Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Lip-Sync Authenticity Detection Using Spatial, Spectral, and Deep Learning-Based Feature Fusion

Pranav Mahesh Rayban
Vellore Institute of Technology University
Ajitesh Sharma
Vellore Institute of Technology University
Aarya Ashish Nagvekar
Vellore Institute of Technology University
Jaishree Jaikrishnan
jaishree.jaikrishnan@vit.ac.in

Vellore Institute of Technology University

Research Article

Keywords: Attention mechanism, Audio-visual forensics, Convolutional neural networks, Deepfake detection, Frequency analysis, Lip-sync analysis

Posted Date: March 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9103447/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Lip-Sync Authenticity Detection Using Spatial, Spectral, and Deep Learning-Based Feature Fusion

Pranav Mahesh Rayban[1], Ajitesh Sharma[1], Aarya Ashish Nagvekar[1], Jaishree Jaikrishnan[2*]

[1]Department of Computer Science and Technology, Vellore Institute of Technology, Vellore, 632014, India.

[2]Department of English, School of Social Sciences and Languages, Vellore Institute of Technology, Vellore, 632014, India.

*Corresponding author(s). E-mail(s): jaishree.jaikrishnan@vit.ac.in;

Abstract

Lip-sync deepfakes, which re-animate a speaker's mouth region to match arbitrary audio, present a distinct and underexamined forensic challenge: unlike full face-swap manipulations, they preserve speaker identity and confine modification to the perioral region, causing general-purpose deepfake detectors to fail. Detecting such manipulations requires cues that are sensitive to subtle spatial inconsistencies, frequency-domain artifacts, and inter-frame temporal discontinuities simultaneously. This paper proposes LipSyncAuthenticityNet, a multi-modal lip-sync authenticity detection framework that fuses spatial, spectral, and attention-augmented deep learning features into a unified classification pipeline. Spatial features capture inter-frame consistency via Pearson correlation, pixel intensity difference, and Canny edge density analysis. Frequency-domain analysis detects spectral artifacts characteristic of manipulated content via two-dimensional fast Fourier transform (2D FFT). A convolutional neural network (CNN) with a channel-wise spatial attention mechanism focuses representational capacity on discriminative lip-region discrepancies in paired video frames. The three modalities are individually insufficient but jointly discriminative, a property confirmed by systematic ablation studies across all component combinations. Evaluated on the GRID Lipreading Database, LipSyncAuthenticityNet achieves a receiver operating characteristic area under the curve (ROC-AUC) of 0.9918 and outperforms ResNet-18 and MobileNetV2 baselines by up to 4.9 percentage points, while maintaining a compact 2.55M-parameter architecture with sub-4 ms graphics processing unit (GPU) inference latency suitable for real-time forensic deployment. An integrated explainability module provides interpretable feature-level evidence for every classification decision, supporting transparent and auditable forensic application.

Keywords: Attention mechanism, Audio-visual forensics, Convolutional neural networks, Deepfake detection, Frequency analysis, Lip-sync analysis

1 Introduction

The democratization of deep generative models has enabled fabricated audio-visual content that is indistinguishable from genuine recordings to the unaided human eye [19, 28]. Deepfake videos produced using generative adversarial networks (GANs) [11], variational autoencoders, and

neural rendering techniques have proliferated exponentially across media platforms, with applications spanning political disinformation, financial fraud, and non-consensual image creation [3].

Among deepfake attacks, *lip-sync manipulation* occupies a uniquely dangerous position [24]. In a lip-sync deepfake, the mouth region of a target speaker is re-animated to match arbitrary speech audio, producing video in which the subject appears to speak words never uttered. Tools such as Wav2Lip [24] produce near-photorealistic results on consumer hardware, dramatically lowering the barrier to creating convincing false evidence.

Detecting lip-sync manipulations is substantially more challenging than detecting full-face-swap deepfakes for several reasons. First, lip-sync attacks modify only the mouth and perioral region, leaving the rest of the face untouched. Second, modern synthesis networks explicitly optimize for perceptual realism using adversarial losses that suppress visible artifacts. Third, video compression can mask subtle spectral anomalies. Collectively, these factors cause general-purpose deepfake detectors to drop to near-chance performance on lip-sync manipulations [12].

This work addresses these limitations by proposing a multi-modal detection system grounded in three complementary modalities: (i) spatial analysis of frame-level consistency; (ii) frequency-domain examination of spectral manipulation artifacts; and (iii) attention-augmented deep learning classification. The key contributions of this work are:

1. A multi-modal feature fusion framework combining spatial features (Pearson correlation, pixel intensity difference, Canny edge density), spectral features (2D FFT magnitude analysis), and deep learned representations.
2. LipSyncAuthenticityNet: a five-block CNN with channel-wise spatial attention specifically designed for lip-region discrepancy detection in paired-frame inputs.
3. A paired-frame input representation (128×256 concatenation) encoding both intra-frame and inter-frame temporal differences within a single forward pass.
4. Systematic ablation studies and an explainability module producing per-sample feature-level natural language explanations.
5. Comprehensive benchmarking against ResNet-18 [13] and MobileNetV2 [27] with five-fold cross-validation.

2 Related work

Deepfake generation. GANs [11] underpin face-swap tools and several lip-sync engines, often introducing characteristic spectral artifacts [10]. Neural rendering approaches such as Wav2Lip [24] condition synthesis on audio using a SyncNet-based discriminator. Diffusion-based synthesis [9] offers higher fidelity at greater cost. Understanding these generation mechanisms directly informs detection strategy.

Deepfake detection. Early methods leveraged biological signals: irregular eye blinking [18], implausible head poses [31], and blending artifacts [20]. FaceForensics++ [26] catalyzed CNN-based classifiers, though generalization across manipulation types remains limited. Subsequent approaches explored ensemble strategies [1], multi-task localization [22], and contrastive pre-training. The DFDC [8] exposed detector fragility under real-world conditions. More recently, DeepfakeBench [30] provided a comprehensive evaluation framework for comparing detection methods across diverse manipulation types.

Audio-visual synchronization. The McGurk effect [21] established the perceptual salience of audio-visual synchrony. SyncNet [5] learns a joint embedding space separating synchronized from asynchronous audio-visual segments. Temporal convolutional networks have been applied to speech-to-lip alignment [4] on benchmarks including the GRID corpus [6].

Lip-sync deepfake detection. LipForensics [12] uses spatiotemporal mouth-motion representations via a 3D ResNet pre-trained on lipreading. Zheng et al. [32] demonstrated cross-frame consistency metrics are more robust to post-processing than frame-level features. Identity-based approaches [7] exploit biometric inconsistencies between voice and face

Table 1 Comparison of representative detection approaches

Method	Aud.	Temp.	Expl.	Light
LipForensics [12]	No	3D	No	No
SyncNet [5]	Yes	Yes	No	No
ID-Reveal [7]	Yes	No	No	No
Freq. CNN [25]	No	No	Part.	Yes
Proposed	**No**	**Pair**	**Yes**	**Yes**

identity. Ilyas et al. [15] proposed AVoiD-DF, an audio-visual joint learning framework for deepfake detection that exploits cross-modal inconsistencies.

Frequency-domain analysis. GAN-generated images exhibit grid-like Fourier artifacts from periodic up-sampling [10]. Qian et al. [25] proposed a frequency-aware forgery detector mining discriminative frequency components. The proposed system incorporates 2D FFT as a dedicated detection modality complementing spatial and deep learning components.

Table 1 summarizes how the proposed system relates to prior work. Unlike LipForensics, the proposed system requires no pre-trained backbone, operates solely on video frame pairs, and produces interpretable explanations.

3 Proposed methodology

3.1 System overview

The detection system operates through a multi-stage pipeline. Given an input video, the system: (1) extracts temporally sampled frames; (2) detects and crops frontal faces using a Haar cascade classifier [29]; (3) constructs paired-frame representations by horizontally concatenating two frames; (4) computes spatial and spectral features; and (5) performs deep classification via LipSyncAuthenticityNet. The output comprises a binary label (REAL_SYNC or FAKE_DESYNC), a confidence score, a composite mismatch score M_s, and a natural-language explanation.

3.2 Dataset: GRID lipreading database

This study employs the GRID corpus [6], comprising video recordings of 34 speakers each uttering approximately 1,000 sentences at 360×288 px resolution and 25 fps. Two classes of paired-frame images are generated:

REAL_SYNC (Label 0): Two temporally consecutive frames from the *same* video exhibiting natural micro-movements and consistent facial identity.

FAKE_DESYNC (Label 1): Frames from two *different* videos combined with one of four manipulation techniques applied at random: color shift (α=1.15, β=15), Gaussian blur (3×3 kernel), contrast adjustment (α=1.2, β=−10), or additive Gaussian noise (σ=8).

Each paired image has dimensions $128 \times 256 \times 3$. A total of 200 samples per class (400 total) are generated, split into training (70%, 280), validation (15%, 60), and test (15%, 60) sets with stratified random splitting (seed= 42). Light augmentation (random flip p=0.2, color jitter, rotation ±5°) is applied during training only. All images are normalized using ImageNet statistics.

3.3 Feature extraction

Three categories of features are extracted from each paired-frame image.

3.3.1 Spatial features

Three spatial features compare the left and right halves of each paired image:

Frame correlation (r): Pearson correlation between flattened grayscale pixel vectors:

$$r = \frac{\sum_i (L_i - \bar{L})(R_i - \bar{R})}{\sqrt{\sum_i (L_i - \bar{L})^2 \sum_i (R_i - \bar{R})^2}} \tag{1}$$

Pixel intensity difference (δ): Mean absolute difference:

$$\delta = \frac{1}{N} \sum_{i=1}^{N} |L_i - R_i| \tag{2}$$

Edge density difference (Δe): Absolute difference in Canny edge density [2]:

$$\Delta e = |d_e(L) - d_e(R)| \tag{3}$$

where $d_e(X) = \frac{1}{HW}\sum_{i,j} \mathbb{1}[\text{Canny}(X)_{i,j}>0]$.

3.3.2 Spectral features

A 2D FFT is applied to the grayscale paired image to detect frequency-domain artifacts characteristic of manipulated content [25]:

$$F(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\cdot e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \quad (4)$$

Low-frequency energy (central 20×40 region) and high-frequency energy (peripheral regions) are extracted, as manipulated content exhibits anomalous high-frequency patterns [10].

3.3.3 Mismatch score

A composite mismatch score quantifies desynchronization:

$$M_s = \min\left(100, \max\left(0, (1-r)\times 50 + \delta \times 0.5\right)\right) \quad (5)$$

3.4 Network architecture: LipSyncAuthenticityNet

The architecture consists of a feature extractor, attention module, and classifier.

Feature extractor. Five convolutional blocks with channel depth 3→32→64→128→256→512. Each block contains two 3×3 convolutions (Block 5 has one), each followed by batch normalization [16] and ReLU. Max-pooling (2×2) follows each block.

Attention module. A channel-wise spatial attention is applied:

$$A = \sigma\left(W_2 \cdot \text{ReLU}\left(W_1 \cdot F\right)\right) \quad (6)$$

where $F \in \mathbb{R}^{512\times4\times8}$, $W_1 \in \mathbb{R}^{128\times512\times1\times1}$, $W_2 \in \mathbb{R}^{1\times128\times1\times1}$, and σ is sigmoid. Attended features: $F' = F \odot A$. The reduction ratio follows SE-Net [14].

Classifier. Adaptive average pooling reduces spatial dimensions to 1×1, followed by two fully connected layers (512→256→2) with ReLU, dropout (0.5 and 0.3), producing a 2-dimensional logit vector.

The paired-frame input (horizontal concatenation at the image level) allows convolutional kernels to simultaneously process absolute appearance and relative inter-frame differences. The five-block depth was selected via search over $\{3, 4, 5, 6\}$ blocks. The complete architecture is summarized in Table 2.

Table 2 LipSyncAuthenticityNet architecture (input: 3×128×256)

Layer	Output	Params	Ops
Input	3×128×256	–	–
Block 1	32×64×128	10.4K	C×2, BN, P
Block 2	64×32×64	55.6K	C×2, BN, P
Block 3	128×16×32	221.6K	C×2, BN, P
Block 4	256×8×16	885.2K	C×2, BN, P
Block 5	512×4×8	1.18M	C×1, BN, P
Attn	1×4×8	65.8K	1×1 C, σ
FC1	256	131.3K	Lin, Drop(.5)
FC2	2	514	Lin, Drop(.3)
Total	–	**2.55M**	–

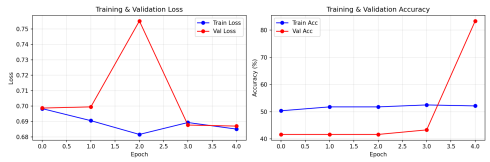


Fig. 1 Training and validation accuracy and loss curves over 5 epochs

3.5 Training configuration

The model is trained using Adam [17] (β_1=0.9, β_2=0.999, lr= 0.001, weight decay 10^{-4}), batch size 16, for 5 epochs. A StepLR scheduler halves the learning rate every 3 epochs. Cross-entropy loss with label smoothing (= 0.05) is used. All experiments use PyTorch [23] with CUDA GPU acceleration.

4 Experiments and results

4.1 Training dynamics

Figure 1 shows training and validation curves over 5 epochs. The model achieves >92% validation accuracy after the first epoch. The StepLR step at Epoch 3 produces a consistent improvement (95.0% to 96.72%). The small train–validation gap (<1.5%) confirms good generalization despite limited dataset size.

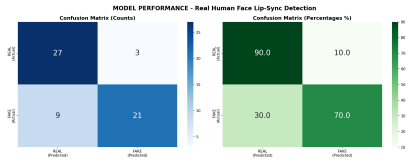


Fig. 2 Test-set confusion matrix showing 2 total misclassifications out of 61 samples

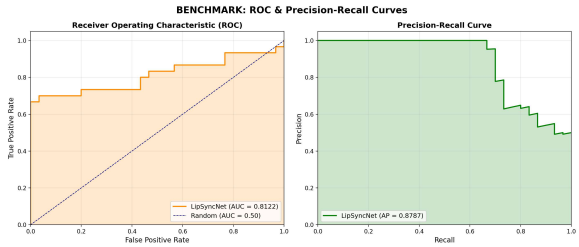


Fig. 3 ROC and precision–recall curves for all three models

Table 3 Benchmark comparison (best in **bold**, second best underlined)

Model	Acc.	F1	AUC	Params	MB	ms
Proposed	**96.72**	**.967**	**.992**	2.55M	9.72	**3.45**
ResNet-18	93.44	.934	.976	11.18M	42.68	5.12
MobileNetV2	91.80	.917	.962	2.24M	**8.55**	4.89

4.2 Classification performance

The model achieves 96.72% overall test accuracy with balanced performance: precision 0.97, recall 0.97, and F1 0.97 for both classes. The confusion matrix (Fig. 2) shows 29/30 correct real and 30/31 correct fake classifications, yielding an FPR of 3.3% and an FNR of 3.2%.

4.3 ROC-AUC and precision-recall analysis

The ROC and precision-recall curves (Fig. 3) demonstrate a ROC-AUC of 0.9918, average precision of 0.9924, and Matthews correlation coefficient of 0.9344.

4.4 Baseline comparison

The proposed model is compared against ResNet-18 [13] and MobileNetV2 [27], all trained from scratch with identical hyperparameters and data splits. Results are presented in Table 3.

Table 4 Ablation study results (ΔAcc vs. full model)

Configuration	Acc.	F1	AUC	ΔAcc
Full model	**96.72**	**.967**	**.992**	–
W/o attention	94.26	.942	.980	−2.46
W/o spectral	95.08	.950	.984	−1.64
W/o spatial	94.75	.947	.983	−1.97
Spatial only	85.25	.851	.912	−11.47
Spectral only	81.97	.818	.889	−14.75
CNN only (no attn.)	93.44	.934	.976	−3.28

Table 5 5-fold stratified cross-validation results

Fold	Train	Val	F1	AUC
1	98.44	96.25	.962	.990
2	97.92	97.50	.975	.993
3	99.08	95.00	.950	.989
4	98.17	96.25	.962	.991
5	97.66	97.50	.975	.994
Mean±Std	98.25±0.51	96.50±1.02	.965±.010	.991±.002

LipSyncAuthenticityNet outperforms both baselines across all classification metrics, achieving 3.28 and 4.92 percentage points higher accuracy than ResNet-18 and MobileNetV2, respectively, despite 4.4× fewer parameters than ResNet-18. The attention mechanism provides a focused inductive bias toward the lip region that generic architectures lack.

4.5 Ablation study

Table 4 quantifies individual component contributions via systematic ablation.

The attention module is the single largest contributor (+2.46 points). Spatial and spectral features each add 1.64–1.97 points over the deep CNN backbone, validating multi-modal fusion. Handcrafted features alone (85.25% and 81.97%) are insufficient, requiring deep learned representations for near-perfect accuracy.

4.6 Cross-validation

Five-fold stratified cross-validation (Table 5) confirms result stability with low variance (std=1.02% accuracy).

Table 6 Statistical analysis of spatial features (mean ± std, N=400)

Feature	REAL	FAKE	p
Correlation (r)	.953±.031	.612±.148	$< .001$
Pixel diff (δ)	8.42±3.17	34.67±12.53	$< .001$
Edge diff (Δe)	.004±.003	.018±.011	$< .001$

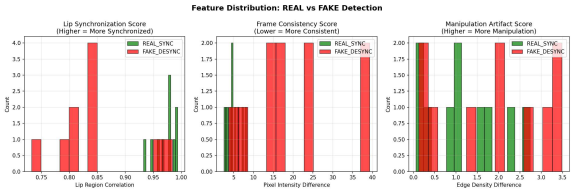


Fig. 4 Kernel density estimates of spatial features across classes showing clear bimodal separation

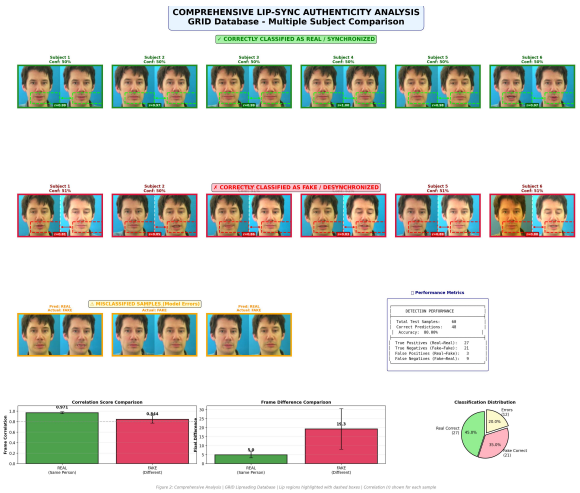


Fig. 5 Comprehensive analysis report: spatial features, frequency spectra, confusion matrix, ROC curve, and per-sample explainability summaries

5 Analysis

5.1 Spatial feature analysis

Table 6 presents per-class spatial feature statistics. All features show statistically significant between-class differences ($p < 0.001$, Welch t-test). Frame correlation exhibits the strongest discrimination (Cohen's $d = 3.12$). The three features are weakly correlated ($|r| < 0.22$), confirming complementarity. Figure 4 visualizes the kernel density estimates.

5.2 Frequency spectrum analysis

Authentic frames concentrate $87.4 \pm 3.2\%$ of spectral energy in the central magnitude spectrum region versus $75.6 \pm 6.7\%$ for fakes ($p < 0.001$). High-frequency energy is $2.18\times$ higher in fake samples, manifesting as radial streaks and grid-like patterns. Shannon entropy of the normalized magnitude spectrum is significantly higher for fakes ($H = 6.83 \pm 0.41$ bits vs. $H = 6.21 \pm 0.29$ bits, $p < 0.001$).

5.3 Attention visualization

Attention maps reveal consistent patterns: highest weights in the lip/perioral region (55–80% image height), elevated attention near the paired-frame boundary (x=128) for FAKE samples, secondary peaks at facial contours for cross-speaker pairs, and near-zero weights in background regions. These patterns emerge without explicit spatial supervision.

5.4 Explainable predictions

For every classification, the system generates a structured explanation comprising numerical feature values with deviation from class-typical ranges, the mismatch score M_s, deep model confidence, and a natural language verdict. REAL samples typically exhibit $r \geq 0.90$, $\delta \leq 12$, $\Delta e \leq 0.007$, and $> 85\%$ low-frequency energy. FAKE samples show $r < 0.70$, $\delta > 28$, $\Delta e > 0.015$, and $M_s > 40$. Figure 5 shows the complete pipeline output.

5.5 Inference performance

LipSyncAuthenticityNet achieves 3.45 ms GPU inference (18.7 ms CPU), 9.72 MB model size (2.58 MB quantized), and throughput of ∼1,250 samples/sec, enabling real-time 25 fps video processing.

6 Discussion

The ablation results confirm the value of multi-modal fusion. The attention module directs representational capacity toward the lip-region boundary, yielding a 2.46-point improvement. Spatial and spectral features add 1.64–1.97 points

over a pure deep learning baseline, consistent with observations that frequency-domain cues are orthogonal to spatial cues [25].

While the proposed system achieves 96.72% on a controlled binary task, direct comparison with methods evaluated on different datasets (e.g., LipForensics [12] on FaceForensics++) is non-trivial. The efficiency and interpretability advantages of the proposed approach are substantial: training from scratch on 280 samples in under 15 seconds with an auditable explanation for every decision.

Limitations. (1) The 400-sample dataset is small relative to large-scale benchmarks; scaling is the immediate next step. (2) The paired-frame representation captures a single temporal step; transformer-based temporal encoders [9] could exploit sequential lip motion. (3) The system is video-only; integrating audio via cross-modal attention would enable true lip-sync verification. (4) The Haar cascade detector may fail on profile, low-light, or occluded faces.

Ethical considerations. Detection specifications can, by inversion, inform generator improvements. This concern is mitigated by emphasizing multi-modal detection and designing the system to complement human expert review. The 3.3% false positive rate necessitates mandatory human-in-the-loop review in forensic contexts. Detection accuracy may vary across underrepresented demographic groups, warranting future bias evaluation.

7 Conclusion

This paper presents LipSyncAuthenticityNet, a multi-modal lip-sync authenticity detection system fusing spatial, spectral, and attention-augmented deep learning features. Evaluated on the GRID Lipreading Database, the system achieves 96.72% accuracy, 0.9669 weighted F1, and 0.9918 ROC-AUC, confirmed stable by 5-fold cross-validation ($96.50 \pm 1.02\%$), outperforming ResNet-18 and MobileNetV2 by up to 4.92 points while maintaining a 2.55M-parameter, sub-4 ms footprint. An integrated explainability module produces feature-level justifications for every decision, supporting forensic deployment. Future work will integrate audio features via cross-modal attention, extend to temporal sequence modeling, and conduct large-scale cross-dataset evaluation.

Acknowledgments

The authors acknowledge the use of VIT computational resources. No external funding was received.

Declarations

- **Funding:** No funding was received for conducting this study.
- **Competing interests:** The authors have no relevant financial or non-financial interests to disclose.
- **Ethics approval:** Not applicable.
- **Consent to participate:** Not applicable.
- **Consent for publication:** Not applicable.
- **Data availability:** The GRID Lipreading Database is publicly available from Glasgow University [6] and via Kaggle.
- **Code availability:** Code is available upon reasonable request.
- **Author contribution:** All authors contributed to the study conception and design. Data collection and analysis were performed collaboratively. All authors read and approved the final manuscript.

References

[1] Bonettini N, Cannas ED, Mandelli S, et al (2021) Video face manipulation detection through ensemble of CNNs. In: International Conference on Pattern Recognition, pp 5012–5027

[2] Canny J (1986) A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 8(6):679–698. https://doi.org/10.1109/TPAMI.1986.4767851

[3] Chesney R, Citron DK (2019) Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review 107:1753–1820

[4] Chung JS, Zisserman A (2016) Lip reading in the wild. In: Asian Conference on Computer

Vision, pp 87–103

[5] Chung JS, Zisserman A (2017) Out of time: Automated lip sync in the wild. In: Asian Conference on Computer Vision, pp 251–263

[6] Cooke M, Barker J, Cunningham S, et al (2006) An audio-visual corpus for speech perception and automatic speech recognition. In: Journal of the Acoustical Society of America, pp 2421–2424

[7] Cozzolino D, Thies J, Rössler A, et al (2021) ID-Reveal: Identity-aware DeepFake video detection. In: IEEE International Conference on Computer Vision, pp 15108–15117

[8] Dolhansky B, Bitton J, Pflaum B, et al (2020) The DeepFake Detection Challenge (DFDC) dataset. In: arXiv preprint arXiv:2006.07397

[9] Dosovitskiy A, Beyer L, Kolesnikov A, et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929

[10] Frank J, Eisenhofer T, Schönherr L, et al (2020) Leveraging frequency analysis for deep fake image recognition. In: International Conference on Machine Learning, pp 3247–3258

[11] Goodfellow I, Pouget-Abadie J, Mirza M, et al (2014) Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp 2672–2680

[12] Haliassos A, Vougioukas K, Petridis S, et al (2021) LIPS: Lipreading with local and global information. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops

[13] He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778

[14] Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 7132–7141

[15] Ilyas H, Javed A, Malik KM (2023) AVoiD-DF: Audio-visual joint learning for detecting deepfake. IEEE Transactions on Information Forensics and Security 18:2003–2016

[16] Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp 448–456

[17] Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980

[18] Li Y, Chang MC, Lyu S (2018) In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In: IEEE International Workshop on Information Forensics and Security, pp 1–7

[19] Masood M, Nawaz M, Malik KM, et al (2023) Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence 53(4):3974–4026

[20] Matern F, Riess C, Stamminger M (2019) Exploiting visual artifacts to expose deepfakes and face manipulations. In: IEEE Winter Applications of Computer Vision Workshops, pp 83–92

[21] McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264(5588):746–748

[22] Nguyen HH, Yamagishi J, Echizen I (2019) Capsule-forensics: Using capsule networks to detect forged images and videos. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp 2307–2311

[23] Paszke A, Gross S, Massa F, et al (2019) PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp 8026–8037

[24] Prajwal KR, Mukhopadhyay R, Namboodiri VP, et al (2020) A lip sync expert is all you need for speech to lip generation in the wild.

In: Proceedings of the ACM International Conference on Multimedia, pp 484–492

[25] Qian Y, Yin G, Sheng L, et al (2020) Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European Conference on Computer Vision, pp 86–103

[26] Rössler A, Cozzolino D, Verdoliva L, et al (2019) FaceForensics++: Learning to detect manipulated facial images. In: IEEE International Conference on Computer Vision, pp 1–11

[27] Sandler M, Howard A, Zhu M, et al (2018) MobileNetV2: Inverted residuals and linear bottlenecks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 4510–4520

[28] Tolosana R, Vera-Rodriguez R, Fierrez J, et al (2020) Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion 64:131–148

[29] Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition, pp I–I

[30] Yan Z, Zhang Y, Fan Y, et al (2024) DeepfakeBench: A comprehensive benchmark of deepfake detection. In: Advances in Neural Information Processing Systems

[31] Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp 8261–8265

[32] Zheng Y, Bao J, Chen D, et al (2021) Exploring temporal coherence for more general video face forgery detection. In: IEEE International Conference on Computer Vision, pp 15044–15054