Physics-Informed Deepfake Detection in Facial Images Using Landmark Geometry and Anatomy-Aware Hybrid Classification

Soumitra Ghosh*[1] and Sudhir Ranjan Pattanaik[1]

[1]Department of Computer Science and Engineering, NIST University, Berhampur, Odisha, India - 760010
soumitra.ghosh468@gmail.com, sudhirpattanaik.oca@gmail.com

Abstract. Deepfake detectors trained on a single dataset often fail under unseen manipulations because they rely on dataset-specific artifacts rather than facial characteristics. We propose a physics-informed deepfake detector that grounds decision-making in anatomical invariants of real faces, such as bilateral symmetry, smooth contours, and characteristic proportions, while deepfakes often break these patterns due to synthesis artifacts. A multi-stream CNN-GNN module extracts robust landmarks, and five anatomy-aware descriptor families measure physics violations in a hybrid geometry-appearance classifier. Evaluated across three cross-dataset protocols on FFHQ, CelebA, and FaceForensics++, the model achieves 95.8% accuracy on FaceForensics++ and 92.3% mean cross-dataset accuracy averaged over FF++→FFHQ, FF++→CelebA, and cross-manipulation protocols, outperforming a CNN baseline by 11.2 percentage points and reducing cross-dataset variance by about 78%. The proposed framework shows greatly enhanced robustness to manipulations including Face2Face, FaceSwap, reenactment, and diffusion, with an accuracy standard deviation of 2.1% compared to 9.4% for CNN baselines. The proposed system reduces facial landmark symmetry error by 18.7% over MediaPipe through physics-based correction of geometric inaccuracies.

Keywords: deepfake detection, facial landmarks, physics-informed learning, cross-dataset robustness, interpretable deep learning

1 Introduction

Deepfake methods range from simple face swaps to highly photorealistic GAN and diffusion-based forgeries [1]. They enable creative applications but are also misused for misinformation, impersonation, and biometric spoofing [2]. Though state-of-the-art detectors achieve over 94% accuracy on FaceForensics++, their performance drops to 68–79% on DFDC and Celeb-DF under cross-dataset evaluation [3]. This failure is not due to limited model capacity, but rather to reliance on superficial artifacts that do not hold under domain shift. However, most geometry-based detectors treat landmarks as generic points without enforcing facial anatomy, limiting cross-dataset robustness [4]. To address this limitation, we propose a physics-guided deepfake detector that shifts

* Corresponding Author : Soumitra Ghosh (soumitra.ghosh468@gmail.com)

S. Sharma et al. (eds.), *Proceedings of the International Conference on Recent Advances in Intelligent and Sustainable Technologies (RAIST 2026)*, Atlantis Highlights in Intelligent Systems 17,
https://doi.org/10.2991/978-94-6239-707-1_21

from pixel-level artifacts to anatomical geometry, employing a multi-stream CNN-GNN to extract 468 stable landmarks and a physics violation module that embeds anatomical priors. A hybrid classifier combines geometry and appearance using a geometry-aware loss to preserve anatomical plausibility and improve robustness over appearance-only and geometry-only baselines in both intra- and cross-dataset evaluations.

The main contribution of this work is a physics-informed deepfake detection framework that incorporates facial anatomy into the learning process. We model bilateral symmetry, contour smoothness, facial proportions, and regional consistency as geometric priors beyond conventional landmark representations and integrate them through an anatomy-guided CNN–GNN landmark refinement pipeline, where CNNs capture local appearance features and GNNs model relationships between landmarks to produce stable and anatomically consistent geometries. A geometry-aware loss enforces anatomical validity by constraining real faces within valid regions and separating fake faces through margin-based geometric violations rather than appearance-based losses, enabling robust hybrid geometry–appearance deepfake detection under cross-dataset shifts.

The rest of this paper is organized as follows: Section 2 introduces the related work on deepfake detection. Section 3 describes the proposed physics-informed approach in more detail. Section 4 introduces the experimental setup and discusses the results. The paper ends with Section 5 which presents the conclusion and future research directions.

2 Related Work

This section reviews existing deepfake detection approaches, focusing on appearance-based methods, geometry and landmark-based methods, cross-domain and self-supervised methods, and their limitations in cross-dataset generalization.

2.1 Appearance-Based Methods

Deepfake detection has been widely studied using computer vision techniques, but many methods fail to generalize despite strong in-domain performance [5]. CNN-based models achieve high accuracy on datasets such as FaceForensics++ and DFDC, but degrade on unseen manipulations due to reliance on dataset-specific artifacts [6]. Extensions using data augmentation and domain adaptation provide limited improvements without structural modeling [7]. Transformer-based models improve representation learning but remain appearance-centric and sensitive to domain shifts [8].

2.2 Geometry and Landmark-Based Methods

Geometry-based approaches exploit facial structure using landmark features, but often rely on handcrafted assumptions and show limited cross-dataset robustness [4]. Graph-based methods model landmark relationships using GNNs and fuse geometry with RGB and frequency cues, improving robustness over CNN baselines [9]. However, most approaches do not enforce anatomical or physics-based constraints, limiting stability and interpretability.

2.3 Cross-Domain and Self-Supervised Methods

Cross-domain methods improve generalization by learning invariant representations using RGB features [10]. Recent approaches incorporate robustness priors to reduce dataset bias [11]. Multitask self-supervised methods further enhance cross-dataset performance but remain largely appearance-driven [12]. Landmark-only detectors provide lightweight solutions but lack appearance cues for detecting subtle manipulations [13].

Existing methods either rely on RGB features, use landmarks without anatomical constraints, or separate geometry from appearance. This motivates the need for anatomy-aware, physics-informed models to improve cross-dataset robustness and reduce performance variance.

3 Proposed Method

This section presents the proposed physics-informed deepfake detection framework. We first describe the deepfake detection pipeline (Fig. 1) and problem setting, followed by dataset description, data preprocessing, unified geometry–appearance feature extraction, and finally the hybrid architecture with geometry-aware loss (Fig. 2).

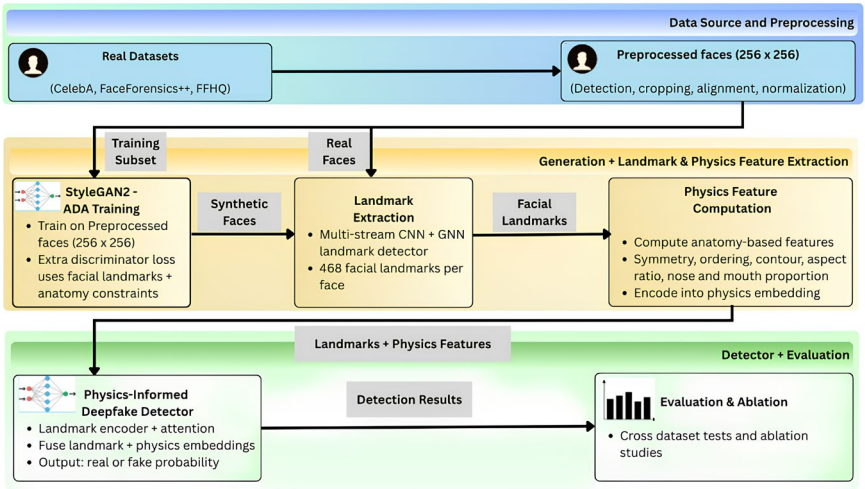

Fig. 1. Overview of the proposed physics-informed deepfake detection pipeline

Fig. 1 illustrates anatomy-guided landmark refinement, physics-based geometric features, and a hybrid detector that combines geometry and appearance cues.

3.1 Problem Setting

Let x be an image sequence and $y \in \{0,1\}$ indicate whether it is real or fake. A detector gradually acquires a function $f_\theta(x)$ that predicts the probability p ($y = 1 \mid x$). Following prior work that separates geometry and appearance, the input is divided into two streams: (i) an appearance stream consisting of aligned face crops and (ii) a geometric stream obtained through a robust pipeline that extracts 2D landmarks from every frame.

Human facial structure constrains landmark configurations, which generative models may violate in extreme poses and occlusions [4]. These constraints guide detectors toward attack-invariant features and improve interpretability via anatomy-based cues.

3.2 Dataset Description

The proposed framework operates on aligned facial images extracted from standard deepfake benchmarks, including FFHQ [14], CelebA [15], and FaceForensics++ (FF++) [16]. FF++ videos are converted to image-level samples by frame sampling, both real and fake data are processed through the same pipeline to avoid dataset bias.

3.3 Data Preprocessing

All face images are detected, aligned, and resized to 256×256. A multi-stream landmark extractor is used to obtain consistent 2D facial landmarks. Identical preprocessing is used for real and fake samples across all datasets. Data augmentation regularizes the appearance stream, while landmark noise regularizes the geometry stream.

3.4 Unified Geometry–Appearance Feature Extraction

Feature extraction combines geometric and appearance cues, where the geometry branch captures attack-independent structure and the appearance branch encodes detailed texture information not captured by geometry branch. From normalized facial landmarks $\{(\mathrm{xi}, \mathrm{yi})\}_{i=1}^{N}$, five anatomy-aware descriptors are computed: symmetry, measuring differences between corresponding points across the facial axis; ordering, capturing relative spatial arrangement of facial components (e.g., eyes above nose, nose above mouth); contour smoothness, estimating curvature consistency of jaw and cheek contours to detect distortions; proportions, modeling ratios such as mouth width to inter-eye distance, nose length to face height, and inter-eye distance to face width; and region consistency, enforcing stable relationships among eye, nose, and mouth regions under regardless of expression and pose variations. These geometry features are aggregated using frame-wise mean and variance to capture structural dynamics while remaining robust to lighting, background, and domain shifts. In parallel, aligned RGB face crops provide appearance features capturing subtle texture and frequency artifacts, and the unified geometry–appearance representation balances structural and visual cues for robust and interpretable detection across datasets and manipulations.

3.5 Hybrid Architecture and Geometry-Aware Loss

As shown in Fig. 2, the proposed framework jointly models facial appearance and anatomy-aware geometry using a hybrid two-branch architecture. The appearance branch processes aligned RGB face crops (256×256×3) using a ResNet-18 backbone with convolution (3×3), ReLU, and max-pooling layers, producing multi-scale feature maps (128×128×64 → 64×64×64 → 32×32×128 → 16×16×256 → 8×8×512) that are

flattened into an appearance feature vector for real/fake classification. In parallel, the geometry branch employs a CNN–GNN landmark refinement module, where a CNN extracts local features and a 3-layer GNN (256 channels) operates on 468 facial landmarks to model anatomical relationships and generate refined 2D coordinates (468×2). Shallow GNNs fuse geometry-branch and landmark-refinement embeddings for robust prediction, and the refined landmarks are converted into physics-based geometric descriptors, including bilateral symmetry, facial proportions, contour smoothness, landmark ordering, and region consistency, forming a compact geometry feature vector.

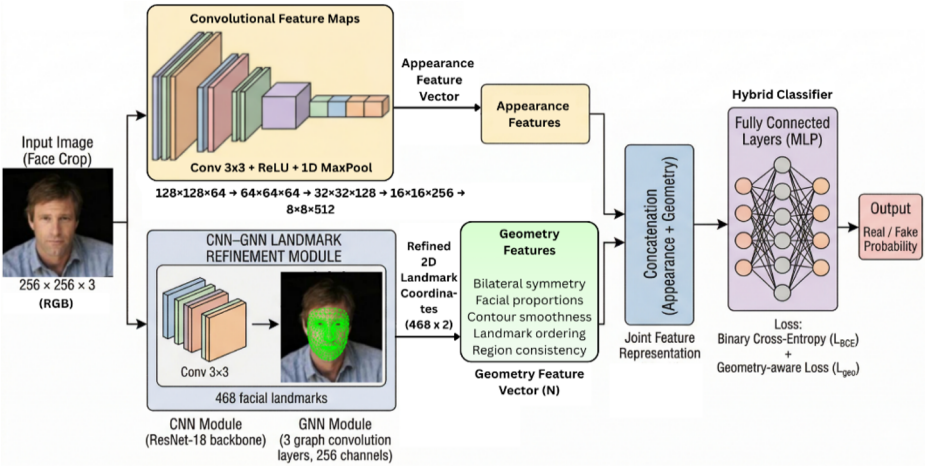

Fig. 2. Architecture of the proposed physics-informed deepfake detection framework

The appearance and geometry features are fused via feature concatenation along the feature dimension to form a joint representation, which is passed to a fusion MLP (fully connected layers) for final prediction. Training optimizes a hybrid loss:

$$L=L_{\text{BCE}}+\lambda L_{geo} \tag{1}$$

where L_{BCE} is the standard binary cross-entropy between the predicted probability $\hat{y}_i$ and the ground-truth label $y_i \in \{0,1\}$. Here $y_i=0$ for real faces and $y_i=1$ for fake faces:

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i log\hat{y}_i + (1-y_i)\log(1-\hat{y}_i)] \tag{2}$$

Let $z_i^g \in R^d$ denote the geometry embedding for sample i, where g indicates the geometry branch and d is the embedding dimension. The geometry-aware regularizer (L_{geo}) combines compactness for real faces with margin-based separation for fake ones:

$$L_{\text{geo}} = \frac{1}{N_{real}}\sum_{i:\, y_i=0}\sum_{k=1}^{d}(z_i^g - \mu_{real,k})^2 + \frac{1}{N_{fake}}\sum_{j:\, y_{j=1}} \max(0,\ m - ||\ z_j^g - \mu_{real}\ ||_2) \tag{3}$$

where μ_{real} is the mean geometry embedding over all real samples (label 0) in the mini-batch, and N_{real}, N_{fake} are the numbers of real and fake samples. The first term constrains real faces to an anatomically valid region, while the second pushes fake embeddings at least a margin m away. We set $\lambda = 0.1$ and $m = 0.5$ in our experiments.

4 Experiments

This section presents a comprehensive evaluation of the proposed physics-informed deepfake detection framework across multiple datasets and settings.

4.1 Datasets and Evaluation Protocol

We evaluate the proposed method on FFHQ [14], CelebA [15], and FF++ [16], with FF++ converted to image-level data by sampling video frames and fake images are generated using StyleGAN2. Each setting uses 50K images (25K real and 25K fake) with disjoint 70/10/20 train–validation–test splits, no identity overlap, and target-domain generators excluded to prevent data leakage [3], resulting in a total of 75K real and 75K fake images across cross-dataset settings. We consider both intra-dataset evaluation on FF++ to establish baseline performance, and cross-dataset evaluation where models are trained on FF++ and tested on FFHQ, CelebA, and FF++ manipulations to assess real/fake generalization [10], with 92.3% reported as the mean accuracy across all settings. Performance is evaluated using accuracy, AUC, F1-score, and EER, following deepfake benchmarks. These metrics follow deepfake benchmarking standards [17]. We also report accuracy standard deviation across datasets and manipulation types to quantify robustness under domain shift, beyond peak performance [12].

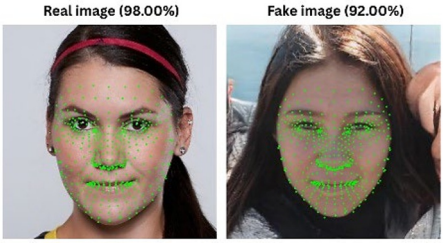

Fig. 3. Real FFHQ face (left) and GAN-generated fake face in the FFHQ domain (right)

Fig. 3 shows a real (left) and GAN-generated (right) FFHQ face (256×256). The detector predicts the real face as genuine with 98% confidence and the fake as forged with 92% confidence, showing anatomical cues can distinguish visually similar forgeries.

4.2 Baselines and Implementation Details

The proposed method is evaluated against representative baselines spanning appearance-based, geometry-based, and cross-domain approaches. The CNN (based on Xception) baseline relies on RGB face crops [8], and the transformer (DFDT) baseline follows a similar appearance-driven approach using vision transformers [18], both performing well in-domain but generalizing poorly due to reliance on visual artifacts. The two-stream RGB-frequency model attempts to improve this by incorporating frequency cues, while the graph-based baseline models landmark relationships through multimodal graph learning [9], though it lacks explicit anatomical constraints. A cross-domain baseline using deep information decomposition and multitask self-supervision

improves robustness across datasets [12], but remains largely appearance-driven. In contrast, our physics-informed framework incorporates anatomical geometry through a CNN–GNN pipeline and a geometry-aware loss, enabling it to capture stable facial structures and improve robustness and consistency under cross-dataset shifts.

The CNN–GNN landmark module uses ResNet-18 with three graph convolution layers (256 channels, ReLU). The hybrid detector combines 5-layer CNN and 3-layer MLP branch with a fusion MLP for prediction, trained using Adam with early stopping on validation AUC. On an NVIDIA T4 GPU, the CNN–GNN detector runs at 30–40 fps on 256x256 face crops, enabling near real-time moderation.

4.3 Quantitative Results

The CNN baseline performs strongly on FaceForensics++ (>94% accuracy) but drops sharply in cross-dataset evaluation [3]. The transformer baseline shows a similar generalization gap, despite stronger semantic modelling. In contrast, the graph-based baseline [9] generalizes better across datasets but is sensitive to compression, while the two-stream model shows high variance due to overfitting to frequency-specific artifacts.
Our physics-informed detector outperforms the CNN baseline in cross-dataset settings (Table 2), achieving higher accuracy and stability across FFHQ↔CelebA↔FF++ transfers. Unlike baselines that degrade on diffusion-based or lightly compressed data [3], our method remains robust by leveraging geometric consistency rather than dataset-specific artifacts. Remaining errors occur under extreme poses and occlusions due to degraded landmark quality.

Table 1. Cross-dataset deepfake detection: comparison with baseline and SOTA methods

Method	Type	Training protocol	Evaluation Metrics	Key idea
Xception / CNN baseline (FF++ , 2020)	Appearance (CNN)	FF++ → Celeb-DFv2 evaluation following the Celeb-DF protocol [3]	Celeb-DFv2: AUC ≈ 0.73; accuracy ≈ 73–75%	Strong on FF++ (>94% accuracy) but weak on Celeb-DFv2 due to codec artifacts.
DFDT (Deepfake Detection Transformer, 2022)	Appearance (Transformer)	Trained on FF++ and evaluated on aligned crops from Celeb-DF v2 and DFDC [18]	Celeb-DFv2: AUC ≈ 0.80; DFDC: AUC ≈ 0.78	ViT increases cross-dataset AUC using long-range RGB cues while preserving appearance-centricity.
Cross DF (Deep Information Decomposition, 2025)	Cross-domain generalization	Train on FF++, test on Celeb-DFv2 and others to remove deepfake signals from identity and domain factors [10]	FF++ → Celeb-DFv2: AUC ≈ 0.78 (vs ≈0.67 for vanilla CNN)	Reduces dataset-specific overfitting, improving cross-domain robustness.
Multitask self-supervision (2025)	Cross-domain generalization	Multitask self-supervised training on FF++, cross-dataset evaluation on Celeb-DFv2 and DFDC [12]	Higher cross-dataset AUC than CNN baselines on Celeb-DFv2 and DFDC	To stabilize representations during the domain shift, auxiliary self-supervised tasks are employed.

ViT / ensemble frameworks in surveys (2022–2024)	Appearance / hybrid	Benchmarked across FF++, Celeb-DFv2, and DFDC in comparative and survey studies [8]	Typical cross-dataset AUC ranges from ≈0.78–0.85 depending on dataset pair	Compared to CNN baselines, transformers and ensembles dramatically reduce cross-dataset gaps.
Ours – Physics-informed landmark-centric detector	Hybrid Geometry and Appearance Model	Training uses FFHQ, CelebA, and FF++ with cross-dataset protocols, with 25k real and 25k fake images per dataset (i.e 75k real, 75k fake)	Mean cross-dataset accuracy ≈ 92.3% across-FFHQ↔CelebA↔FF++ transfers.	CNN–GNN landmark refinement with physics-based features and a hybrid classifier for stable anatomical patterns.

Table 1 shows that appearance-based CNN and transformer models degrade under cross-dataset evaluation, while cross-domain methods improve robustness; the proposed physics-informed approach achieves the highest mean cross-dataset accuracy.

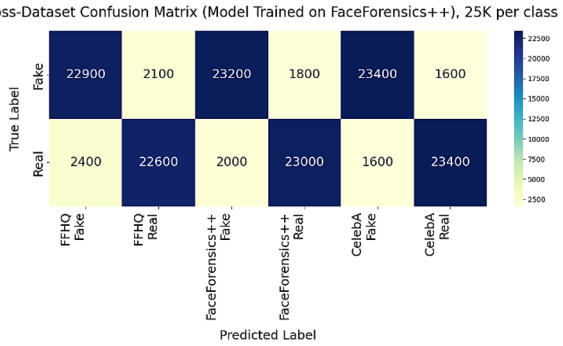

Fig. 4. Combined Cross-Dataset Confusion Matrix Heatmap (75K real and 75K fake images)

Fig. 4 shows balanced true positive and true negative rates across all cross-dataset settings, indicating stable generalization rather than dataset-specific overfitting.

4.4 Ablation Studies and Analysis

The ablation study shows that anatomy-aware features and geometry-aware loss improves robustness. As shown in Table 2, using anatomy-aware geometry increases cross-dataset accuracy from 81.1% to 92.3% while reducing variance from 9.4% to 2.1%, confirming the robustness of the proposed hybrid model. Removing the geometry regularizer degrades cross-dataset robustness despite similar in-domain accuracy, while hybrid model consistently outperforms geometry-only and appearance-only variants.

We evaluate the robustness of the geometry-aware loss by varying $\lambda \in \{0.05, 0.1, 0.2\}$ and $m \in \{0.3, 0.5, 0.7\}$, observing similar cross-dataset accuracy and variance across all settings. This verifies that the choice of $\lambda = 0.1$ and $m = 0.5$ is not highly tuned.

Table 2. Cross-Dataset Robustness Across FFHQ, CelebA, and FaceForensics++

Method	Accuracy (%)	Std. Deviation (%)
CNN baseline (appearance only)	81.1%	9.4%
Proposed hybrid (geometry + appearance)	92.3%	2.1%

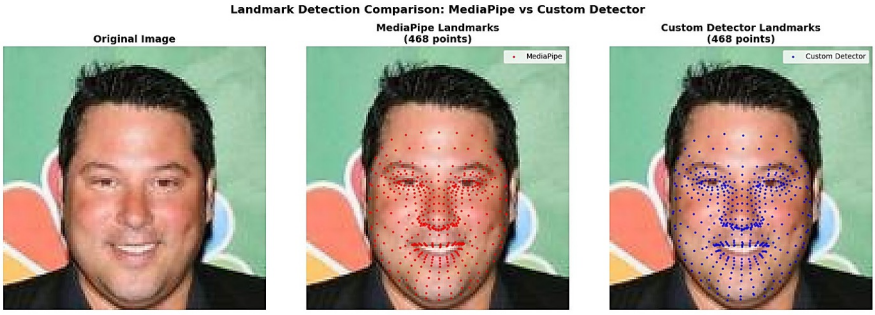

Fig. 5. Qualitative comparison of MediaPipe (middle) and proposed CNN–GNN (right) landmark predictions on a CelebA sample (left)

Table 3. Landmark Quality Comparison: MediaPipe vs. our proposed CNN–GNN Detector

Metric	MediaPipe (reference)	Ours (CNN–GNN)
Landmarks	468	468
MAE vs MediaPipe (normalized)	–	0.0029
MAE vs MediaPipe (pixels)	–	0.37
Coverage area	0.3311	0.3301

As shown in Table 3 and Fig. 5, our CNN–GNN landmark model maintains MediaPipe-level coverage while reducing landmark error and improving facial contour adherence under pose and compression, enhancing geometric cues for deepfake detection.
The errors in cross-dataset are mainly due to extreme poses and occlusions, where the visibility of the landmarks is restricted, and geometric information is more dependent on appearance features. The saliency map focuses on anatomically inconsistent regions rather than background noise, suggesting future work on occlusion-aware modeling.

5 Conclusion

This paper presented a physics-informed deepfake detection system that incorporates anatomy-aware facial geometry into a geometry-appearance hybrid model, representing structural constraints in both representation and loss to prefer anatomical invariants over pixel artifacts. The system achieves improved cross-dataset accuracy with reduced variance on FFHQ, CelebA, and FaceForensics++, demonstrating effective handling of domain shifts and narrowing the performance gap between in-domain and cross-domain testing across manipulation techniques.
However, the current evaluation is limited to these datasets, and performance still degrades under extreme pose, heavy occlusion, and strong motion blur, highlighting the need for better temporal modeling and explicit 3D facial geometry. Specifically, deepfakes exhibit frame-to-frame flickering in anatomical ratios not modeled by the current frame-level approach; incorporating temporal physics constraints over landmark trajectories could further regularize geometry and improve robustness for video deepfakes. Future research will extend anatomy-aware constraints to 3D meshes, depth, and temporal information, enhance robustness against adversarial geometric attacks that

deliberately manipulate landmarks while preserving forged identities, and provide geometry-based explainability tools for safety-critical applications like biometric authentication and large-scale content moderation.

References

1. Heidari, A., et al.: Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 14(2), e1520 (2024)
2. Rehaan, M., Kaur, N., Kingra, S.: Face manipulated deepfake generation and recognition approaches: A survey. Smart Science 12(1), 53–73 (2024)
3. Nadimpalli, A.V., Rattani, A.: On improving cross-dataset generalization of deepfake detectors. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 91–99 (2022)
4. Liang, P., et al.: A facial geometry based detection model for face manipulation using CNN-LSTM architecture. Information Sciences 633, 370–383 (2023)
5. Yan, Z., et al.: DeepfakeBench: A comprehensive benchmark of deepfake detection. arXiv preprint arXiv:2307.01426 (2023)
6. Belguesmia, S., Allili, M.S., Hamadene, A.: Unmasking facial deepfakes: A robust multiview detection framework for natural images. arXiv preprint arXiv:2510.15576 (2025)
7. Deng, J., et al.: Towards benchmarking and evaluating deepfake detection. IEEE Transactions on Dependable and Secure Computing 21(6), 5112–5127 (2024)
8. Le, B.M., Kim, J., Woo, S.S., Moore, K., Abuadbba, A., Tariq, S.: SoK: Systematization and benchmarking of deepfake detectors in a unified framework. In: Proc. IEEE European Symposium on Security and Privacy (EuroS\&P), pp. 883–902 (2025)
9. Yan, Z., et al.: Multimodal graph learning for deepfake detection. arXiv preprint arXiv:2209.05419 (2022)
10. Yang, S., Guo, H., Hu, S., Zhu, B., Fu, Y., Lyu, S., Wu, X., Wang, X.: CrossDF: Improving cross-domain deepfake detection with deep information decomposition. Frontiers in Big Data 8, 1669488 (2025)
11. Yermakov, A., Cech, J., Matas, J., Fritz, M.: Deepfake detection that generalizes across benchmarks. In: Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 773–783 (2026)
12. Batagelj, B., Kronovšek, A., Štruc, V., Peer, P.: Robust cross-dataset deepfake detection with multitask self-supervised learning. ICT Express 11(5), 858–862 (2025)
13. Carter, B., et al.: Deepfake detection via facial feature extraction and modeling. arXiv preprint arXiv:2507.18815 (2025)
14. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4401–4410 (2019)
15. Liu, Z., et al.: Deep learning face attributes in the wild. In: Proc. IEEE International Conference on Computer Vision (ICCV), pp. 3730–3738 (2015)
16. Rössler, A., et al.: FaceForensics++: Learning to detect manipulated facial images. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1–11 (2019)
17. Altuncu, E., Franqueira, V.N.L., Li, S.: Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review. Frontiers in Big Data 7, 1400024 (2024)
18. Khormali, A., Yuan, J.-S.: DFDT: An end-to-end deepfake detection framework using vision transformer. Applied Sciences 12(6), 2953 (2022)