help distinguish live faces from replay and mask-based attacks.[12,13] Nevertheless, temporal methods may struggle under low frame rates, motion blur, or limited facial movement, and explicit spatio-temporal modeling often increases computational complexity.

The observations on appearance, depth, and motion cues providing complementary information for FacePAD; modeling texture, geometric structure, and dynamic behavioral pattern features, respectively, from image or video frames provide the motivation for this work. However, most existing approaches exploit these cues in isolation or rely on additional sensing hardware, limiting their robustness and scalability.

In this work, we propose a unified multimodal FacePAD framework that jointly integrates RGB appearance, synthetic depth maps, and optical flow dynamics using only monocular video input. The proposed approach captures complementary depth–motion–appearance correlations by fusing spatial, geometric, and temporal representations, that improve discrimination between bonafide and spoofed faces across diverse attack scenarios. The proposed framework requires no additional sensors while achieving strong generalization and state-of-the-art performance on multiple benchmark datasets.

2. PROPOSED METHOD

2.1 Proposed Frame Processing and Augmentation

The input to the proposed framework consists of monocular RGB frames from video clips. From each clip, a reference RGB frame is sampled at time t $X_t^{(i)} \in \mathbb{R}^{3\times H\times W}$, where i is the video index from dataset. During training, the frame index t is selected randomly to increase data variability, while at inference time the first frame of each clip is used for consistency.

For capturing temporal motion cues, a temporally adjacent frame at time $t+\Delta t$ (with $\Delta t = 1$), $X_{t+\Delta t}^{(i)} \in \mathbb{R}^{3\times H\times W}$ is paired with the reference frame $X_t^{(i)}$. Given $\mathbf{U}_{t,t+\Delta t}^{(i)} \in \mathbb{R}^{2\times H\times W}$ represents pixel-wise displacement fields in horizontal and vertical directions, dense optical flow is estimated as

$$\mathbf{U}_{t,t+\Delta t}^{(i)} = (u, v) \;\leftarrow\; \text{FlowEngine}\Big(X_t^{(i)},\, X_{t+\Delta t}^{(i)}\Big)\,,$$

We instantiate FlowEngine using the *UniMatch* model[14] with the MixData-pretrained checkpoint and UniMatch default configuration, including a multi-scale transformer encoder, iterative refinement, and windowed attention. Input frames are scaled to the $[0, 255]$ range when required, transposed to ensure width $\geq$ height, and resized to multiples of 32 to match the training configuration of UniMatch. The predicted flow fields from frozen FlowEngine are resized back to the original resolution with magnitude correction, and any spatial transposition is reversed to preserve pixel-wise alignment. For compatibility with CNN-based backbones, the two-channel flow field hereby acquired is encoded into a three-channel representation using HSV-to-RGB colorwheel mapping $X_{\text{FLOW},t}^{(i)} \in \mathbb{R}^{3\times H\times W}$, which serves as input to the flow-branch of the proposed architecture.

For geometric cues extract, synthetic depth maps are generated from the reference RGB frame $X_t^{(i)}$ using the *Depth Anything V2* model.[15] This monocular depth estimation network employs a transformer-based

encoder–decoder architecture and is trained on large-scale synthetic and real-world data, providing strong generalization across diverse scenes. In this work, we use the ViT-B variant, which offers a favorable trade-off between accuracy and computational efficiency. Each reference RGB frame $X_t^{(i)}$ is passed through the pretrained and frozen depth estimation network to produce a corresponding synthetic depth map, $X_{\mathrm{DEPTH},t}^{(i)} \in \mathbb{R}^{1\times H\times W}$, which serves as input to the depth-branch of the proposed architecture.

Identical geometric pre-processing is applied to each of RGB, flow, and depth modalities to preserve spatial alignment. The frames are first center-cropped to the largest possible square region and then resized to $\rho \times \rho$ with $\rho = 224$. When resizing optical flow, vector magnitudes are scaled proportionally to maintain physical consistency. During training, synchronized random augmentations of in-plane rotation and isotropic scaling are applied with shared parameters across modalities to maintain alignment. At inference time, only deterministic center cropping and resizing are used. This unified processing pipeline enables consistent fusion of appearance, geometric, and temporal cues for robust face presentation attack detection.

2.2 Proposed Architecture

The proposed framwork adopts a multi-stream architecture to jointly model appearance, geometric, and temporal cues using RGB images, optical flow representations, and synthetic depth maps derived from monocular video. The architecture is illustrated in Fig. 1. The design follows a parallel processing paradigm, where each modality is encoded independently and fused at the feature level to exploit complementary information.

Each pre-processed and augmented input sample provides three aligned modality-specific inputs:

$$\tilde{X}_{\mathrm{RGB}}^{(i)} \in \mathbb{R}^{3\times\rho\times\rho}, \quad \tilde{X}_{\mathrm{FLOW}}^{(i)} \in \mathbb{R}^{3\times\rho\times\rho}, \quad \tilde{X}_{\mathrm{DEPTH}}^{(i)} \in \mathbb{R}^{1\times\rho\times\rho},$$

corresponding to the RGB frame, synthetic depth map, and encoded optical flow, respectively.

Each modality is processed by an independent CNN encoder. We adopt MobileNetV3-Large backbones for all three streams due to its favorable trade-off among representational capability and computational efficiency. All backbones are initialized with ImageNet-pretrained weights.[16] For the depth branch, the first convolutional layer is adapted to accept single-channel input. Each encoder outputs a D-dimensional feature embedding:

$$\mathbf{f}_{\mathrm{RGB}}^{(i)} = \mathrm{CNN}_{\mathrm{RGB}}(\tilde{X}_{\mathrm{RGB}}^{(i)}), \quad \mathbf{f}_{\mathrm{FLOW}}^{(i)} = \mathrm{CNN}_{\mathrm{FLOW}}(\tilde{X}_{\mathrm{FLOW}}^{(i)}), \quad \mathbf{f}_{\mathrm{DEPTH}}^{(i)} = \mathrm{CNN}_{\mathrm{DEPTH}}(\tilde{X}_{\mathrm{DEPTH}}^{(i)}).$$

The flow branch processes the encoded optical flow image, which implicitly encodes both motion magnitude and direction, enabling the network to capture subtle temporal irregularities associated with spoofing attacks.

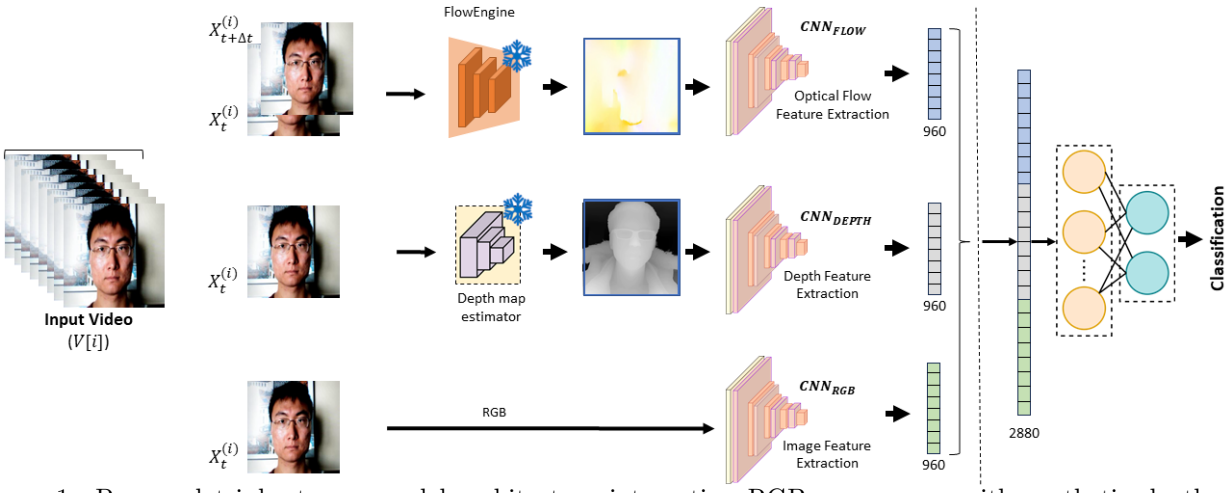

Figure 1: Proposed triple-stream model architecture integrating RGB appearance with synthetic depth and optical flow cues derived from monocular video input, processed by modality-specific CNN branches. The fused representation exploits complementary depth–motion–appearance correlations for face presentation attack detection.

The depth branch emphasizes geometric consistency and surface structure, while the RGB branch captures fine-grained appearance and texture cues.

Feature fusion is performed via late fusion by concatenating the modality-specific embeddings:

$$\mathbf{f}^{(i)} = \left[\, \mathbf{f}^{(i)}_{\mathrm{RGB}},\ \mathbf{f}^{(i)}_{\mathrm{FLOW}},\ \mathbf{f}^{(i)}_{\mathrm{DEPTH}} \,\right] \in \mathbb{R}^{3D}.$$

The fused representation is passed to a lightweight multi-layer perceptron (MLP) classifier, which outputs logits $\mathbf{s}^{(i)} \in \mathbb{R}^2$ corresponding to bonafide and spoof classes, with final predictions obtained using a softmax layer.

This unified multi-stream design enables the proposed model to jointly reason about static appearance, dynamic motion patterns, and 3D structure, leading to improved robustness and generalization against diverse presentation attack types.

We report both *classification* and *computational* metrics. Classification metrics include Accuracy, AUC-ROC, and Half Total Error Rate (HTER). HTER is defined as HTER = (FAR+FRR)/2, with FAR = $FP/(FP+TN)$ and FRR = $FN/(FN+TP)$, where TP, TN, FP, and FN denote true/false positives and negatives.

AUC-ROC is computed by threshold sweeping. We also report Equal Error Rate (EER), defined at the threshold where FAR = FRR, and the Youden index max(TPR − FPR) as a complementary summary.

2.3 Evaluation Metrics

We evaluate the proposed FacePAD framework using standard metrics commonly adopted in face anti-spoofing literature. Performance is primarily reported in terms of Half Total Error Rate (HTER), which balances false acceptance rates (FAR) and false rejection rates (FRR), defined as HTER = $\frac{\mathrm{FAR}+\mathrm{FRR}}{2}$, where the FAR and FRR are given by FAR = $\frac{FP}{FP+TN}$, FRR = $\frac{FN}{FN+TP}$.. Here, TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively.

In addition to HTER, we report Accuracy and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The ROC curve is obtained by sweeping the decision threshold over predicted scores. We also report the Equal Error Rate (EER), defined as the operating point where FAR = FRR, which provides a threshold-independent measure of system performance. These metrics together provide a comprehensive assessment of performance and robustness.

2.4 Implementation Details

The proposed model is trained using the Adam optimizer with an initial learning rate of 10^{-4} and cross-entropy loss. Training is performed with a batch size of 16 for a maximum of 100 epochs. Early stopping is applied based on validation loss with a patience of 20 epochs to prevent overfitting. All CNN backbones are initialized with ImageNet-pretrained weights. During training, synchronized data augmentations are applied consistently across RGB, depth, and flow modalities, as described in Section 2.1. At inference time, only deterministic pre-processing is used. All experiments follow the official subject-disjoint protocols provided by each dataset.

3. EXPERIMENTS AND RESULTS

3.1 Datasets

The proposed method is evaluated on three widely used FacePAD benchmark datasets: Replay-Attack (RA),[17] Replay-Mobile (RM),[18] and ROSE-Youtu (RY).[19] These datasets cover a broad range of presentation attack instruments, capture devices, and environmental conditions. The dataset statistics are summarized in Table 1.

Table 1: Summary of datasets used in this study.

Dataset	Year	Subjects	Genuine / Spoof Clips
Replay-Attack[17]	2012	50	300 / 1000
Replay-Mobile[18]	2016	39	540 / 624
ROSE-Youtu[19]	2018	20	1000 / 2350

Table 2: Performance comparison of different modality configurations on Replay-Attack, Replay-Mobile, and ROSE-Youtu datasets. Results are reported for image-only (I), image with synthetic-depth (ID), image with optical-flow (IF), and the proposed multi-stream image + depth + flow model (IFD).

Dataset / Method / Metric	Replay-Attack				Replay-Mobile				ROSE-Youtu			
	I	ID	IF	IFD	I	ID	IF	IFD	I	ID	IF	IFD
Accuracy (%)	100	100	100	100	100	100	100	100	98.06	99.03	98.57	99.49
AUC-ROC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.999	0.969	0.9993	0.9998
EER (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.39	1.15	0.98	0.46
HTER (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.47	1.023	0.94	0.45
FAR (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.39	1.15	0.92	0.46
FRR (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.56	0.89	0.97	0.45
Youden's Index	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.971	0.980	0.981	0.991

The Replay-Attack dataset consists of 1,300 short video clips from 50 subjects, including bonafide presentations and spoofing attempts using printed photos and replayed videos. The data are captured under both controlled and adverse lighting conditions using multiple cameras. Replay-Mobile is designed for mobile authentication scenarios and contains 1,164 video clips from 39 subjects, captured using mobile devices under varying illumination conditions, with both print and digital replay attacks. ROSE-Youtu comprises 3,350 video clips from 20 subjects, recorded using multiple camera models and diverse lighting environments, and additionally includes paper mask attacks, making it a particularly challenging benchmark. All datasets provide official training, validation, and test splits with subject-disjoint protocols, which are followed in our experiments.

3.2 Results

The quantitative results for the proposed augmented flow and depth FacePAD framework are presented in Table 2, with comparision of different modality configurations. We evaluate four model/modality variants if image-only (I), image with synthetic depth (ID), image with optical flow (IF), and the proposed model combining image, depth, and flow (IFD). On the RA and RM datasets, the proposed framework as well as rest of configurations achieve perfect presentation attack detection with 0.0% HTER and 100% Accuracy.

On the RY dataset, incorporating either synthetic depth (ID, 1.023% HTER) or optical flow (IF, 0.94% HTER) consistently improves performance over image-only model (I, 1.47% HTER). Whereas, the proposed framework (IFD) achieves the best overall performance, attaining an HTER of 0.45% and an EER of 0.46%.

3.3 Comparison with State-of-the-Art

Table 3 compares the proposed IFD framework with recent state-of-the-art FacePAD approaches on the RA, RM, and RY datasets. On the RA and RM datasets, the proposed method achieves ideal HTER with 0.0%, matching the best reported results in the recent literature. On the RY dataset. The proposed model achieves an HTER of **0.45%** and an EER of **0.46%**, outperforming prior approaches based on depth-augmented methods,[29]

Table 3: Comparison with recent state-of-the-art FacePAD methods on RA, RM, and RY datasets. Results are reported as HTER (%) for RA and RM, and HTER/EER (%) for RY.

Method	Year	RA	RM	RY
3D ConvNet[20]	2023	11.7	8.7	21.3/-
Image Diffusion[21]	2023	0.09	1.14	4.92/4.95
ResNet50V2[22]	2023	0.03	0.0	2.53/2.64
ViViT[23]	2023	–	–	13.28/2.46
EfficientNet-B0[24]	2024	36.88	4.62	9.54/-
CA-FAS[25]	2024	–	–	1.37/-
DeformMobileNet[26]	2025	0.0	0.0	1.26/0.80
Spatio-Temporal[27]	2025	0.0	0.0	1.47/0.85
AdvSpoofGuard[28]	2025	0.0	0.0	1.97/1.08
Depth-aug. Teacher[29]	2025	0.0	0.0	1.02/1.15
Depth-aug. Student[29]	2025	0.0	0.0	1.01/1.39
Proposed (IFD)	2026	**0.0**	**0.0**	**0.45/0.46**

spatio-temporal models,[27] and recent confidence-aware and adversarially robust approaches.[25,28] The consistent gains on RY demonstrate the effectiveness of the proposed FacePAD multimodal fusion framework in handling diverse attack types and capture conditions to enable more robust liveness detection performance.

4. CONCLUSION

This paper presents a multi-stream multimodal framework for face presentation attack detection that jointly integrates RGB appearance, optical flow dynamics, and synthetic depth maps using only monocular video input. By explicitly modeling complementary texture, geometric structure, and motion cues, the proposed framework provides a more comprehensive characterization of facial liveness than image- or video-frame-based, dual-modality, and conventional spatio-temporal approaches. Extensive experiments on three benchmark datasets demonstrate the effectiveness of the proposed framework. The model achieved perfect performance on the Replay-Attack and Replay-Mobile datasets and established state-of-the-art results on the more challenging ROSE-Youtu dataset, attaining an HTER of 0.45%. These results confirm that while depth and motion cues individually enhance robustness over RGB-only models, their joint integration yields the strongest generalization. Importantly, both depth and flow representations are synthesized directly from monocular RGB video, eliminating the need for additional sensing hardware or depth cameras. This design choice makes the proposed approach practical and scalable for real-world biometric authentication systems.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum and Minerals (KFUPM) under SDAIA-KFUPM Joint Research Center for support and resources. Shujaat Khan also acknowledges the support from the King Fahd University of Petroleum & Minerals (KFUPM) under Early Career Research Grant no. EC241027.

REFERENCES

[1] Kong, C., Zheng, K., Wang, S., Rocha, A., and Li, H., "Beyond the pixel world: A novel acoustic-based face anti-spoofing system for smartphones," *IEEE Transactions on Information Forensics and Security* **17**, 3238–3253 (2022).

[2] Ye, W., Chen, W., and Fortunati, L., "Mobile payment in china: A study from a sociological perspective," *Journal of Communication Inquiry* **47**(3), 222–248 (2023).

[3] Lee, H., Park, S.-H., Yoo, J.-H., Jung, S.-H., and Huh, J.-H., "Face recognition at a distance for a stand-alone access control system," *Sensors* **20**(3), 785 (2020).

[4] Fang, H., Liu, A., Wan, J., Escalera, S., Zhao, C., Zhang, X., Li, S. Z., and Lei, Z., "Surveillance face anti-spoofing," *IEEE Transactions on Information Forensics and Security* **19**, 1535–1546 (2023).

[5] Jiang, K., Wang, Z., Yi, P., Lu, T., Jiang, J., and Xiong, Z., "Dual-path deep fusion network for face image hallucination," *IEEE Transactions on Neural Networks and Learning Systems* **33**(1), 378–391 (2020).

[6] Jiang, K., Wang, Z., Yi, P., Wang, G., Gu, K., and Jiang, J., "Atmfn: Adaptive-threshold-based multi-model fusion network for compressed face hallucination," *IEEE Transactions on Multimedia* **22**(10), 2734–2747 (2019).

[7] Lucena, O., Junior, A., Moia, V., Souza, R., Valle, E., and Lotufo, R., "Transfer learning using convolutional neural networks for face anti-spoofing," in [*Image Analysis and Recognition: 14th International Conference, ICIAR 2017, Montreal, QC, Canada, July 5–7, 2017, Proceedings 14*], 27–34, Springer (2017).

[8] Nagpal, C. and Dubey, S. R., "A performance evaluation of convolutional neural networks for face anti spoofing," in [*2019 international joint conference on neural networks (IJCNN)*], 1–8, IEEE (2019).

[9] Gan, J., Li, S., Zhai, Y., and Liu, C., "3d convolutional neural network based on face anti-spoofing," in [*2017 2nd international conference on multimedia and image processing (ICMIP)*], 1–5, IEEE (2017).

[10] Liu, Y., Jourabloo, A., and Liu, X., "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 389–398 (2018).

[11] Atoum, Y., Liu, Y., Jourabloo, A., and Liu, X., "Face anti-spoofing using patch and depth-based cnns," in [*2017 IEEE international joint conference on biometrics (IJCB)*], 319–328, IEEE (2017).

[12] Feng, L., Po, L.-M., Li, Y., Xu, X., Yuan, F., Cheung, T. C.-H., and Cheung, K.-W., "Integration of image quality and motion cues for face anti-spoofing: A neural network approach," *Journal of Visual Communication and Image Representation* **38**, 451–460 (2016).

[13] Shao, R., Lan, X., and Yuen, P. C., "Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing," *IEEE Transactions on Information Forensics and Security* **14**(4), 923–938 (2018).

[14] Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Yu, F., Tao, D., and Geiger, A., "Unifying flow, stereo and depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(11), 13941–13958 (2023).

[15] Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., and Zhao, H., "Depth anything v2," (2024).

[16] "Pretrained imagenet-1k weights for mobilenet-v3-large: MobileNet_V3_Large_Weights.IMAGENET1K_V1," (2024). [Accessed: 2025-05-27].

[17] Chingovska, I., Anjos, A., and Marcel, S., "On the effectiveness of local binary patterns in face anti-spoofing," in [*2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*], 1–7, IEEE (2012).

[18] Costa-Pazo, A., Bhattacharjee, S., Vazquez-Fernandez, E., and Marcel, S., "The replay-mobile face presentation-attack database," in [*2016 international conference of the Biometrics Special Interest Group (BIOSIG)*], 1–7, IEEE (2016).

[19] Li, H., Li, W., Cao, H., Wang, S., Huang, F., and Kot, A. C., "Unsupervised domain adaptation for face anti-spoofing," *IEEE Transactions on Information Forensics and Security* **13**, 1794–1809 (2018).

[20] Giurato, S., Ortis, A., and Battiato, S., "Real-time multiclass face spoofing recognition through spatiotemporal convolutional 3d features," in [*International Conference on Image Analysis and Processing*], 356–367, Springer (2023).

[21] Alassafi, M. O., Ibrahim, M. S., Naseem, I., AlGhamdi, R., Alotaibi, R., Kateb, F. A., Oqaibi, H. M., Alshdadi, A. A., and Yusuf, S. A., "A novel deep learning architecture with image diffusion for robust face presentation attack detection," *IEEE Access* (2023).

[22] Alassafi, M. O., Ibrahim, M. S., Naseem, I., AlGhamdi, R., Alotaibi, R., Kateb, F. A., Oqaibi, H. M., Alshdadi, A. A., and Yusuf, S. A., "Fully supervised contrastive learning in latent space for face presentation attack detection," *Applied Intelligence* **53**(19), 21770–21787 (2023).

[23] Marais, M., Brown, D., Connan, J., and Boby, A., "Facial liveness and anti-spoofing detection using vision transformers," in [*Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*], **8**, 2023 (2023).

[24] Huszár, V. D. and Adhikarla, V. K., "Securing phygital gameplay: Strategies for video-replay spoofing detection," *IEEE Access* (2024).

[25] Long, X., Zhang, J., and Shan, S., "Confidence aware learning for reliable face anti-spoofing," *IEEE Transactions on Information Forensics and Security* (2025).

[26] Ibrahim, S. M., Ibrahim, M. S., Khan, S., Ko, Y.-W., and Lee, J.-G., "Improving face presentation attack detection through deformable convolution and transfer learning," *IEEE Access* (2025).

[27] Khan, S., Siddique, T. H. M., Ibrahim, M. S., Siddiqui, A. J., and Huang, K., "Spatio-temporal deep learning for improved face presentation attack detection," *Knowledge-Based Systems* , 113059 (2025).

[28] Siddique, T. H. M., Khan, S., Wang, Z., and Huang, K., "Advspoofguard: Optimal transport driven robust face presentation attack detection system," *Knowledge-Based Systems* , 113759 (2025).

[29] Jabbar, M. S., Siddique, T. H. M., Huang, K., and Khan, S., "Knowledge distillation with predicted depth for robust and lightweight face presentation attack detection," *Knowledge-Based Systems* , 114325 (2025).