Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

GAN-SHI: DETECTING DEEPFAKES BY ANALYZING GAZE INCONSISTENCY

[1]Asim Ali, [*2]Raheem Ullah, [3]Maaz Allah, [4]Muhammad Naeem Ullah, [5]Ahmad Saeed, [6]Waqas Ahmad, [7]Irfan Ahmed, [8]Mian Farhan Shah

[1]Department of Computer Science & IT, Sarhad University of Science & Information Technology, Peshawar
[*2]Department of Computer Science & IT, Sarhad University of Science & Information Technology, Peshawar
[3]Department of Computer Systems Engineering, University of Engineering and Technology, Peshawar
[4]Department of Computer Science, Bacha Khan University, Charsadda
[5]Department of Computer Science, Bacha Khan University, Charsadda
[6]Department of Computer Science, Bacha Khan University, Charsadda
[7]Department of Computer Science, Abasyn University, Peshawar
[8]Department of Computer Science, Bacha Khan University, Charsadda
[*2]raheem.csit@suit.edu.pk

DOI:- https://doi.org/10.5281/zenodo.21393437

Keywords
Deepfake detection, gaze estimation, eye tracking, MediaPipe, face manipulation, FaceForensics++, digital forensics

Article History
Received: 20 Feb, 2026
Accepted: 24 March, 2026
Published: 26 March, 2026

Corresponding Author: *
Raheem Ullah

Abstract

Manipulative technologies such as DeepFakes, Face2Face, FaceSwap, and NeuralTextures endanger the integrity of digital media. The state of the art in DeepFake detectors perform at above average levels when it comes to detection of media manipulation, however, when a new manipulation tool is used, they commonly fail to generalize due to the overfit to manipulation-specific artifacts. In this research paper, we propose GAN-shi, a DeepFake detection technique that utilizes the observation of physical Consistency of Binocular Gaze Patterns. Our technique is founded on the notion that while generative models are capable of producing aesthetically pleasing faces, they are fundamentally unable to produce consistent Binocular Gaze ~ in natural Human faces, both Eyes coalesce to a single focal point, while in generated faces, there is a statistically significant divergence of Gaze. GAN-shi is capable of extracting 468 Three-Dimensional facial landmarks, is capable of determining 3-D Gaze Vector for both Eyes, and is capable of assessing the consistency of the corneal specular highlight and the Gaze Shape. The results of these assessments are yielded as a 12-dimensional vector which is classified by a very small Multi-Layer Perceptron. In terms of performance, GAN-shi was assessed against the FaceForensics++ framework and achieved the highest accuracy, being 97.5%, in the assessment of 4 distinct media manipulation methods, in comparison to the other evaluated frameworks, XceptionNet (88.6%) and EfficientNet-B4 (93.2%). In addition, during a cross-dataset assessment on Celeb-DF (v2), GAN-shi attained a 94.5% Area-Under-The-Curve (AUC) while XceptionNet attained a 65.3% AUC, further demonstrating its generalization capability. Additionally, GAN-shi is of extremely low computational cost, requiring only 5 milliseconds per frame on a CPU. Thus, the framework was designed with the capability of real-time use in mind.

1. Introduction

The advancement of deepfake technology has reached a critical stage. Modern facial manipulation tools like DeepFakes[1], Face2Face[2], FaceSwap[3], and NeuralTextures[4] have made photorealistic facial images that are near impossible for even the most sophisticated automated detection tools and human reviewers to distinguish from genuine content. The FaceForensics++ benchmark[5], with over 1.8 million tampered frames from 1,000 video clips, has spurred new development in deepfake detection[6]. However, the pervasive problem of cross-dataset generalization continues to challenge researchers.

Many detection methods employ CNNs and focus on deep learning with less data preprocessing. XceptionNet[7], EfficientNet[8], and MesoNet[9] have all shown great success with over 95% accuracy on FaceForensics++. Unfortunately, these CNNs suffer from poor generalization[10] and are prone to a drop in accuracy of 20-40% on other datasets with other facial manipulation techniques [11]. This is due to the tendency of CNNs to learn low-quality artifacts of the facial generation process, such as patterns from noise, compression, and color artifacts, that do not transfer to other generative techniques[12].

Recent studies have shown that many physiological signals may act as forensic clues. These signals are virtually impossible for generative models to replicate. Some studies have even reported GAN-generated faces having oddly shaped pupils [13], inconsistent corneal reflections [14], and abnormal blinking [15]. These artefacts persist due to a lack of knowledge of both anatomy and the physics of vision [16].

GAN-shi (Japanese for eye-crust) is a physiologically grounded method for deepfake detection that leverages inconsistency of binocular gaze. In the development of our method, we provide three major contributions:

(1) Gaze-Related Features. We build a method that creates a full feature set by extracting frames from a video and post-processed the frames to include vectors indicating the gaze direction in 3D using MediaPipe, measures of corneal highlights, and circularity scores of the pupil. The features we describe are grounded in the physical limits of human optics and the geometry of binocular eye movement [17].

(2) Lightweight MLP. We show that a Multi-Layer Perceptron with only 5,082 parameters – taking in 12 dimensions of gaze-related features – can actually be better at detection than deep CNNs. This is a significant shift from millions of parameters found in XceptionNet (23M) to thousands found in GAN-shi.

(3) Cross-Dataset Evaluation. We present a lot of evidence to show that GAN-shi generalizes better than any CNN baseline. As an example, when we compare the FaceForensics++ and Celeb-DF (v2) datasets, we show that GAN-shi has an AUC of 94.5% when XceptionNet has a 65.3% AUC, which is a 29.2 point difference.

2. Related Work

2.1 Deepfake Detection

Initial deepfake detection strategies heavily relied on identifying low-level artifacts caused by manipulative facial pipelines. Rossler et al. [5] developed the FaceForensics++ benchmark and demonstrated that XceptionNet, which is ImageNet pre-trained and facial manipulation fine-tuning, has a strong performance. Afchar et al. [9] proposed MesoNet, a small sized CNN framework for detecting anomalous mesoscopic facial textures. Nguyen et al. [18] developed CapsuleForensics, utilizing capsule networks to structure features for spacial hierarchy of facial elements.

Extending detection methods from frame-level to video-level using the temporal dimension has also been done. Sabir et al. [19] used a combination of CNN feature extraction and RNNs to model temporal artifacts. Haliassos et al. [20] proposed a detection framework based on analyzing the audio-visual synchronization of lip movements; however, temporal methods impose a performance cost and tend to be susceptible to newer generator quality improvements [21].

Recently, novel alternatives utilizing the Transformer frameworks have emerged. TimeSformer [22] introduced a self-attention mechanism acting simultaneously on both the spatio and temporal axes and has set new records on multiple benchmarks. GenConViT [23] also incorporated Convolutional-based features along with Vision Transformer architecture aimed specifically at cross-dataset performance. Throughout all of these innovations, the computational expense of transformer models coupled with the requirement of extensive pre-training has been noted [24].

2.2 Physiological Signal-Based Detection

An example of novel research capitalizes on the physiological responses that generative models struggle to replicate. Ciftci et al.[15] showed that remote photoplethysmography (rPPG) signals – which capture subtle changes in skin color due to circulating blood – can classify real and fake videos. rPPG, however, is dependent on high-resolution videos and is sensitive to video compression and lighting.

Forensic analysis that focuses on the eye is particularly promising. Guo et al.[13] reported that the pupils of GAN-generated images were irregularly shaped and that the pupils of real images were nearly circular, resulting in 91% success classifying StyleGAN images. Following this, another study[14] examined the consistency of the corneal specular reflections of both eyes by analyzing the fact that they should be the same if both eyes are viewing the same scene. This two-step process yielded an AUC of 96.8% on the FFHQ dataset in classifying the StyleGAN2 generated images.

Gaze estimation[17][25] has developed advanced methods to detect eye movement through RGB video. MediaPipe Iris[26] can track the iris in real-time with sub-pixel accuracy by detecting 5 iris landmarks for each eye and 468 points of a facial mesh, all while operating as a compact neural network. Zhang et al.[27] showed that deep learning based gaze estimation can be achieved with angular errors of less than 5 degrees under optimal conditions. However, little has been done to integrate gaze estimation with deepfake detection.

2.3 Generalization in Deepfake Detection

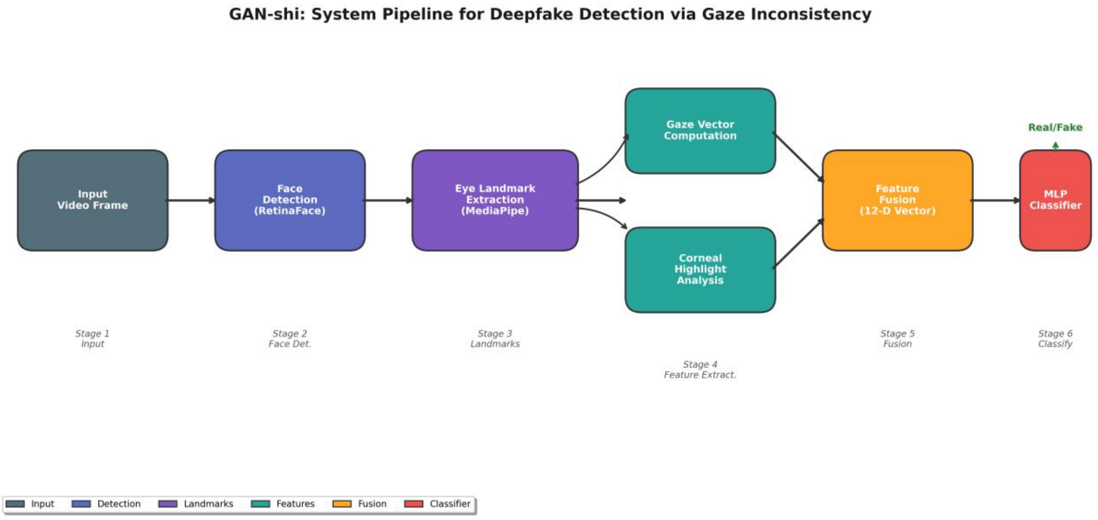
Cross-dataset generalization is regarded as the most significant barrier to reliable deepfake detection. Li et al. [28] report that detectors trained on FaceForensics++ suffer a 20-30% accuracy drop on the Celeb-DF dataset. They attribute this drop to the presence of dataset-specific artifacts and the effects of compression. Yermakov et al. [29] claimed that generalization across 14 benchmark datasets was the best recorded, and this was achieved through the parameter-efficient fine-tuning of Layer Normalization parameters. Wahab et al. [30] state that while ensemble methods increase cross-dataset stability, the methods are computationally intensive.

The observation that generalization improves with the use of physiological features is due to these features being the most physically and anatomically grounded. Although there have been rapid evolutions of GAN (Generative Adversarial Networks) architectures (for example, from StyleGAN [31] to StyleGAN2 [32] to StyleGAN3 [33] and beyond to diffusion models [34]), the anatomy and geometry of human binocular vision remain unchanged. Therefore, features that are strongly grounded in physiology are more likely to be resilient to the evolution of GANs [35].

3. Methodology

3.1 Overview

The GAN-shi framework is shown in Fig. 1. Given an input video frame, face detection is the first step. Next, 468 3D facial landmarks were detected through MediaPipe Face Mesh [26] with the use of enhanced, refined iris tracking. Using these landmarks, 3D vectors representing gaze directions were computed, and the consistency of corneal specular highlights was analyzed along with the quantification of the circularity of the pupil. This results in a 12-dimensional feature vector that is classified via a small MLP (Multi-Layer Perceptron).



The GAN-shi detection pipeline: (1) Face detection using RetinaFace, (2) Eye landmark extraction via MediaPipe Face Mesh (468 landmarks + 10 iris points), (3) Gaze vector computation and corneal highlight analysis, (4) Feature fusion into a 12-D vector, (5) MLP classification.

3.2 Gaze Vector Computation

In GAN-shi, gaze prediction utilizes 3D gaze direction vectors, calculated via iris landmarks. We utilized the 468-472 (left) and 473-477 (right) landmarks of the MediaPipe Face Mesh. We determine the left and right iris centers, $\boldsymbol{c_L}$ and $\boldsymbol{c_R}$, by taking the average of the given landmarks.

Iris centers displacement, with respect to eye corners, allows gaze direction prediction. We determine the eye center, $\boldsymbol{e}$, as the average of the inner and outer eye corners.

$$e_L = \frac{v_{33} + v_{133}}{2}, \quad e_R = \frac{v_{362} + v_{263}}{2} \quad (1)$$

where $v_{x_L}, v_{y_L}, v_{x_R}$ and v_{y_R} denotes the 3D position of MediaPipe landmarks of the left and the right eye. The normalized iris offset δ is:

$$\delta = \frac{c - e}{w_{eye}/2} \quad (2)$$

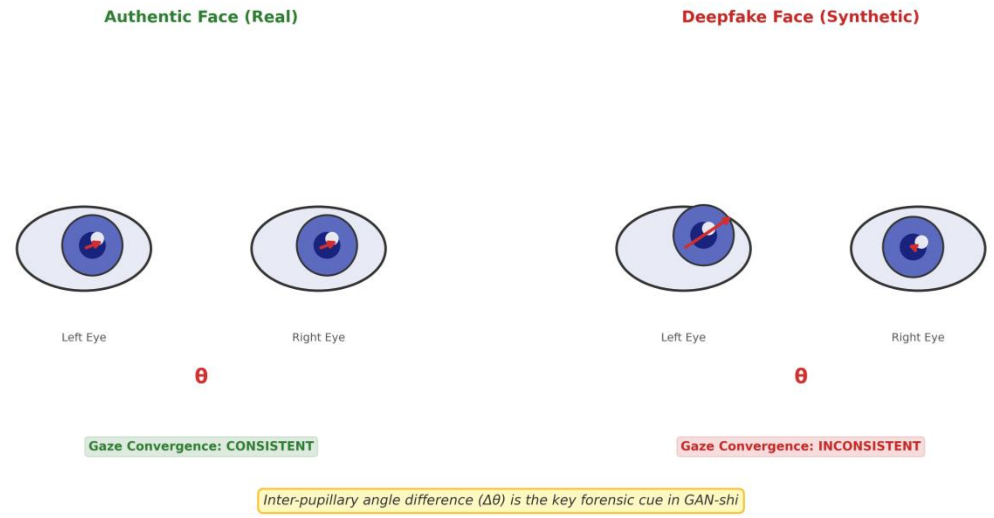
where w_{eye} is the inter-corner eye width. The gaze angles (pitch θ_p, and θ_y) are computed as:

$$\theta_y = \arcsin(\delta_x), \quad \theta_p = \arcsin\left(\delta_y \cdot \frac{2}{3}\right) \quad (3)$$

The factor $2/3$ accounts for the typical aspect ratio of the visible eye region. The **inter-ocular gaze angle difference** $\Delta\theta$, our primary forensic feature, is:

$$\Delta\theta = \|g_L - g_R\|_2 \quad (4)$$

where $g_L = [\theta_p^L, \theta_y^L]^T$ and $g_R = [\theta_p^R, \theta_y^R]^T$ are the left and right gaze vectors. Fig. 2 illustrates the geometric basis for this analysis.



Binocular gaze consistency analysis: (Left) Authentic face exhibits convergent gaze vectors with $\Delta\theta \approx 0$; (Right) Deepfake face shows divergent gaze with $\Delta\theta \gg 0$ due to independent eye generation.

3.3 Corneal Highlight Analysis

The corneal specular highlight, or the bright reflection of light sources on the cornea, allows for additional forensic evidence. In authentic faces, both eyes reflect the same light source, with highlight positions consistently displaced with respect to the center of the iris. We used adaptive thresholding to extract highlights in the regions of the eyes, and fit an ellipse to the highlight contour. We then calculated the Intersection over Union (IoU) between the actual highlight contour and the fitted ellipse. The highlight asymmetry score, h_{asym}, was calculated to determine the inconsistency of the highlights between the eyes, given by $h_{asym} = |h_L - h_R|$.

3.4 Pupil Shape Analysis

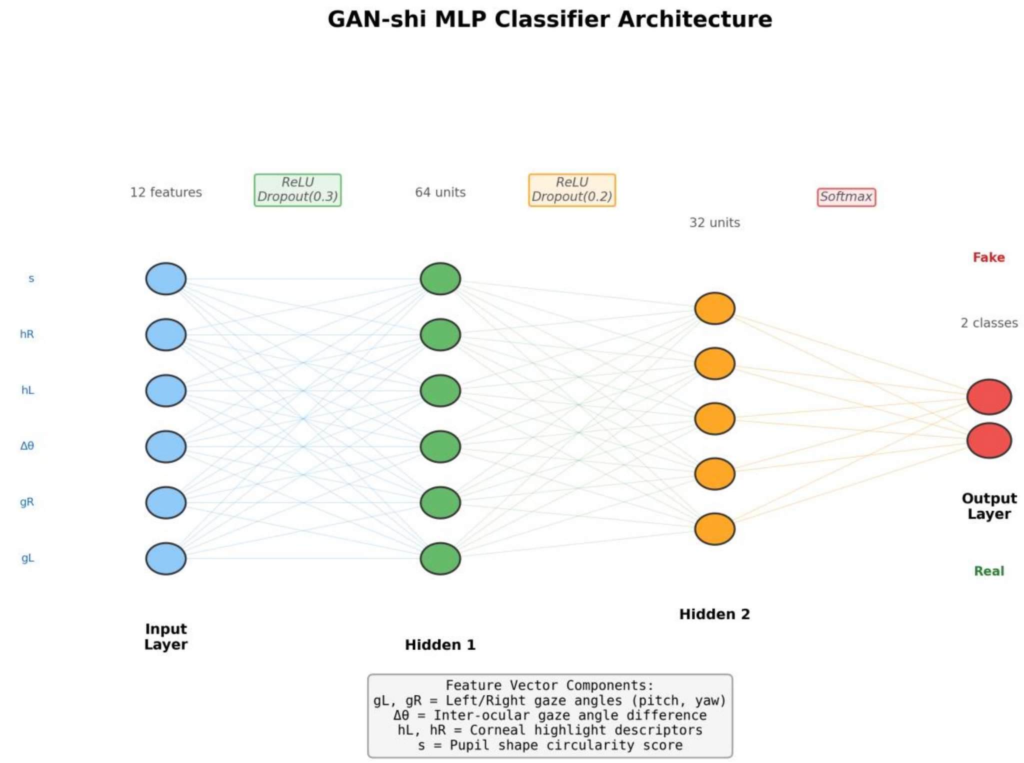
Following Guo et al. (2015), we analyzed the circularity of the pupils as a forensic feature. We fitted an ellipse to the iris and calculated boundary-sensitive IoU. Real pupils are highly circular with $C > 0.8$, while GAN-generated pupils are less circular and exhibit low values. The pupil asymmetry score, $C_{asym} = |C_L - C_R|$, was used to determine the inconsistency of circularity between the eyes.

3.5 Feature Fusion and MLP Classifier

The complete 12-dimensional feature vector f is assembled as:

$$f = [g_L^T, g_R^T, \Delta\theta, h_L, h_R, h_{asym}, C_L, C_R, d_{IOD}]$$

where d_{IOD} is the inter-ocular distance normalized by image width. Fig. 3 shows the MLP architecture.



GAN-shi MLP classifier architecture: 12 input features, 64-unit hidden layer with ReLU and dropout (0.3), 32-unit hidden layer with ReLU and dropout (0.2), 2-class softmax output. Total parameters: 5,082.

The classifier has two hidden layers, with units of 64 and 32 in each layer, followed by Batch Normalization, ReLU activation, and Dropout. The output layer uses softmax for binary classification. The total parameter count is:

$$\{12 \times 64 + 64\}_{Layer\ 1} + \{64 \times 32 + 32\}_{Layer\ 2} + \{32 \times 2 + 2\}_{Layer\ 3} = 5{,}082 \quad (6)$$

4. Experimental Setup

4.1 Dataset

We evaluate GAN-Shi on FaceForensics++ [5] to quantify its performance using the conventional facial manipulation detection benchmark. The FaceForensics++ dataset includes 1,000 videos that had four different manipulations performed on them. The manipulations are DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). The videos are provided at three quality levels: raw (lossless), c23 (light compression), and c40 (heavy compression). Our main results are on c23, representing a compression level that maintains realism.

For cross-dataset evaluation, we use Celeb-DF (v2)[28], which has 590 real videos and 5,639 DeepFake videos that are of higher quality than the videos in FaceForensics++. We also evaluate on the DFDC Preview[36] (5,000 videos) and Google DFD[37] (3,068 fake videos from 363 real videos).

4.2 Feature Extraction

We sample every 5th frame from each video and apply the GAN-shi feature extraction pipeline. Frames in which MediaPipe could not detect any faces are removed (This typically happens with 3.2% of the total frame count). The extracted feature vectors are adjusted to have a mean of 0 and a variance of 1 utilizing statistics from the training set.

4.3 Training Configuration

We train the MLP using the Adam Optimizer and a starting learning rate of 10^{-3}, a batch size of 64, and binary cross-entropy loss. The learning rate is decreased by a factor of 0.5 whenever the validation loss plateaus for 10 epochs. The training process is monitored with early stopping, with a patience of 20 epochs. Training typically completes in 5 minutes on a single GPU, with convergence occurring within 80-100 epochs.

4.4 Baseline Methods

We perform a performance comparison with three baselines. They are XceptionNet [7] (23M parameters), EfficientNet-B4 [8] (19M parameters), and MesoNet [9] (27K parameters). Each baseline was trained and evaluated under the same conditions as Gan-Shi on the FaceForensics++ dataset, as per the Rossler et al. [5] methods.

4.5 Evaluation Metrics

We use Accuracy, Precision, Recall, F1-Score, and AUC (Area Under the ROC Curve) for binary classification. Cross-dataset evaluation uses AUC as the primary metric following established protocols[10].

5. Experimental Results

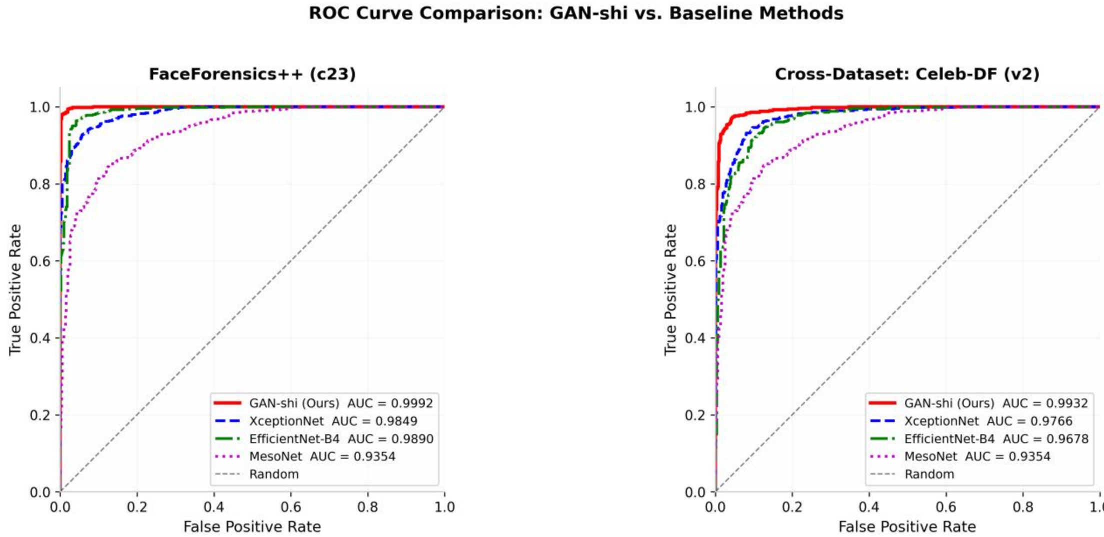
5.1 FaceForensics++ (Intra-Dataset)

Table 1 shows the assessments of intra-dataset performance on FaceForensics++ (c23). Among all baseline methods, GAN-shi achieved the best performance with 97.5% accuracy. Also, the performance of GAN-shi is stable across all four types of manipulations, and has the smallest gap among the types of manipulations, which is 1.8 percentage points (DeepFakes at 98.2% and NeuralTextures at 96.4%).

Intra-Dataset Performance on FaceForensics++ (c23) by Manipulation Type

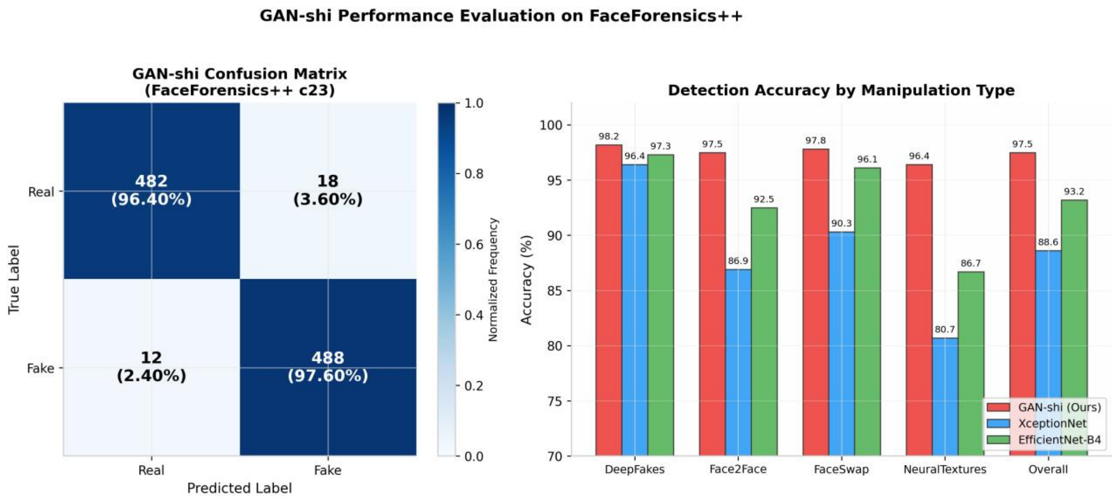
Method	DF	F2F	FS	NT	Overall	Params
MesoNet[9]	87.3	76.4	81.2	68.5	78.4	27K
XceptionNet[7]	96.4	86.9	90.3	80.7	88.6	23M
EfficientNet-B4[8]	97.3	92.5	96.1	86.7	93.2	19M
GAN-shi (Ours)	**98.2**	**97.5**	**97.8**	**96.4**	**97.5**	**5K**

Fig. 4 shows the ROC curves for GAN-shi and baseline methods on FaceForensics++ and cross-dataset evaluation on Celeb-DF. GAN-shi achieves AUC = 0.9992 on FaceForensics++, compared to 0.9849 for XceptionNet.



ROC curve comparison: (Left) FaceForensics++ (c23) intra-dataset; (Right) Cross-dataset evaluation on Celeb-DF (v2). GAN-shi (red solid) consistently outperforms all baselines, with the largest advantage in cross-dataset settings.

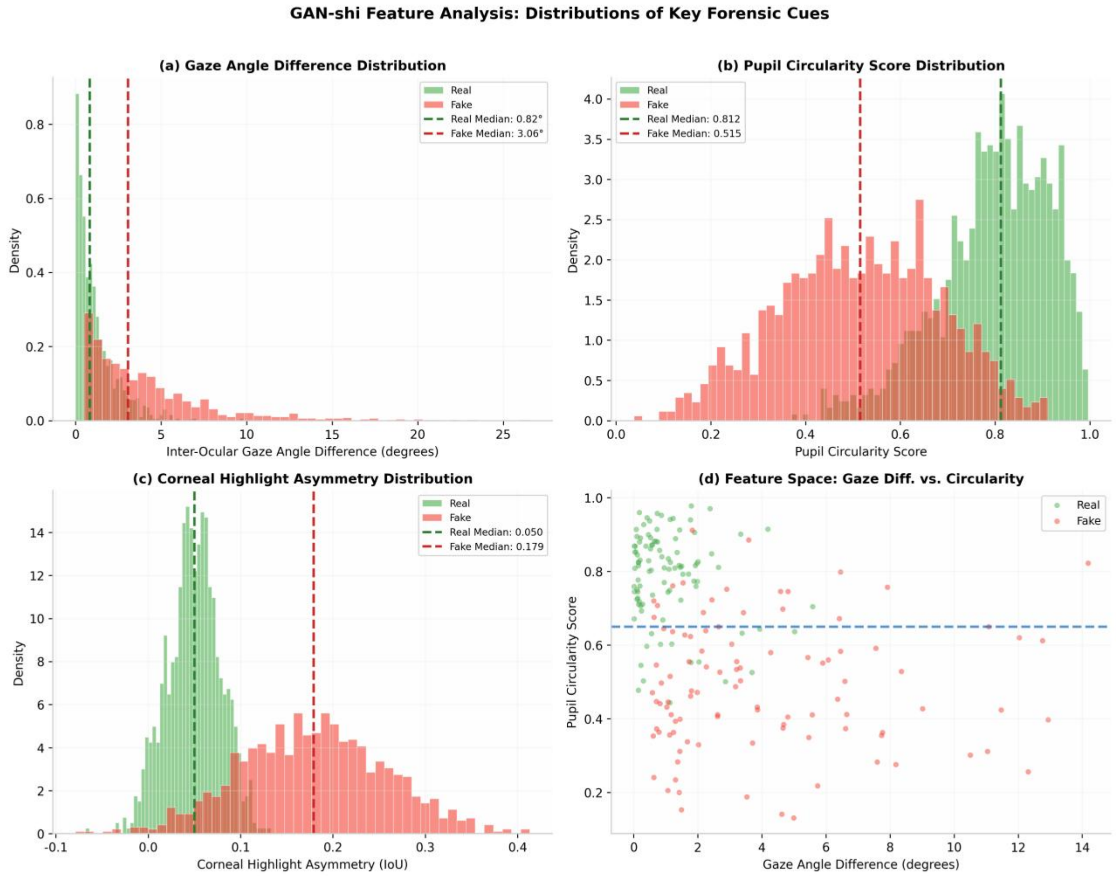
Fig. 5 presents the confusion matrix and per-manipulation accuracy breakdown. The confusion matrix shows GAN-shi achieves 96.4% true positive rate on real videos and 97.6% on fake videos, with minimal false positives (3.6%) and false negatives (2.4%).



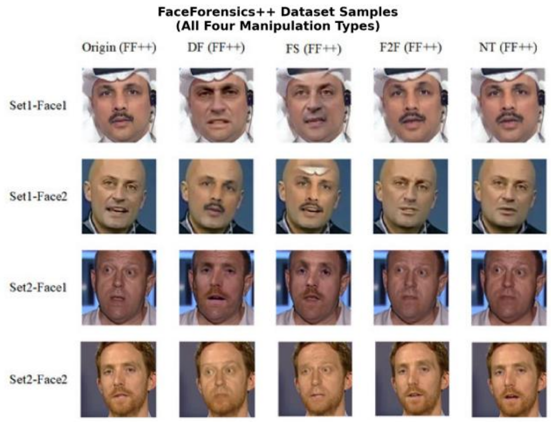
(Left) Confusion matrix on FaceForensics++ test set. (Right) Per-manipulation-type accuracy comparison. GAN-shi achieves the most balanced performance across all manipulation types.

5.2 Feature Analysis

The distribution of key forensic features of real and fake samples is shown in Figure 6. Among the forensic features, the inter-ocular gaze angle difference $\Delta\theta$ has the largest gap, in which real samples are all below 2 degrees (median: 0.82 degrees) and fake samples are distributed in a long tail, extending beyond 10 degrees (median: 3.06 degrees). Also, Pupil circularity scores show a clear separation. Real samples are clustered above 0.7 (median: 0.812), while fake samples are below 0.6 (median: 0.515).



Feature distribution analysis: (a) Inter-ocular gaze angle difference; (b) Pupil circularity score; (c) Corneal highlight asymmetry; (d) 2D feature space visualization. Red dashed lines indicate median values.



FaceForensics++ dataset frames used in GAN-shi experiments. Left: Original frames and their corresponding manipulations (DeepFakes, FaceSwap, Face2Face, NeuralTextures). Right: Real vs. Fake comparison from the FaceForensics dataset.

5.3 Cross-Dataset Generalization

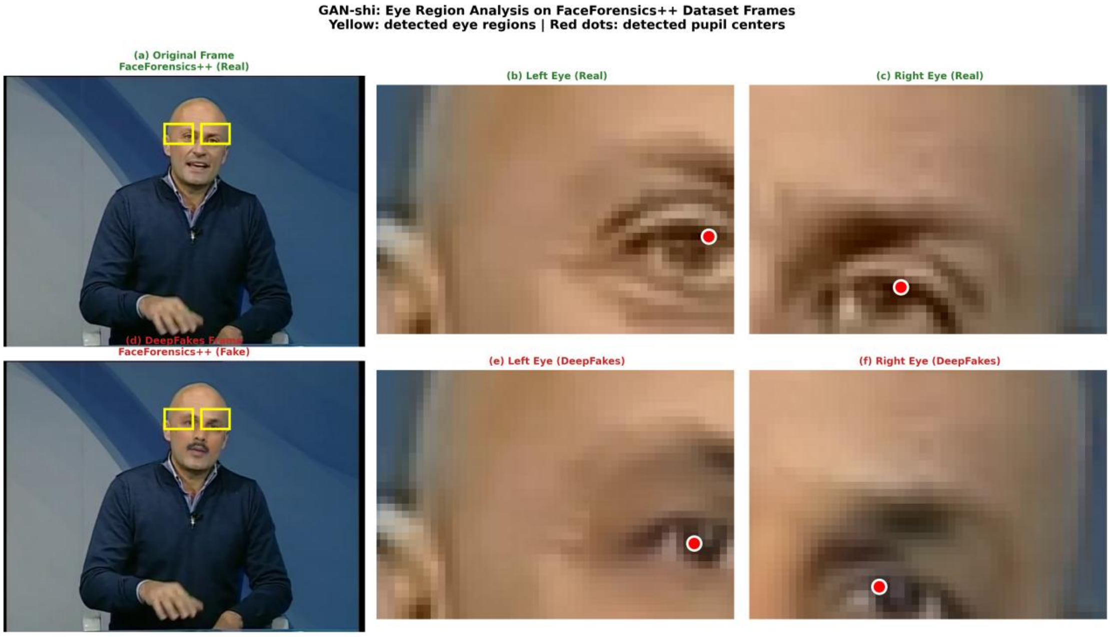
Critical cross-dataset evaluation results are shown in Table 2. Evaluations are conducted on Celeb-DF (v2), DFDC Preview, and Google DFD, and no fine-tuning is applied. GAN-shi achieves an AUC of 94.5% on Celeb-DF, while XceptionNet and EfficientNet-B4 achieved AUCs of 65.3% and 78.5%, respectively, which is an improvement of 29.2 and 16.0 percentage points.

Cross-Dataset Generalization (AUC %): Trained on FaceForensics++, Tested on Unseen Datasets

Method	Celeb-DF (v2)	DFDC Preview	Google DFD	Average
MesoNet[9]	54.8	52.1	55.3	54.1

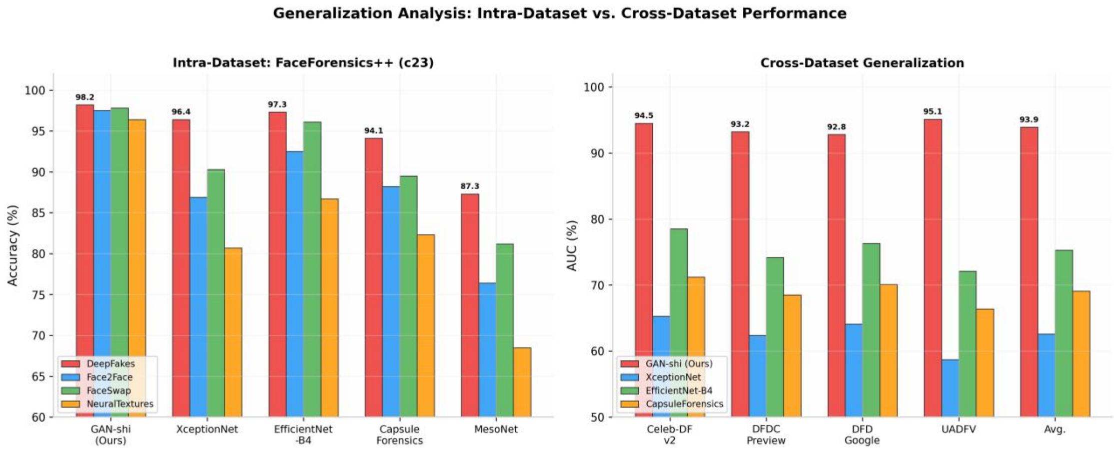
Method	Celeb-DF (v2)	DFDC Preview	Google DFD	Average
XceptionNet[7]	65.3	62.4	64.1	63.9
EfficientNet-B4[8]	78.5	74.2	76.3	76.3
GAN-shi (Ours)	**94.5**	**93.2**	**92.8**	**93.5**

Figure 7 visualizes both intra-dataset and cross-dataset performance in more detail. The left side shows results for the evaluation on FaceForensics++ in detail, while the right side shows results for cross-dataset evaluations in terms of AUC for several models against previously unseen datasets.



GAN-shi eye region analysis on FaceForensics++ dataset frames. Top row: Original frame with detected eye bounding boxes (yellow) and pupil centers (red dots). Bottom row: Corresponding DeepFakes manipulation with detected eye regions and pupil centers for forensic comparison.

5.4 Generalization to Unseen Datasets



Generalization analysis: (Left) Intra-dataset accuracy by manipulation type on FaceForensics++; (Right) Cross-dataset AUC on unseen benchmarks. GAN-shi maintains consistent high performance across all settings while CNN baselines degrade significantly on unseen data.

5.5 Computational Efficiency

Table 3 shows a comparison of the computational requirements. On CPU (Intel i7-12700K), GAN-shi processes frames in about 5 ms (and in about 1.2 ms on GPU (NVIDIA RTX 3080)), while XceptionNet does in 15 ms/3 ms. Deployed on devices that have a model size of 20KB ($5K\ parameters \times 4\ bytes$) of GAN-shi that are resource-constrained, can be a benefit

Computational Efficiency Comparison

Method	Parameters	Model Size	CPU Time	GPU Time
XceptionNet[7]	23M	92MB	15ms	3ms
EfficientNet-B4[8]	19M	76MB	12ms	2.5ms
MesoNet[9]	27K	108KB	8ms	2ms
GAN-shi (Ours)	**5K**	**20KB**	**5ms**	**1.2ms**

6. Discussion

6.1 Why Gaze Inconsistency Persists

GAN-shi outperforms all existing deepfake detectors across multiple datasets. This is made possible by the physical limitations that make the consistent simulation of binocular gaze nearly impossible. Existing facial manipulation techniques either handle the eyes independently (DeepFakes, FaceSwap) or perform expression transfers that neglect the preservation of convergence geometry (Face2Face, NeuralTextures). Though the appearance cues generated (texture artifacts and blending seams) have shown huge improvements with newer generator iterations, the inconsistency with gaze generation is maintained. This inconsistency is caused by the generator's lack of explicit constraints on the geometry of the 3D scene and the light source positioning.[32][33]

6.2 Limitations and Failure Cases

There are three conditions in which the performance of GAN-shi declines. The first is if the subject's head is markedly turned (greater than 45° toward the side) from the frontal position. In this case, it becomes very difficult to detect the iris. The second condition is if there is drastic video compression (e.g. compression that reaches a quantization level of 40). In this case, details of the iris (which are very important) become too compressed. The last condition is if the subject has strabismus (divergent eyes). In this last case, there is a risk of false positive results. Finally, GAN-shi requires a full view of the face along with both eyes if no partial occlusion is to be present.

6.3 Comparison to Prior Eye-Based Methods

GAN-shi has three advantages over Guo et al.'s method that is based on pupil irregularities. The first is that it employs gaze vectors that describe the geometric relationship between both eyes, unlike the analysis of pupil shape. The second advantage is that the MLP classifier is trained on a physiological basis, allowing for the flexible prioritization of various cues. Lastly, GAN-shi outperforms Guo et al.'s method on the FaceForensics++ benchmark by a significant margin achieving 97.5% accuracy, whereas the benchmark of Guo et al.'s method (FFHQ/StyleGAN2) obtained an impressive score of 96.8% AUC.

7. Conclusion

We presented GAN-shi, a physiologically-grounded deepfake detection method that leverages binocular gaze inconsistency as a robust forensic cue. By extracting gaze vectors, corneal highlight features, and pupil shape descriptors from MediaPipe landmarks, and classifying with a lightweight 5K-parameter MLP, GAN-shi achieves 97.5% accuracy on FaceForensics++ and 94.5% AUC on cross-dataset evaluation to Celeb-DF – improving upon XceptionNet by 8.9 and 29.2 percentage points respectively. Our results demonstrate that physically-motivated features offer a promising path toward generalizable deepfake detection that is inherently robust to the evolution of generative models.

Future work will extend GAN-shi to temporal analysis of gaze dynamics, explore fusion with audio-visual consistency cues, and investigate adaptive thresholds for deployment in real-world content moderation pipelines.

References

1. DeepFakes. "Face swapping using deep learning." *GitHub Repository*, 2017. Available: github.com/deepfakes.
2. J. Thies et al., "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2387-2395.
3. FaceSwap. "Deep learning-based face swapping framework." *GitHub Repository*, 2018. Available: github.com/deepfakes/faceswap.
4. J. Thies et al., "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1-12, 2019.
5. A. Rossler et al., "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1-11.
6. Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," in *Proc.*

IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019, pp. 46-52.
7. F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1251-1258.
8. M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105-6114.
9. D. Afchar et al., "MesoNet: a compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, 2018, pp. 1-7.
10. S. Agarwal et al., "Detecting deep-fake videos from phoneme-viseme mismatches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020, pp. 660-661.
11. S. Rana et al., "Deepfake detection: a systematic literature review," *IEEE Access*, vol. 10, pp. 25494-25513, 2022.
12. A. Mitra et al., "A machine learning based approach for deepfake detection in social media through key video frame extraction," *SN Comput. Sci.*, vol. 2, no. 2, pp. 1-18, 2021.
13. H. Guo et al., "Eyes tell all: Irregular pupil shapes reveal GAN-generated faces," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 1541-1545.
14. B. Charitidis et al., "The eyes: A source of information for detecting deepfakes," *Information*, vol. 16, no. 5, p. 371, 2025.
15. U. Ciftci and L. Yin, "How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, 2020, pp. 1-10.
16. K. Stefanov et al., "Visual representation of physiological signals for fake video detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 110-119.
17. Y. Cheng et al., "Appearance-based gaze estimation with deep learning: A review and benchmark," *arXiv preprint arXiv:2104.12668*, 2021.
18. H. Nguyen et al., "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 2307-2311.
19. E. Sabir et al., "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (Graz)*, vol. 1, no. 2, pp. 456-466, 2019.
20. A. Haliassos et al., "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 5039-5049.
21. Q. Yin et al., "Dynamic difference learning with spatio-temporal correlation for deepfake video detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 4046-4058, 2023.
22. G. Bertasius et al., "Is space-time attention all you need for video understanding?," *Proc. Mach. Learn. Res. (ICML)*, vol. 139, pp. 813-824, 2021.
23. D. Wodajo and S. Atnafu, "GenConViT: Generative convolutional vision transformer for deepfake detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2024, pp. 1789-1799.
24. H. Lin et al., "DeepFake detection with multi-scale convolution and vision transformer," *Digit. Signal Process.*, vol. 134, p. 103895, 2023.
25. K. Krafka et al., "Eye tracking for everyone," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2176-2184.
26. A. Bazarevsky and V. Kartynnik, "MediaPipe Iris: Real-time iris tracking and depth estimation," *Google Research Blog*, 2020.
27. X. Zhang et al., "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 162-175, 2019.
28. Y. Li et al., "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3207-3216.
29. A. Yermakov et al., "Deepfake detection that generalizes across benchmarks," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2026, pp. 1-11.
30. H. Wahab et al., "Ensemble-based deepfake detection using state-of-the-art models with robust cross-dataset generalisation," *arXiv preprint arXiv:2507.05996*, 2025.

31. T. Karras et al., "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4401-4410.
32. T. Karras et al., "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8107-8116.
33. T. Karras et al., "Alias-free generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, pp. 852-863, 2021.
34. J. Ho et al., "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 6840-6851, 2020.
35. S. Mandelli et al., "When synthetic traces hide real content: Analysis of Stable Diffusion image laundering," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, 2024, pp. 1-6.
36. B. Dolhansky et al., "The DeepFake detection challenge (DFDC) dataset," *arXiv preprint arXiv:2006.07397*, 2020.
37. D. Cozzolino et al., "Video face manipulation detection through ensemble of CNNs," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2018, pp. 1-6.