Ministry of Science and Higher Education of the Russian Federation

FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION OF HIGHER EDUCATION

National Research University ITMO

(University ITMO)

Faculty: Information Security

Educational program: Master's in Information Security

Field of study (specialty): 10.04.01 Information Security

REPORT
of the research work

Name of the topic: **Analysis of DeepFake video editing and Detection Techniques**
Student: **Juan Pablo Sierra Useche, N4150c**

Agreed:
Thesis supervisor: **Коржук Виктория Михайловна, ITMO University, Faculty of Secure Information Technologies, associate professor (qualification “full associate professor”)**

Research work completed with a grade ___________

Date **2026-01-24**

Saint Petersburg

2026

CONTENT

Introduction 2
Chapter 1: Deepfake Video Editing and Manipulation Techniques 3
1.1 Types of Video Deepfakes 3
1.2 Partial versus Full Video Synthesis 3
1.3 Object Insertion, Removal, and Scene Alteration 4
1.4 Spatiotemporal Consistency Challenges 4
Chapter 1 Summary 5
Chapter 2: Forensic Features and Detection Signals 6
2.1 Spatial Artifacts in Manipulated Frames 6
2.2 Temporal Inconsistencies Across Frames 6
2.3 Frequency-Domain and Compression Traces 7
2.4 Physical Inconsistencies: Motion, Lighting, and Shadows 7
2.5 Cross-Frame Coherence and Identity Preservation 7
Chapter 2 Summary 9
Chapter 3: Machine Learning Approaches for Detection 10
3.1 CNN-Based Frame Analysis 10
3.2 Temporal Models: RNNs, LSTMs, and Transformers 10
3.3 Spatiotemporal Architectures 11
3.4 Dataset Dependency and Generalization Challenges 11
3.5 Robustness to Compression and Post-Processing 12
Chapter 3 Summary 13
Chapter 4: Detection Errors, Trust Implications, and Evaluation Challenges . . . 14
4.1 False Positives in Authentic Videos 14
4.2 False Negatives in Edited Deepfakes 14
4.3 Trust Degradation in Forensic Contexts 15
4.4 Dataset Bias and Evaluation Limitations 15
4.5 Practical Constraints in Deployment 16
Chapter 4 Summary 17
Conclusion 18
References 19

INTRODUCTION

Courts admit video as evidence. Journalists publish footage to document events. People share clips expecting them to show what actually happened. This trust assumes video captures reality. Generative AI breaks that assumption. Tools that once required professional expertise now let anyone swap a face, change a license plate, or alter a timestamp while leaving the rest of the frame untouched. A prosecutor cannot tell if the defendant's face was swapped onto someone else's body. A viewer cannot tell if the politician actually said those words.

Current deepfake detectors do not help much with partial manipulations. They output a single probability for the whole video. When 95% of pixels are authentic and 5% contain a fabricated face, the detector often misses it. Even when detection succeeds, the system cannot say which part was faked or why it flagged the video.

This research investigates element-level deepfake detection with selective confidence estimation: systems that identify specific manipulated objects within frames and refuse to make predictions when uncertainty is high.

The objectives are:

a) Analyze existing deepfake detection methods and their limitations for partial manipulation scenarios.
b) Review approaches for manipulation localization and object-level analysis in video.
c) Examine uncertainty quantification and selective prediction techniques applicable to evidence verification.
d) Structure collected materials as groundwork for developing an element-wise detection framework.

The expected outcome is a conceptual foundation for deepfake detection that prioritizes reliability and interpretability over whole-video classification accuracy.

CHAPTER 1: DEEPFAKE VIDEO EDITING AND MANIPULATION TECHNIQUES

Video manipulation has transformed dramatically. What once required professional editors and weeks of work now happens in minutes through AI. The term “deepfake” originally described AI face swaps but has expanded to cover all neural network-based video manipulation. GANs dominated early work; diffusion models now produce even more convincing results with different artifact signatures [1], [2]

1.1 Types of Video Deepfakes

Face-centric deepfakes dominate research because faces carry identity—they determine who we trust and believe. Four main subcategories exist [1]:

Face swapping replaces one person’s identity with another while preserving lighting, pose, and background. The boundary where synthetic meets authentic remains forensically significant. **Face reenactment** keeps identity constant but transfers motion—expressions, head movements, gaze—from a source video. Implementations use 3DMM models and latent feature decoupling. **Facial attribute editing** modifies age, makeup, or expression without changing identity. Subtler than swapping, harder to detect. **Face super-resolution** enhances low-quality imagery but can trigger false positives when enhancement textures mimic manipulation signatures.

Beyond faces, body animation techniques extend manipulation to posture and gesture [1]. An alternative taxonomy organizes forgery by scope: intra-frame (morphing, inpainting), inter-frame (insertion, deletion, shuffling), and spatiotemporal (combined) [3].

1.2 Partial versus Full Video Synthesis

Partial manipulation modifies specific regions while leaving surroundings intact —face swapping being the prime example [1]. The boundary between synthetic and authentic content creates forensic opportunity through mismatched noise distributions, compression histories, and textures. Inpainting fills targeted regions with synthesized content, enabling evidence removal [4].

Full video synthesis generates everything artificially [2]. Systems like Sora create content from text prompts alone. No authentic reference exists for comparison. Detection must identify generative properties rather than boundary discontinuities.

Temporal manipulation alters frame sequences without modifying individual frames [3]. Deletion removes evidence; insertion adds fabricated events; reordering changes apparent causation. Each frame may pass authenticity tests while the sequence fails coherence analysis.

1.3 Object Insertion, Removal, and Scene Alteration

Object manipulation enables evidence tampering. Inpainting removes specified elements—people, vehicles, documents—filling voids with plausible content [4], [5]. Different approaches leave different artifacts: diffusion methods over-smooth, exemplar methods create repetitions, deep learning methods introduce checkerboard patterns.

Copy-move duplicates content internally; splicing integrates external material [3], [6]. Scene alteration modifies object properties—license plates, clothing colors, weapon appearances—while maintaining overall plausibility.

1.4 Spatiotemporal Consistency Challenges

Video represents continuous observation of a 3D world through time. Spatial consistency requires perspective-appropriate proportions, realistic texture variation, and physically coherent lighting [6], [7]. Temporal consistency demands plausible motion trajectories, smooth evolution, and stable identity [1].

Early methods processed frames independently, producing flickering and motion discontinuities. Temporal modeling improved quality substantially but artifacts persist [1]. The L3DE framework trains models to distinguish videos satisfying real-world constraints from those violating them [7].

Motion analysis through optical flow detects temporal anomalies—unnatural relative motion, impossible velocities, apparent movement in static regions [3], [7]. Even advanced systems produce subtle inconsistencies: minute flickering, marginal physics violations. Generation improves continuously. Detection must keep pace.

CHAPTER 1 SUMMARY

Video deepfakes encompass far more than face swaps. The taxonomy includes identity replacement, motion transfer, attribute modification, and full scene generation—each with distinct technical characteristics and forensic implications. Face-centric manipulations dominate research attention because faces carry trust, but object manipulation and temporal tampering enable equally consequential evidence fabrication.

The distinction between partial and full synthesis reshapes detection strategy fundamentally. Partial manipulation—face swapping, inpainting, splicing—creates boundaries where synthetic meets authentic. These boundaries leak forensic information: mismatched noise, divergent compression histories, inconsistent textures. Full synthesis offers no such boundaries. Everything is generated. Detection must find properties intrinsic to the generation process rather than discontinuities with authentic surroundings.

Spatiotemporal consistency provides the common thread across manipulation types. Video depicts a 3D world evolving through time. That structure imposes constraints. Perspectives must be geometrically valid. Shadows must follow light sources. Motion must respect physics. Identity must remain stable. Manipulation that violates these constraints—however subtly—becomes detectable. The challenge lies in building systems capable of recognizing violations that human observers miss.

CHAPTER 2: FORENSIC FEATURES AND DETECTION SIGNALS

Manipulation leaves traces. The forensic challenge lies in identifying which traces reliably indicate tampering versus natural video characteristics. Detection signals span multiple domains: spatial artifacts within frames, temporal inconsistencies across sequences, frequency-domain signatures, physical constraint violations, and coherence failures.

2.1 Spatial Artifacts in Manipulated Frames

Blending boundaries appear where manipulated regions meet authentic content [1]. Face swapping creates perimeter artifacts despite feathering and color harmonization. Subtle color shifts, texture discontinuities, and geometric misalignments persist at fusion boundaries.

Noise distribution anomalies provide powerful indicators [1], [3]. Camera sensors produce characteristic noise patterns throughout authentic footage. Manipulation disrupts local noise statistics. Sensor Pattern Noise (SPN) and Color Filter Array (CFA) artifacts offer device-specific signatures that forgeries disturb.

Compression artifacts reveal tampering when DCT coefficient distributions shift unexpectedly [6], [8]. Block-level inconsistencies appear at macroblock edges. Error Level Analysis (ELA) visualizes compression behavior differences between manipulated and authentic regions [4].

GANs produce checkerboard patterns from transpose convolutions; diffusion models leave different signatures [1], [5]. Unrealistic coloring, over-saturation, and geometric deformations characterize AI-generated content [2], [7].

2.2 Temporal Inconsistencies Across Frames

Inter-frame tampering—insertion, deletion, shuffling—disrupts natural temporal dependencies [3], [6]. Motion vector anomalies manifest as sudden position jumps, impossible velocity changes, or reversed direction without cause. Optical flow analysis quantifies these discontinuities.

Jitter and flickering characterize frame-by-frame manipulation [1], [4]. Independent processing produces slight inter-frame variations that accumulate into visible oscillations. Frame discontinuity appears when adjacent frames differ more than physical evolution permits.

Static scenes pose detection challenges [3]. With minimal motion, frame manipulation leaves few visible traces. Detection must rely on noise evolution, compression progression, or sensor-level temporal signatures.

ConvLSTM architectures accumulate evidence across frames, identifying regions behaving inconsistently over time [4]. Cross-frame attention mechanisms relate information across temporal distance [2].

2.3 Frequency-Domain and Compression Traces

DCT analysis reveals manipulation through coefficient distribution changes [3], [5], [6]. Double compression produces statistical patterns distinct from single-pass encoding. Quantization errors create periodic traces when compression parameters differ between original and re-encoded content [4].

High-frequency anomalies from GAN upsampling appear as spectral peaks at frequencies corresponding to upsampling factors [1], [2]. Constrained convolutional layers with high-pass filters extract these forensic traces while suppressing irrelevant semantic content [8].

H.264 bitstream analysis examines encoding parameters and motion vectors without full decoding [6]. Inconsistent quantization settings or anomalous motion prediction patterns indicate potential tampering.

2.4 Physical Inconsistencies: Motion, Lighting, and Shadows

Lighting direction must follow consistent physical principles within authentic footage [1], [5], [7]. Manipulated content may exhibit shadows pointing different directions, mismatched reflections, or contradictory brightness gradients. Shadow consistency provides strong forensic evidence—shadows follow strict geometric relationships with light sources.

Motion must respect physics [3], [7]. AI-generated video produces impossible trajectories, velocity changes without force, movements exceeding physical limits. Physics violations extend to interactions: unnatural fluid behavior, objects passing through surfaces, collisions without momentum transfer.

Biological signals constrain face manipulation [1]. Blinking follows characteristic frequency patterns. Gaze direction must follow possible eye movements. Head pose must respect anatomical limits. Mouth movements must match speech physics.

2.5 Cross-Frame Coherence and Identity Preservation

Identity stability matters for face manipulation [1], [2]. Authentic video maintains consistent facial identity with natural variations in expression, pose, and lighting.

Imperfect manipulation produces drift—gradual appearance shifts—or instability—flickering features.

Object permanence characterizes authentic video [2], [7]. Objects enter frames, move consistently, exit appropriately. Generated video may violate permanence through spontaneous appearance, disappearance, or inconsistent transformation.

Style and feature consistency reveal manipulation when synthesis fails to maintain uniform visual characteristics [1]. Mismatched color palettes, inconsistent noise, varying detail levels indicate regions of different origin. Decoupling identity from manipulation artifacts improves detection generalization [6].

CHAPTER 2 SUMMARY

Forensic analysis draws from multiple signal domains, each revealing different manipulation traces. Spatial artifacts—blending boundaries, noise anomalies, compression inconsistencies—expose tampering within individual frames. GANs leave checkerboard patterns; diffusion models produce distinct signatures. Temporal analysis catches what spatial examination misses: the jitter of frame-by-frame processing, motion vector discontinuities from inserted or deleted frames, flickering that betrays independent rather than coherent generation.

Frequency-domain forensics operates where human vision fails. DCT coefficient distributions shift after manipulation and recompression. GAN upsampling creates spectral peaks invisible to casual inspection but obvious under Fourier analysis. Physical constraint violations offer another avenue entirely—shadows pointing wrong directions, impossible trajectories, biological signals that don't match human physiology.

No single signal type suffices. Sophisticated manipulation may satisfy spatial constraints while violating temporal ones. Compression may obscure frequency signatures while preserving physical inconsistencies. Robust detection requires integration across domains, combining complementary evidence to catch what any single approach would miss.

CHAPTER 3: MACHINE LEARNING APPROACHES FOR DETECTION

Machine learning translates forensic knowledge into automated detection. The field has evolved from frame-level CNN analysis through temporal sequence modeling to integrated spatiotemporal architectures. Throughout this progression, generalization remains the central challenge—models that excel on training data often fail on novel manipulations.

3.1 CNN-Based Frame Analysis

CNNs provide foundational feature extraction for deepfake detection [1]. The same architectural elements enabling GANs to synthesize faces—convolutional layers, encoder-decoder structures—can identify synthesis artifacts. VGG, ResNet, and Xception architectures adapted from image classification serve as detection backbones [3], [4], [6].

Constrained convolutional layers focus attention on forensic traces rather than semantic content [8]. High-pass filter initialization emphasizes high-frequency components where manipulation artifacts often reside. Dual-stream architectures combine forensic noise analysis with visual context processing—one stream extracts noise patterns and compression artifacts, another processes scene content and lighting.

CNNs enable manipulation localization through probability masks indicating where forgery occurred [8]. Perceptual similarity metrics (LPIPS) and identity embeddings (ArcFace) provide supporting measurements [9].

Frame-by-frame analysis has fundamental limitations [10]. Processing frames independently discards temporal information. Manipulation maintaining frame-level plausibility while violating temporal coherence evades single-frame detection. Performance depends heavily on training-test distribution match.

3.2 Temporal Models: RNNs, LSTMs, and Transformers

Recurrent networks process video as sequences, maintaining hidden states carrying information across frames [3], [6]. Bidirectional LSTMs incorporate context from both past and future frames. ConvLSTM integrates convolution into recurrence, processing spatiotemporal features jointly [4].

Transformer self-attention relates all sequence elements directly rather than propagating through hidden states [1], [10]. TimeSformer and VideoSwin apply transformer architectures to video understanding. In-and-Across Frame Attention (IAFA) structures attention for both spatial and temporal dimensions [2].

Temporal coherence losses from generation research illuminate what properties manipulation attempts to maintain—and where it fails [9]. Detection can target precisely these temporal properties.

3.3 Spatiotemporal Architectures

3D CNNs extend 2D convolutions to operate over space and time simultaneously [3], [6], [10]. I3D, SlowFast, and C3D employ kernels spanning spatial and temporal dimensions. These architectures capture motion and temporal evolution that neither spatial nor temporal analysis alone accesses.

Spatiotemporal models exhibit detection patterns uncorrelated with visual quality [10]. They identify motion inconsistencies beyond what quality metrics measure. Trident networks fuse parallel streams—SRM for spatial forensics, 3D convolutions for temporal dynamics [6].

The L3DE framework uses 3D convolutions for appearance, motion, and geometry analysis [7]. VIDNet combines VGG spatial encoding with ConvLSTM temporal decoding [4]. SlowFast processes multiple temporal resolutions simultaneously.

Computational cost limits deployment. 3D convolutions require substantially more resources than 2D, constraining real-time applications.

3.4 Dataset Dependency and Generalization Challenges

Current detection suffers severe generalization limitations [10]. Models trained on specific manipulation methods learn particular artifacts rather than generalizable forensic features. Cross-model evaluation shows detection success depends more on training-test statistical match than model sophistication.

Detectors achieving high accuracy on FaceForensics++ often fail on other datasets [1]. StyleGAN inherits biases from training data (FFHQ), and detection systems may learn these biases rather than manipulation signatures [9].

CLIP-based backbones demonstrate improved generalization over traditional CNNs [10]. Large-scale vision-language pretraining creates broader visual representations. General-purpose detection systems targeting diverse attack types achieve more consistent cross-dataset performance [8].

Comprehensive datasets spanning diverse manipulations, demographics, and sources remain research priorities [3]. The AIGVDBench covers 31 generation models with 440,000+ videos [10].

3.5 Robustness to Compression and Post-Processing

Compression acts as adversary to forensic analysis [1], [6]. Aggressive compression scrubs subtle traces distinguishing manipulated from authentic content. Detection systems performing well on high-quality video may fail after compression.

H.264 standardization ensures consistent compression across training data, eliminating spurious correlations [10]. Modern codecs apply content-adaptive compression varying by region, complicating analysis [8]. Region-aware quality estimation weights analysis based on trace reliability.

Face restoration techniques can remove manipulation artifacts, functioning as anti-forensic processing [9]. Compression-aware training with augmented data improves real-world robustness [1], [3].

CHAPTER 3 SUMMARY

Detection architectures have progressed through three generations. Frame-level CNNs extract spatial features—blending artifacts, noise anomalies, compression traces—but discard temporal information entirely. A manipulation that looks plausible frame-by-frame can still violate temporal coherence. RNNs, LSTMs, and transformers address this gap by modeling sequence dependencies. They catch flickering, motion discontinuities, and identity drift that single-frame analysis misses.

Spatiotemporal architectures represent the current frontier. 3D CNNs process space and time jointly, learning features that neither dimension reveals alone. SlowFast captures multiple temporal resolutions. Trident networks fuse parallel spatial and temporal streams. These approaches detect motion inconsistencies uncorrelated with visual quality—subtle violations invisible to quality metrics but revealing to properly trained models.

Yet architecture sophistication cannot overcome data limitations. Models trained on FaceForensics++ fail on novel datasets. Detectors learning StyleGAN artifacts miss diffusion-generated content. CLIP-based backbones show promise through broader pretraining, but generalization remains fundamentally unsolved. Compression compounds the problem—aggressive encoding scrubs the very traces detection relies upon. Real-world deployment demands robustness that current benchmarks don't measure and current models don't achieve.

CHAPTER 4: DETECTION ERRORS, TRUST IMPLICATIONS, AND EVALUATION CHALLENGES

Detection accuracy on benchmarks tells an incomplete story. Real-world deployment introduces asymmetric error costs, extreme class imbalance, and trust dynamics that laboratory evaluation ignores. Understanding these factors determines whether detection technology helps or harms.

4.1 False Positives in Authentic Videos

False positives—flagging authentic content as manipulated—create serious problems. Vision-Language Models sometimes hallucinate artifacts in legitimate footage [10]. Low-quality authentic video mimics generation artifacts. DeepSeek-VL2 and similar systems cannot reliably distinguish poor capture quality from synthetic content.

Human detection shows analogous patterns [11]. Viewers mistake unusual lighting, awkward angles, or compression artifacts for deepfake signatures. Anger increases suspicion. Warning labels raise general distrust without improving discrimination.

The base-rate fallacy devastates operational performance [12]. Deepfakes are rare—anomalies among vast authentic content. Even 99% accuracy produces overwhelming false positive volumes at scale. Call center simulation: 333 flagged calls, 332 authentic, 1 actual deepfake. The detection system becomes operationally useless despite strong benchmark performance.

Detection fails on unusual but legitimate content [8]. Small regions, poor illumination, heavy texture provide insufficient forensic information. Systems produce confident but incorrect predictions rather than abstaining. Extreme conditions—heavy blur, noise, unusual capture—systematically generate false positives [2].

Frame manipulation using authentic frames from the same video poses particular challenges [3]. Shuffled or duplicated genuine frames share statistical properties with unmanipulated footage. Distinguishing temporal manipulation from benign editing or compression artifacts increases false positive risk.

4.2 False Negatives in Edited Deepfakes

False negatives—missing actual manipulation—enable the harms deepfakes threaten: fraud, defamation, evidence fabrication. Human detection proves inadequate: up to 70% of participants failed to identify fakes in some studies [11].

Audio-only deepfakes evade human detection particularly often. Lip-syncing is harder to detect than face swapping despite both being face-centric.

Sophisticated manipulation minimizes forensic artifacts. As generation improves, the gap between synthetic and authentic narrows. Detection trained on older methods misses newer techniques entirely. Processing pipelines that downsample for efficiency lose subtle forgery traces [2].

Static scenes enable temporal manipulation without visible discontinuity [3]. Frame deletion or insertion in low-motion footage leaves minimal traces. Each frame remains authentic; only the sequence is manipulated.

False positive and false negative rates trade off. Reducing one typically increases the other. Appropriate balance depends on deployment context and relative error costs.

4.3 Trust Degradation in Forensic Contexts

Deepfake capability affects trust in all digital media, not just manipulated content. Hyper-realistic AI video contributes to “synthetic skepticism”—increasing doubt about legitimate recordings [10].

The “Liar’s Dividend” lets bad actors dismiss authentic evidence as fabricated [12]. When any video could be fake, genuine recordings can be denied without evidence of manipulation. This defense has already appeared in legal proceedings. Courts traditionally treated video as reliable evidence; deepfake capability challenges that foundation.

Journalism faces parallel pressures [3]. Video documentation becomes uncertain. Rapid publication decisions occur without sophisticated forensic access. Unreliable detection might suppress legitimate footage or inadvertently publish manipulations.

Detection systems promising reliability but delivering uncertainty may harm trust rather than help [2]. High false positive rates undermine confidence by incorrectly flagging authentic content. High false negative rates fail when manipulations are later revealed. Context-aware detection with uncertainty quantification better supports appropriate trust calibration [12].

4.4 Dataset Bias and Evaluation Limitations

Benchmarks suffer semantic and distributional biases that skew evaluation [10]. Detection systems exploit benchmark-specific patterns rather than learning general manipulation signatures. High accuracy reflects memorization, not generalization.

Inverted class distributions create fundamental mismatch [12]. Training datasets balance authentic and manipulated content; real applications encounter vastly more authentic material. Models optimized for balanced data fail under actual prevalence conditions.

Limited metrics obscure operational failures [12]. EER and AUC appear favorable while hiding problematic threshold behavior. Researchers report strong aggregate scores without examining real-world distribution performance.

Standardization challenges fragment research [11]. Diverse metrics—Likert scales, binary classifications, confidence ratings—resist comparison. Narrow focus on identity-swap deepfakes leaves other manipulation categories under-examined.

New datasets address emerging gaps. DVF covers diffusion-generated video [2]. AIGVDBench spans 31 generation models with 440,000+ videos [10]. But concentration on few datasets (SULFA, VTL) limits generalizability assessment [3].

4.5 Practical Constraints in Deployment

Resource barriers affect both generation and detection [10]. Comprehensive video analysis requires computational resources unavailable in all contexts. Sophisticated methods may be too intensive for real-time content moderation [3].

Resolution limitations constrain implementations [8]. Systems limited to 1080p cannot process higher-resolution content at native quality. Model distillation creates smaller, faster versions with accuracy tradeoffs [9].

Human verification costs matter operationally [12]. Systems generating many flags—true or false—require analyst review. False positive floods make deployment impractical regardless of technical capability.

Platform integration demands processing enormous volumes with minimal latency [11]. Detection adding significant time or requiring specialized infrastructure may be infeasible at scale. Automated protections remain incompletely integrated into platforms.

Humans remain the “first line of defense” because AI detection has clear limitations [11]. While research advances, practical deployment relies substantially on human judgment. Regional restrictions and proprietary limitations prevent comprehensive evaluation of all generation methods [10].

CHAPTER 4 SUMMARY

Benchmark accuracy deceives. A detector achieving 99% on balanced test data becomes operationally useless when deepfakes constitute 0.3% of real traffic—332 false positives for every true detection. The base-rate fallacy transforms impressive laboratory performance into deployment failure. False positives damage legitimate creators and erode trust in detection itself. False negatives enable the frauds, defamations, and evidence fabrications that motivate detection research in the first place.

Trust dynamics extend beyond individual classification errors. The mere existence of convincing manipulation capability changes how all video evidence is perceived. The "Liar's Dividend" allows dismissal of authentic recordings without proof of fabrication—a defense already deployed in courtrooms. Detection systems promising certainty but delivering ambiguity may accelerate rather than reverse this trust erosion. Unreliable tools are worse than no tools at all.

Evaluation frameworks fail to capture these dynamics. Benchmarks suffer distributional biases that reward memorization over generalization. Metrics like EER and AUC hide operationally critical threshold behavior. Training on balanced datasets prepares models for conditions they'll never encounter. Meanwhile, practical constraints—computational costs, resolution limits, latency requirements, human review capacity—further constrain what laboratory-proven methods can actually deliver. Detection matters. But only if it works reliably, communicates uncertainty honestly, and integrates appropriately into verification workflows where stakes are real and errors have consequences.

CONCLUSION

Generative AI has undermined video as evidence. Face swaps, altered license plates, and modified timestamps now pass casual inspection. Courts and journalists can no longer assume footage is authentic.

This research examined element-level deepfake detection with selective confidence estimation. The goal: build the theoretical foundation before experiments begin.

The investigation traced manipulation methods through detection signals to deployed systems. Deepfake techniques now include face replacement, motion transfer, object synthesis, and temporal reordering. Each leaves distinct traces. Spatial artifacts appear at edit boundaries. Frequency anomalies emerge in synthetic regions. Physical violations break lighting consistency. Compression mismatches mark where authentic and fabricated content meet.

Detection architectures have evolved to catch these signals. Frame-level CNNs gave way to spatiotemporal models with attention mechanisms. Yet cross-dataset tests expose a persistent gap: benchmark scores do not predict real-world performance. Deployed systems face even harder problems. The base-rate fallacy turns 99% accuracy into floods of false positives. The “Liar’s Dividend” lets people dismiss authentic evidence as fabrication.

These problems demand a different approach. Whole-video classification fails when only a face or a timestamp was manipulated. The forensic signal gets diluted across millions of authentic pixels. Binary outputs tell users nothing about which regions look suspicious or why. High-confidence predictions on uncertain inputs destroy trust when they turn out wrong.

Element-level detection with selective confidence estimation solves these problems. Identifying specific manipulated objects within frames preserves signal strength for partial fakes. Abstaining when uncertainty runs high avoids the confident errors that erode trust. Showing which elements triggered concern, and with what confidence, supports human verification when errors carry consequences.

This research establishes the theoretical and methodological foundation for such a framework. It identifies what detection must recognize, what signals detection can exploit, what architectures can implement detection, and what conditions detection must satisfy. From this groundwork, experimental investigation of element-wise detection with uncertainty quantification can proceed.

REFERENCES

[1] G. Pei *et al.*, “Deepfake Generation and Detection: A Benchmark and Survey.” Accessed: Jan. 22, 2026. [Online]. Available: http://arxiv.org/abs/2403.17881

[2] X. Song *et al.*, “On Learning Multi-Modal Forgery Representation for Diffusion Generated Video Detection.” Accessed: Jan. 23, 2026. [Online]. Available: http://arxiv.org/abs/2410.23623

[3] M. M. Ali, N. I. Ghali, H. M. Hamza, K. M. Hosny, E. Vrochidou, and G. A. Papakostas, “Interframe Forgery Video Detection: Datasets, Methods, Challenges, and Search Directions,” *Electronics*, vol. 14, no. 13, p. 2680, July 2025, doi: 10.3390/electronics14132680.

[4] P. Zhou, N. Yu, Z. Wu, L. S. Davis, A. Shrivastava, and S. N. Lim, “Deep Video Inpainting Detection,” 2021.

[5] A.-A. Barglazan, R. Brad, and C. Constantinescu, “Image Inpainting Forgery Detection: A Review,” *Journal of Imaging*, vol. 10, no. 2, p. 42, Feb. 2024, doi: 10.3390/jimaging10020042.

[6] A. Diwan, S. Dixit, R. Subbiah, and R. Mahadeva, “Systematic analysis of video tampering and detection techniques,” *Cogent Engineering*, vol. 11, no. 1, p. 2424466, Dec. 2024, doi: 10.1080/23311916.2024.2424466.

[7] C. Chang *et al.*, “How Far are AI-generated Videos from Simulating the 3D Visual World: A Learned 3D Evaluation Approach.” Accessed: Jan. 23, 2026. [Online]. Available: http://arxiv.org/abs/2406.19568

[8] T. D. Nguyen, S. Fang, and M. C. Stamm, “VideoFACT: Detecting Video Forgeries Using Attention, Scene Context, and Forensic Traces,” in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, Jan. 2024, pp. 8548–8558. doi: 10.1109/WACV57701.2024.00837.

[9] A. Melnik *et al.*, “Face Generation and Editing with StyleGAN: A Survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3557–3576, May 2024, doi: 10.1109/TPAMI.2024.3350004.

[10] L. Ma *et al.*, “Your One-Stop Solution for AI-Generated Video Detection.” Accessed: Jan. 23, 2026. [Online]. Available: http://arxiv.org/abs/2601.11035

[11] K. Somoray, D. J. Miller, and M. Holmes, “Human Performance in Deepfake Detection: A Systematic Review,” *Human Behavior and Emerging*

Technologies, vol. 2025, no. 1, p. 1833228, Jan. 2025, doi: 10.1155/hbe2/1833228.

[12] S. Layton, T. Tucker, D. Olszewski, K. Warren, K. Butler, and P. Traynor, "SoK: The Good, The Bad, and The Unbalanced: Measuring Structural Limitations of Deepfake Media Datasets."