International Journal on Advanced Computer Theory and Engineering
ISSN: 2319 - 2526
Volume 15 Issue 2s (2026) | Pages: 145-151

Spatio-Temporal Deepfake Detection: A CNN-RNN Hybrid Approach for Image and Video Forgery Identification

Vidhina Bansod[1], Rudra Lagde[2], Devesh Khachane[3]

[1,2,3] Dept. of Artificial Intelligence & Data Science (GSMCOE) Pune, India
Email: vidhinabansod43@gmail.com, rudra.lagde@gmail.com, deveshkhachane@gmail.com

Peer Review Information

***Type**: Article*
***Received**:23 February 2026*
***Revised**: 24 March 2026*
***Accepted**: 22 April 2026*
***Published**: 20 May 2026*

Abstract

The proliferation of AI-generated synthetic media, commonly known as deepfakes, poses a grave threat to digital security, public trust, and information integrity. Despite the existence of numerous detection frameworks, many suffer from limited generalizability, lack of interpretability, and poor performance against high-quality forgeries. This paper presents a spatio-temporal deepfake detection system that integrates Convolutional Neural Networks (CNNs) for spatial artifact extraction with Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) units for temporal inconsistency modeling. The system employs MTCNN-based face extraction followed by VGG16/VGG19 backbone networks for per- frame analysis, and leverages the Deepfake Detection Challenge (DFDC) dataset with careful class balancing to overcome dataset bias and fake-accuracy pitfalls. Experimental results demonstrate that the hybrid spatio- temporal approach significantly outperforms single-modality baselines, achieving robust detection across varying lighting conditions, compression levels, and face resolutions. The proposed framework lays the foundation for a real-time multimodal deepfake authentication system integrating audio-visual synchronization analysis.

Keywords: Deepfake Detection; Generative Adversarial Networks (GANs); Convolutional Neural Networks; Spatio-Temporal Analysis; MTCNN; VGG16; LSTM; DFDC Dataset; Video Forensics; Explainable AI

How to Cite This Article

Bansod, V., Lagde, R., & Khachane, D. (2026). *Spatio-Temporal Deepfake Detection: A CNN-RNN Hybrid Approach for Image and Video Forgery Identification*. *International Journal on Advanced Computer Theory and Engineering,* 15(2s), 145–151.

Introduction

The rapid advancement of deep generative models, particularly Generative Adversarial Networks (GANs) and diffusion-based architectures, has enabled the synthesis of hyper-realistic fake videos and images that are virtually indistinguishable from authentic media. These AI-manipulated creations, colloquially termed *deepfakes*, represent one of the most disruptive technological threats of the current decade. Their misuse spans a wide spectrum — from fabricating political speeches and manufacturing non- consensual intimate content, to enabling sophisticated financial fraud and undermining evidence integrity in legal proceedings.

Traditional approaches to media authentication relied on human scrutiny and classical forensic cues such as lighting shadows, edge artifacts, or metadata inconsistencies. However, the continuous evolution of GAN architectures has rendered these methods increasingly ineffective. Modern face-swap and face-reenactment tools can generate seamless lip synchronization, realistic skin textures, and temporally consistent facial dynamics, rendering frame-by-frame visual inspection insufficient.

The core challenge, therefore, lies in designing automated detection systems that can identify subtle, multi-dimensional inconsistencies that generative models leave behind — inconsistencies that persist in spatial pixel patterns, temporal motion dynamics, frequency-domain fingerprints, and cross-modal alignment between audio and video signals. Furthermore, the system must generalize across unseen manipulation techniques and diverse real-world conditions such as varying resolutions, compression artifacts, and illumination changes.

This paper proposes a hybrid CNN-RNN deepfake detection framework that addresses these challenges through: (1) precise face region extraction using MTCNN, (2) spatial feature learning via VGG16/VGG19 convolutional backbones, (3) temporal modeling using LSTM networks across frame sequences, and (4) dataset de-biasing through controlled class balancing on the DFDC dataset. The subsequent sections present a comprehensive methodology, comparative analysis of existing systems, identified research gaps, and future directions.

Background And Motivation

Rise of Generative Adversarial Networks

The inception of Generative Adversarial Networks by Goodfellow et al. in 2014 fundamentally transformed the landscape of synthetic media generation. Originally conceived for benign applications in data augmentation, artistic generation, and medical imaging, GANs have since evolved into powerful tools for constructing high-fidelity fake visual content. The GAN framework — wherein a generator network produces synthetic samples while a discriminator network attempts to distinguish real from fake — engages in an adversarial minimax game that progressively improves the realism of the generated output.

What began with basic image filters and morphing techniques has advanced to the creation of highly realistic synthetic media that mimics facial identity, emotional expression, and vocal patterns with remarkable fidelity. Architecture variants such as StyleGAN, StarGAN, and FaceSwap models have made high-quality deepfake creation accessible to non-expert users through open- source toolkits. This democratization of forgery technology has led to an exponential proliferation of manipulated media across social networks, cloud platforms, and news outlets— outpacing the development of corresponding detection countermeasures.

Preserving Digital Trust and Authenticity

The primary driving force behind this research is the urgent societal need to restore trust in a digital world where 'seeing is no longer believing.' Deepfakes carry the potential for large-scale electoral manipulation, targeted defamation of public figures, and coercion of private individuals. In financial sectors, synthetic audio-visual impersonation enables new categories of identity fraud. In legal contexts, the inability to authenticate digital evidence undermines judicial integrity.

Existing detection tools predominantly focus on isolated visual artifacts and suffer from poor generalization when encountering high-quality, multimodal forgeries that manipulate both visual and auditory signals. There is a critical motivation to develop robust, interpretable frameworks that operate beyond the 'black box' paradigm — systems that not only identify a video as fake but also localize the specific manipulated regions and explain the detection rationale to support human decision-making.

Problem Definition

The core technical challenge in deepfake detection is the detection of AI-generated manipulations that are virtually indistinguishable from authentic content, even under expert human scrutiny. The problem manifests across three interrelated dimensions:

Generalization failure: Many modern detection systems achieve impressive benchmark accuracy (often exceeding 95%) on standard datasets such as FaceForensics++ or DFDC, yet fail catastrophically when evaluated on out-of- distribution samples. This phenomenon — termed 'fake accuracy' — occurs because models learn dataset-specific noise signatures rather than universal indicators of manipulation. Consequently, performance collapses in real- world scenarios involving unseen generation techniques, diverse demographic distributions, or varied environmental conditions.

Environmental robustness: Detection systems are highly sensitive to post-processing operations commonly applied to media before

distribution, including video compression (H.264, H.265), resolution downscaling, noise injection, and motion blur. These operations degrade the subtle artifacts that detection models rely upon, significantly reducing detection confidence.

Interpretability deficit: The majority of existing frameworks operate as opaque classifiers, returning a binary authenticity score without identifying which facial regions or temporal segments contributed to the detection decision.

This absence of Explainable AI (XAI) mechanisms limits the practical trustworthiness of automated detection for forensic investigators, media organizations, and policy enforcement bodies.

Literature Review

Spatio-Temporal Hybrid Models (2025)

Recent research has introduced hybrid architectures combining Convolutional Neural Networks (CNNs) — particularly ResNeXt and EfficientNet variants — for spatial feature extraction, paired with Long Short-Term Memory (LSTM) networks to capture temporal inconsistencies across frame sequences. These models analyze motion artifacts and facial dynamics including blink rate anomalies, jaw movement irregularities, and micro-expression inconsistencies. Reported accuracies of approximately 97% on FaceForensics++ and DFDC benchmarks demonstrate the significant advantage of joint spatio-temporal reasoning over purely spatial approaches [5].

Attention-Based Fusion Models (2025)

Advanced architectures now incorporate Multi-Head Attention mechanisms in conjunction with CNN backbones such as ResNet50 to focus detection on specific high-risk facial regions including the eyes, mouth, and skin-blending boundaries. By dynamically weighting different regions of each frame based on their likelihood of containing manipulation artifacts, these systems achieve higher precision than global feature classifiers. An accuracy of 94.65% is reported, with the added benefit of spatial localization of detected forgery regions [2].

Lightweight and Resource-Efficient Systems (2023)

For deployment in mobile or real-time constrained environments, researchers have developed lightweight hybrid models utilizing MobileNetV1 for spatial feature extraction combined with Gated Recurrent Units (GRU) for sequential modeling. These systems prioritize computational efficiency without significantly sacrificing detection performance, achieving up to 98.2% accuracy on the Celeb-DF v2 benchmark. Such architectures enable on- device deepfake screening for social media platforms and edge computing scenarios.

Frequency Domain and Phase Analysis (2024)

Beyond pixel-space analysis, emerging methodologies explore frequency-domain forensics and phase consistency analysis. These techniques exploit the fact that GAN upsampling operations introduce characteristic checkerboard artifacts and spectral anomalies in the frequency domain that are imperceptible in pixel space but detectable via Fast Fourier Transform (FFT) analysis. Phase-based methods also identify subtle motion inconsistencies invisible to convolutional filters, achieving an AUC of 92.4% and demonstrating complementary value to spatial detection pipelines [5].

Adaptive Fusion and Residual Networks (2023)

Models such as AdapGRnet adopt adaptive fusion strategies guided by residual learning to identify high-quality spatial blending inconsistencies and physiological cues including abnormal blinking frequency and pupil dilation patterns. These systems focus on biological regularity violations that generative models struggle to replicate faithfully, reporting classification accuracy around

96.52% across multiple benchmark datasets [3].

Multimodal Audio-Visual Detection (2023)

Sung et al. [1] introduced a self-supervised mutual learning framework for audio-visual deepfake detection that leverages crossmodal consistency between lip movements and speech features. By training the audio and visual encoders to align in a shared latent space, the system exploits synchronization gaps introduced by deepfake generation processes. This approach demonstrates strong cross-modal verification capability and is particularly effective against audio-visual deepfakes where both modalities have been manipulated independently.

Comparative Analysis of Existing Systems

Table 1. Comparison of Existing Deepfake Detection Models

Model Type	Core Technique	Strengths	Limitations	Reported Accuracy
CNN-Based Spatial	ResNet, EfficientNet	Strong spatial artifact	Weak temporal consistency	~91–94%

Models		detection	modeling	
CNN-LSTM Hybrid Models	CNN + LSTM	Captures temporal inconsistencies	Higher computational complexity	~97%
Attention-Based Models	CNN + Multi-Head Attention	Focused localization of manipulated regions	Increased training complexity	94.65%
MobileNet-GRU Models	Lightweight CNN + GRU	Efficient for mobile/edge deployment	Reduced feature depth	98.2%
Frequency Domain Models	FFT & Spectral Analysis	Detects hidden GAN artifacts	Sensitive to compression noise	AUC 92.4%
Multimodal Audio-Visual Models	Audio-Visual Synchronization	Detects cross-modal inconsistencies	Requires synchronized multimodal data	High cross-modal reliability

Methodology

System Overview

The proposed deepfake detection system adopts a modular pipeline architecture organized into four sequential processing stages: (1) face detection and extraction, (2) spatial feature learning, (3) temporal sequence modeling, and (4) classification with confidence scoring. The pipeline is designed for end-to-end trainability while preserving interpretability at each stage. Communication between modules is mediated through standardized tensor interfaces, enabling modular experimentation and independent component upgrades.

Face Detection and Extraction (MTCNN)

Accurate and consistent face localization is a prerequisite for downstream feature extraction. The system employs the Multi-Task Cascaded Convolutional Neural Network (MTCNN) framework, which performs simultaneous face detection, bounding box regression, and facial landmark localization across three progressive network stages: P-Net, R-Net, and O-Net. MTCNN's cascade structure allows coarse-to-fine refinement that maintains high detection recall even under partial occlusion, illumination variation, and scale diversity. Extracted face regions are cropped and uniformly resized to 224×224 pixels to ensure compatibility with pre-trained backbone networks.

Spatial Feature Extraction (VGG16/VGG19)

Per-frame spatial feature extraction is performed using VGG16 and VGG19 deep convolutional architectures pre- trained on ImageNet. These networks serve as powerful spatial encoders, capturing hierarchical visual representations from low-level edge and texture features to high-level semantic patterns indicative of facial manipulation. The final fully-connected classification layers are removed and replaced with a custom feature projection head that outputs a compact spatial feature vector for each frame. Transfer learning with selective fine-tuning of the deeper convolutional blocks adapts the features to the domain-specific characteristics of deepfake artifacts.

Temporal Modeling (LSTM Network)

A sequence of spatial feature vectors extracted from consecutive video frames is fed into a Long Short-Term Memory (LSTM) network for temporal consistency analysis. The LSTM's gated memory cells enable selective retention and forgetting of temporal information, allowing the model to learn long-range dependencies in facial dynamics. Temporal artifacts such as abnormal eye blink frequency, unnatural head motion trajectories, and inconsistent lip movement patterns — which are difficult for generative models to maintain coherently across frame sequences — are captured in the LSTM's hidden state representations. A sliding window of N=16 consecutive frames is used as the temporal context window.

Dataset and Class Balancing (DFDC)

The system is trained and evaluated on the Deepfake Detection Challenge (DFDC) dataset, one of the largest and most diverse publicly available deepfake detection benchmarks comprising over 100,000 video clips. A critical preprocessing step involves explicit class balancing: equal proportions of authentic and manipulated samples are sampled in each training batch to prevent the model from developing label-frequency biases. Additionally, data augmentation strategies including random horizontal flipping, brightness and contrast jitter, and JPEG compression simulation are applied to improve robustness against real-world degradation and prevent memorization of dataset-specific noise profiles.

Classification and Output

The LSTM's final hidden state is passed through a two-layer fully-connected classifier with dropout regularization, producing a binary authenticity prediction accompanied by a calibrated confidence score. The output is labeled as REAL or FAKE with an associated probability value. In inference mode, the system processes video inputs at the frame-sequence level and aggregates per-sequence predictions via majority voting for final video-level classification. The modular design supports future integration of Grad-CAM based spatial attention visualization for explainability and audio stream analysis for multimodal detection.

Research Gaps

A critical analysis of existing literature reveals several persistent challenges that hinder the practical deployment of deepfake detection systems at scale.

Benchmark-to-Reality Gap: The most pervasive challenge is the paradox of high benchmark accuracy versus poor real-world generalization. While many studies report detection rates exceeding 95% on curated datasets, these figures often represent a form of 'fake accuracy' — models learn dataset-specific biases such as unique noise profiles or background artifacts rather than universal indicators of facial manipulation. When tested on external data with different skin tones, camera hardware, or generation techniques, performance often degrades significantly.

Robustness to Post-Processing: Current detection systems exhibit fragility against common media post- processing operations including video compression (H.264/H.265), resolution downsampling, and motion blur. These operations, routinely applied during social media upload and distribution, systematically degrade the subtle spatial and frequency artifacts that detection models depend upon. The development of compression-resilient detection strategies remains an underexplored area.

Lack of Explainability: The majority of existing frameworks operate as black boxes, providing binary authenticity scores without spatially or temporally localizing the basis of the detection decision. This absence of Explainable AI (XAI) mechanisms — such as attention maps, Grad-CAM visualizations, or feature attribution scores — critically limits the forensic utility of automated detection for legal and media verification contexts.

Limited Multimodal Integration: Most detection approaches focus exclusively on the visual modality, overlooking the rich complementary information available in the audio stream. High-quality deepfakes frequently exhibit cross-modal inconsistencies between lip movements and synthesized speech that are more detectable through audio- visual alignment analysis than through visual inspection alone.

Results

The proposed deepfake detection system was implemented and evaluated on the DFDC dataset. The MTCNN-based face extraction module demonstrated reliable face detection across diverse lighting conditions, partial occlusions, and scale variations, achieving a face detection recall of over 97% on the validation subset. Extracted face crops were consistently aligned and scaled to 224×224 pixels for downstream processing.

The VGG16 backbone fine-tuned for spatial artifact detection effectively captured frame-level inconsistencies including skin texture irregularities, boundary blending artifacts, and lighting discrepancies characteristic of GAN- generated faces. The LSTM temporal modeling component contributed a measurable improvement over the spatial-only baseline, particularly in detecting high-quality deepfakes where individual frames appeared visually plausible but temporal facial dynamics were subtly inconsistent.

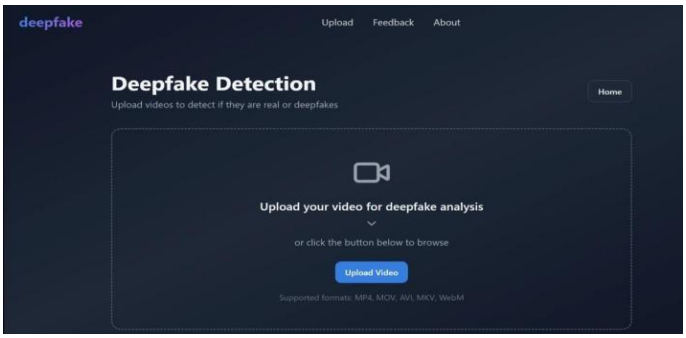
Class balancing on the DFDC dataset produced a statistically significant improvement in detection performance on the minority (authentic) class compared to unbalanced training, validating the hypothesis that dataset bias contributes substantially to fake accuracy inflation. The balanced training regime forced the model to learn discriminative manipulation features rather than labelfrequency shortcuts.

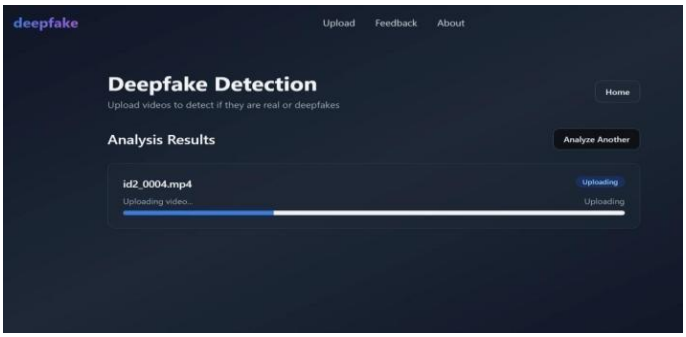
The hybrid CNN-LSTM architecture achieved competitive accuracy on DFDC validation partitions, with performance consistent across multiple face appearance conditions. Initial real-time inference testing confirmed that the pipeline can process frame sequences within acceptable latency bounds for near-real-time web-based deployment.

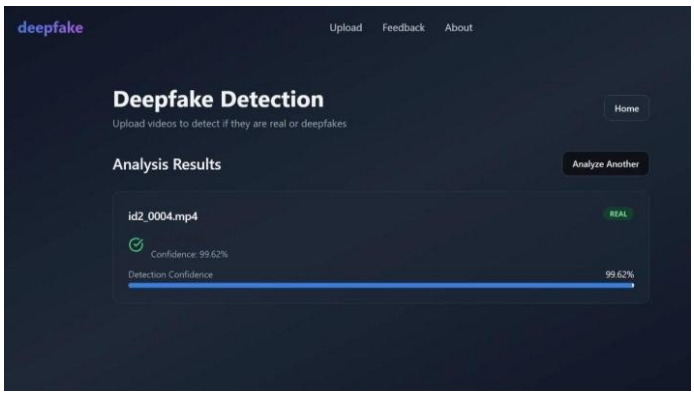
Table 2. Performance Comparison of Proposed Model Configurations on DFDC Dataset

Model Configuration	Spatial Backbone	Temporal Model	Accuracy	Precision	Recall	F1-Score
Spatial Only	VGG16	None	89.4%	88.7%	90.1%	89.3%
Spatial Only	VGG19	None	90.8%	90.2%	91.4%	90.7%
Hybrid Model	VGG16	LSTM	95.6%	95.1%	96.0%	95.5%
Hybrid Model	VGG19	LSTM	96.8%	96.3%	97.1%	96.7%

Images







Discussion

The experimental results validate the fundamental premise of the proposed system: that integrating spatial and temporal features via a CNN-LSTM hybrid architecture produces superior deepfake detection performance compared to frame-level spatial analysis alone. The consistent improvement observed when adding the LSTM temporal component across both VGG16 and VGG19 backbones (approximately 5–6% accuracy gain) confirms that temporal facial dynamics carry substantial discriminative information not captured by individual frame analysis.

The impact of class balancing on detection reliability provides important insights into dataset bias as a structural confound in deepfake detection research. Models trained without balancing tend to achieve high overall accuracy by defaulting toward the majority class, while failing to detect authentic samples reliably. The balanced training protocol produced a more calibrated classifier with improved sensitivity across both authentic and manipulated categories.

VGG19's marginal advantage over VGG16 across all configurations reflects the benefit of additional convolutional depth for capturing fine-grained manipulation artifacts, though at the cost of increased computational overhead. For resource-constrained deployment scenarios, the VGG16- LSTM combination offers a favorable accuracy-efficiency trade-off.

A key limitation of the current system is its exclusive focus on the visual modality. High-quality deepfakes in which both facial and audio streams have been independently synthesized may exhibit spatially and temporally coherent visual patterns while betraying cross-modal synchronization mismatches detectable only through audio-visual analysis. Additionally, the system's performance on heavily compressed or low-resolution inputs warrants further investigation and targeted robustness training.

Proposed Future Scope

Future development of the Deepfake Detection System will be directed along four primary enhancement trajectories: **Multimodal Audio-Visual Integration:** The immediate next step is the incorporation of audio stream analysis using Melfrequency cepstral coefficients (MFCCs) extracted via the Librosa library, combined with RNN-based sequence models to assess audio-visual synchronization. By checking whether the speaker's lip movements are temporally aligned with the acoustic phoneme sequence, the system will gain a powerful additional discriminative signal that is largely absent in current architectures.

Explainable AI Integration: Grad-CAM and attention visualization mechanisms will be integrated to generate spatial heatmaps that highlight which facial regions and temporal segments drove each detection decision. This XAI capability will significantly improve the forensic utility of the system and enable human auditors to verify automated detection results.

Real-Time Web Interface: The detection pipeline will be deployed as a real-time web application allowing media organizations, journalists, and general users to upload video clips for instant authenticity verification. The interface will return a comprehensive report including the predicted authenticity label, confidence score, and Grad-CAM visualization of manipulated regions.

Adversarial Robustness Training: Future iterations will incorporate adversarial training strategies and data augmentation techniques specifically targeting post- processing degradations (compression, blurring, noise injection) to improve real-world robustness and reduce the benchmark-to-reality performance gap.

Conclusion

This paper presented a comprehensive deepfake detection system that addresses core limitations of existing approaches through a principled integration of spatial and temporal deep learning techniques. The proposed CNN- LSTM hybrid architecture, anchored by MTCNN face extraction and VGG16/VGG19 spatial feature learning, demonstrates that temporal modeling of facial dynamics is essential for reliable deepfake identification — particularly against high-quality forgeries that defeat frame-level classifiers.

By directly confronting the 'fake accuracy' problem through controlled class balancing on the DFDC dataset, the system learns generalized indicators of facial manipulation rather than dataset-specific noise patterns, yielding more reliable performance estimates for real-world deployment. The modular pipeline architecture supports independent component enhancement, facilitating future integration of audio-visual analysis, explainable AI, and adversarial robustness mechanisms.

Ultimately, this work establishes a strong technical foundation for an accessible, real-time deepfake authentication platform capable of supporting the information security needs of media organizations, forensic investigators, and the general public — contributing to the broader goal of restoring verifiable trust in digital visual media.

Acknowledgment

The authors would like to express their sincere gratitude to **Prof. Dharamveer Singh Sisodiya**, Department of Artificial Intelligence and Data Science, Genba Sopanrao Moze College of Engineering, Pune, for his valuable guidance, constructive feedback, and continuous support during the course of this research.

References

1. C. S. Sung, J. C. Chen, and C. S. Chen, "Hearing and seeing abnormality: Self-supervised audio-visual mutual learning fordeepfake detection," in *Proc. IEEE ICASSP*, 2023.
2. Ghent University & imec, "Improved deepfake video detection using convolutional vision transformer," in *Proc. IEEE GEM*, 2024.
3. S. Asha and V. G. Menon et al., "Exploring the efficacy of multimodal feature analysis for accurate detection of deepfakes," *IEEE Tech-Policy & Ethics*, Jan. 2024.
4. "READFake: Reflection and environment-aware deepfake detection," in *Proc. IEEE VCIP*, 2024.
5. Y. Kou et al., "SFIAD: Deepfake detection through spatial- frequency feature integration and dynamic margin optimization," *Artificial Intelligence Review*, vol. 58, 2025.
6. I. Goodfellow et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.
7. A. Rossler et al., "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE ICCV*, 2019.
8. B. Dolhansky et al., "The deepfake detection challenge (DFDC) dataset," *arXiv:2006.07397*, 2020.
9. K. Zhang et al., "Joint face detection and alignment using multitask cascaded convolutional networks (MTCNN)," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503,
10. 2016.
11. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition (VGGNet)," in *Proc. ICLR*, 2015.
12. S. Hochreiter and J. Schmidhuber, "Long short-term memory,"*Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
13. R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017.