Rajnandini Birajadar, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

DeepShield: A Hybrid Deep Learning Framework for Real-Time Deepfake Detection Using Spatial and Temporal Cues

Rajnandini Birajadar, Sanika Shinde, Krutika Sane, Shivani Khopkar
Guide Name - Apurva Deshpande
Dept. of AIML Engineering Rasiklal M. Dhariwal Institute of Technology, Pune, India

Abstract- Deep learning has demonstrated remarkable success in solving complex problems across various domains, such as big data analytics, computer vision, and human-level control. However, the same advancements in deep learning have also given rise to applications that pose threats to privacy, democracy, and national security. One such application is deepfake technology, which leverages deep learning algorithms to create convincingly realistic fake images and videos that are indistinguishable from authentic ones. Consequently, the need for technologies capable of automatically detecting and assessing the integrity of digital visual media has become imperative. This paper aims to present a comprehensive survey of the algorithms employed to create deepfakes and, more importantly, the methods proposed in the literature for detecting deep fakes. The survey delves into extensive discussions on the challenges, research trends, and future directions concerning deepfake technologies. By reviewing the background of deepfakes and examining state-of the-art deepfake detection methods, this study provides an inclusive overview of deepfake techniques, thereby facilitating the development of novel and robust methods to combat the increasingly sophisticated deep fake threats in conclusion, this survey paper provides a comprehensive overview of deepfake techniques and detection methods. By synthesizing the existing literature and highlighting research trends and challenges, it aims to support the development of novel and effective approaches to combat the growing threat of deep fakes, ensuring the integrity, privacy, and security of digital visual media in an increasingly complex and interconnected world.

Keywords— Deep learning, deepfake technology, fake images and videos, privacy threats, security risks, detection algorithms, media integrity, research challenges, and future directions—this survey highlights the growing impact of deepfakes and emphasizes the need for advanced detection methods to ensure trustworthy digital media.

I. INTRODUCTION

Deepfakes, in a narrow definition, are a type of artificial content created using deep learning techniques that involve superimposing face images of a target person onto a video of a source person. This manipulation makes it appear as though the target person is performing actions or saying things that the source person actually did. This specific category of deepfakes is commonly known as face swapping. However, in a broader sense, deep fakes encompass other types of AI-generated content as well. Two additional categories of deepfakes are lip-sync and puppet-master. Lip-sync deep fakes involve modifying videos to synchronize the movements of the subject's mouth with a particular audio recording. By altering the original video, the mouth movements of the person in the video are made consistent with the audio, creating a convincing lip-sync effect. Puppet-master deep fakes, on the other hand, consist of videos featuring a target person (the puppet) whose facial expressions, eye movements, and head movements are animated to mimic those of another person (the master) who is situated Infront of a camera. The puppet follows the actions and expressions of the master, resulting in a video where the target person appears to be controlled by

the movements of the master. It is important to note that these categories of a deepfakes are not mutually exclusive, and a deepfake can incorporate elements from multiple categories. The broader definition of deep fakes encompasses not only face swapping but also lip-sync and puppet-master techniques, enabling a wider range of AI-synthesized content that can potentially deceive viewers.

II. RELATED WORK

This section traces the evolution of deepfake detection techniques:

Early / Hand-Crafted Methods: Researchers initially relied on biological cues like eye-blinking frequency [3] and head-pose estimation [4]. These were quickly defeated as GAN training improved and generators learned to replicate these physiological signals naturally.

CNN-Based Spatial Detectors: The release of the FaceForensics++ benchmark (Rössler et al., ICCV 2019) [5] shifted the field toward data-driven approaches. XceptionNet [6] became a dominant detector by learning subtle facial manipulation artefacts from raw pixels. Li et al. [7] showed that face warping creates characteristic geometric distortions that CNNs can identify. However, a critical weakness emerged — JPEG compression and down-sampling severely degraded the performance of these spatial-only models.

Temporal / 3D Approaches: LSTMs on frame sequences [8] and 3D convolutional networks [9] were introduced to model motion inconsistencies across video frames. While effective at capturing temporal patterns, they are computationally expensive and hard to train with limited labeled data.
Transformer-Based Detectors: The Vision Transformer (ViT) adaptation by Wodajo et al. [10] achieved strong accuracy but required large-scale pre-training data and significant GPU memory — limiting real-world usability.

Frequency-Domain Methods: Frank et al. [11] exploited GAN-specific spectral artefacts in the frequency domain. These methods work well against known generators but are fragile under image compression (which redistributes high-frequency energy).

Gap Addressed by DeepShield: DeepShield builds on insights from [7], [9], and [10] while fixing their complementary weaknesses through cross-modal fusion and efficient quantised deployment.

III. SYSTEM ARCHITECTURE

DeepShield follows a dual-stream fusion architecture with five main components:

A. Face Extraction Pipeline

- Uses RetinaFace detector to extract aligned face crops at 224 × 224 pixels.
- A five-landmark affine transform normalises head-pose variation.
- For the temporal stream, consecutive face crops are used to compute Farnebäck dense optical-flow difference maps, producing a 3-channel UV-magnitude representation per frame pair.

B. Spatial Stream

- Backbone: EfficientNet-B4 pre-trained on ImageNet.
- The final classification layer is replaced with a Global Average Pooling (GAP) layer → 512-dimensional FC layer with ReLU, producing spatial feature vector $f_s \in R^{512}$.
- Fine-tuning: last three MBConv blocks and top dense layers are unfrozen; earlier layers are frozen to retain low-level feature detectors.

C. Temporal Stream

- Processes a sequence of T = 16 optical-flow maps through a two-layer Bidirectional LSTM (BiLSTM) with 256 hidden units per direction.
- Captures both forward and backward motion context, producing temporal feature vector $f_t \in R^{512}$ from the final hidden state concatenation.
- A 0.3 dropout layer is applied between BiLSTM layers to prevent overfitting.

D. Cross-Attention Fusion Module

- Instead of simple feature concatenation, a cross-attention mechanism adaptively fuses spatial and temporal features.
- Query (Q), Key (K), and Value (V) projections are computed via learned linear transforms W_Q, W_K, W_V ∈ R^{512×64}.
- Attention weight matrix: A = softmax(Q · Kᵀ / √64)
- Fused representation f_fused = A · V is concatenated with a residual shortcut [f_s ; f_t], producing a 1024-dimensional fused vector.

IV. METHODOLOGY

Three benchmark datasets are used:

Dataset	Details
FaceForensics++	1,000 original videos, 4 manipulation types (Deepfakes, Face2Face, Face Swap, Neural Textures), 3 compression levels (raw, c23, c40)
Celeb-DF v2	590 celebrity videos + 5,639 high-quality deepfakes
DFDC	100,000+ videos from 3,000+ paid actors — most diverse real-world distribution

An 80/10/10 train-validation-test split is used with no subject overlap between splits.

B. Training Protocol

- Spatial stream: Fine-tuned for 30 epochs, Adam optimizer (lr = 2e-4, weight decay = 1e-5), cosine LR annealing.
- Temporal stream: Trained for 20 epochs (lr = 5e-5, lower due to smaller optical-flow dataset).
- End-to-end joint training: 15 additional epochs with combined binary cross-entropy loss.
- Data augmentation: Random horizontal flipping, Gaussian blurring, JPEG re-compression (quality 50–95), colour jitter — all to simulate real-world content moderation inputs.

C. Compression Robustness Evaluation

Models are tested at three compression levels: lossless (c0), low (c23, QP=23), high (c40, QP=40). The robustness metric used is: ΔAcc = Acc(c0) − Acc(c40).

D. Adversarial Robustness

Two adversarial attack strategies are evaluated:

- FGSM perturbations (ε = 4/255 in L∞ norm) on input face crops.
- GAN-based adversarial training [17], where the deepfake generator is jointly optimised to fool the detector.

E. Deployment Optimisation

- PyTorch model exported to ONNX format.
- Quantised to INT8 using ONNX Runtime's post-training static quantisation with a 512-sample calibration set.
- Benchmarked on: NVIDIA RTX 3060 GPU (12 GB VRAM) and Intel Core i7-12700H CPU.

V. RESULTS AND ANALYSIS

All Models Are Evaluated On Held-Out Test Sets Of Faceforensics++ (C23), Celeb-Df V2, And Dfdc. Results Are Macro-Averaged Across Datasets.

Table I: Comparative Performance

Method	Accuracy (%)	AUC-ROC	F1-Score	FPR (%)
XceptionNet (baseline)	91.2	0.934	0.908	8.6
EfficientNet-B4	93.5	0.951	0.931	6.4
LSTM + CNN	89.7	0.916	0.893	9.8
Vision Transformer (ViT)	94.1	0.958	0.937	5.8

Method	Accuracy (%)	AUC-ROC	F1-Score	FPR (%)
DeepShield (Proposed)	97.3	0.981	0.971	2.9

Key Findings:

- Deepshield Surpasses The Next-Best Model (Vit) By 3.2% Accuracy And 0.023 Auc.
- The False Positive Rate Of 2.9% Is Critical — Misclassifying Real Content In Moderation Contexts Has Legal And Reputational Consequences.

Ablation Study:

- Removing The Temporal Stream → Accuracy Drops By 4.1 Points (To 93.2%), Confirming That Motion Inconsistencies Are Crucial.
- Replacing Cross-Attention With Simple Concatenation → Accuracy Drops By 1.8 Points (To 95.5%), Validating The Advantage Of Adaptive Modality Weighting.

Compression Robustness:

- Deepshield Achieves The Lowest Degradation: Δacc = 3.1% Vs. 9.4% (Xceptionnet) And 6.2% (Efficientnet-B4 Alone).
- This Is Attributed To The Temporal Stream's Optical-Flow Representation Being More Compression-Invariant Than Raw Pixel Features.

Adversarial Robustness:

- Under Fgsm Attack (E = 4/255): Deepshield Retains 91.4% Accuracy Vs. Only 78.3% For Xceptionnet — The Dual-Stream Architecture Provides Inherent Robustness Through Feature Diversity.

Deployment:

- Onnx-Quantised Model: 28.3 Fps On Rtx 3060 And 6.1 Fps On Cpu.
- Quantisation Accuracy Loss: Only 0.9% (97.3% → 96.4%), Validating Int8 Deployment Viability.

VI. CONCLUSION

Deep Shield is presented as a hybrid deep learning framework for real-time deepfake video detection combining:

- EfficientNet-B4 (spatial pixel-level artefact detection)
- BiLSTM on optical-flow (temporal motion inconsistency modelling)
- Cross-attention fusion (adaptive modality weighting)

It achieves state-of-the-art performance (97.3% accuracy, AUC 0.981) with robustness to compression and adversarial attacks. Its 28 fps real-time ONNX pipeline bridges the gap between research accuracy and production deploy ability, and its low FPR makes it fit for automated content moderation.

Future Directions:

- Extension to audio-visual deepfakes (lip-sync forgeries).
- Domain adaptation for emerging diffusion-based synthesis methods.
- Federated learning for training on privacy-sensitive forensic datasets.
- Integration with blockchain-based media provenance systems (e.g., C2PA) for end-to-end authenticity guarantees.

Acknowledgment

The authors thank:

- The Department of Artificial Intelligence and Machine Learning Engineering faculty at Rasiklal M. Dhariwal Institute of Technology for guidance.
- Creators of FaceForensics++, Celeb-DF v2, and DFDC datasets for making benchmarks publicly available.
- Open-source communities behind PyTorch, ONNX Runtime, RetinaFace, and EfficientNet.

REFERENCES

1. H. Ajder et al., "The State of Deepfakes: Landscape, Threats, and Impact," Deeptrace Report, 2019
2. F. Chesney and D. Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," California Law Review, vol. 107, pp. 1753–1820, 2019
3. Y. Li et al., "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking," IEEE WIFS, 2018
4. X. Yang et al., "Exposing Deep Fakes Using Inconsistent Head Poses," IEEE ICASSP, 2019
5. A. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," ICCV, 2019
6. F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," CVPR, 2017
7. Y. Li et al., "Face X-Ray for More General Face Forgery Detection," CVPR, 2020
8. D. Guerra and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," AVSS, 2018
9. I. Sabir et al., "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," CVPRW, 2019
10. D. Wodajo and S. Atnafu, "Deepfake Video Detection Using Convolutional Vision Transformer," arXiv:2102.11126, 2021
11. J. Frank et al., "Leveraging Frequency Analysis for Deep Fake Image Recognition," ICML, 2020
12. J. Deng et al., "RetinaFace: Single-Stage Dense Face Localisation in the Wild," CVPR, 2020
13. G. Farnebäck, "Two-Frame Motion Estimation Based on Polynomial Expansion," SCIA, 2003
14. M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for CNNs," ICML, 2019
15. Y. Li et al., "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," CVPR, 2020
16. B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) Dataset," arXiv:2006.07397, 2020
17. N. Neves et al., "Adversarial Examples Against Deep Neural Networks for Deepfake Detection," IEEE Access, vol. 10, pp. 37567–37580, 2022

Author Profile

1. Rajnandini Birajdar

Rajnandini Birajadar is an undergraduate student in the Department of Artificial Intelligence and Machine Learning at Rasiklal M. Dhariwal Institute of Technology, Pune. Her areas of interest include deep learning, computer vision, and data analytics. She is particularly interested in developing AI-based solutions for real-world problems and has worked on research related to deepfake detection techniques.

2. Krutika Sane

Krutika Sane is an undergraduate student in the Department of Artificial Intelligence and Machine Learning at Rasiklal M. Dhariwal Institute of Technology, Pune. Her interests lie in artificial intelligence, machine learning, and user interface design. She focuses on system architecture and integration of AI models, and is keen on building practical and efficient AI-driven applications.

3. Sanika Shinde

Sanika Shinde is an undergraduate student in the Department of Artificial Intelligence and Machine Learning at Rasiklal M. Dhariwal Institute of Technology, Pune. Her academic interests include data preprocessing, machine learning, and computer vision. She has contributed to research work involving dataset analysis and model training strategies for deepfake detection.

4. Shivani khopkar

Shivani Khopkar is an undergraduate student in the Department of Artificial Intelligence and Machine Learning at Rasiklal M. Dhariwal Institute of Technology, Pune. Her interests include software development, UI/UX design, and project documentation. She focuses on presentation, deployment concepts, and ensuring user-friendly implementation of AI systems.