International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 02 | Feb - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Digital Image and Video Forgery Detection: A Comprehensive Review of Techniques, Datasets, and Future Directions

Dr. Pratibha V. Kashid

Associate Professor

Department of Information Technology

Sir Visvesvaraya Institute Of Technology Nashik

Ms.Gudulkar Gayatri Piraji[1]

Department of Information Technology

Sir Visvesvaraya Institute Of Technology Nashik

Ms. Avhad Anisha Balu[3],

Department of Information Technology

Sir Visvesvaraya Institute Of Technology Nashik

Ms.Katale Bhagyashree Shashikant[2]

Department of Information Technology

Sir Visvesvaraya Institute Of Technology Nashik

Ms.Kurhade SakshiBhivaji[4]

Department of Information Technology

Sir Visvesvaraya Institute Of Technology Nashik

ABSTRACT

The rapid advancement of digital media editing tools and generative artificial intelligence has significantly increased the prevalence of manipulated images and videos. From traditional copy–move forgeries to sophisticated deepfake synthesis, digital content manipulation threatens information integrity, cybersecurity, journalism, and legal systems. This review provides a comprehensive analysis of digital image and video forgery detection techniques, benchmark datasets, evaluation metrics, and emerging research challenges. We categorize detection methods into classical forensic approaches and deep learning-based frameworks, examining their performance across detection accuracy, robustness, computational efficiency, interpretability, and multimodal integration. We further analyze major public datasets, highlight domain generalization challenges, and discuss adversarial vulnerabilities. Finally, we outline future research directions including cross-domain learning, physics-informed modeling, lightweight edge architectures, explainable AI, and self-supervised multimodal learning. This review serves as a unified technical reference for researchers, practitioners, and policymakers working in digital media forensics.

Keywords

Digital Forensics, Image Forgery Detection, Video Manipulation Detection, Deepfake Detection, CNN, Vision Transformers, Self-Supervised Learning

1. INTRODUCTION

Digital images and videos serve as primary evidence in journalism, surveillance, social media, and courtrooms. However, modern manipulation techniques—especially deep generative models such as GANs and diffusion-based synthesis—have made forgery creation increasingly realistic and accessible.

Forgery types include:

1. Copy–Move Manipulation
2. Image Splicing
3. Image Retouching
4. Deepfake Face Swapping
5. Expression Reenactment

6. Synthetic Video Generation

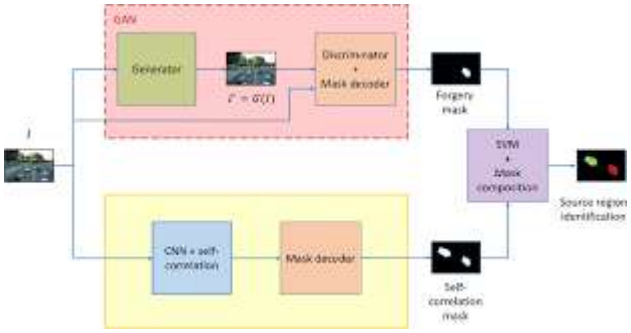

Figure 1.System Architecture

The arms race between forgery creation and detection demands robust forensic systems capable of generalizing across unseen domains.

2. TAXONOMY OF FORGERY DETECTION TECHNIQUES

2.1 Traditional (Hand-Crafted Feature) Methods

Method Type	Feature Used	Strength	Limitation
Block-based copy–move	DCT, PCA	Good for duplication detection	Fails under heavy post-processing
Keypoint-based	SIFT, SURF	Rotation/scale invariant	Computationally expensive
JPEG artifact analysis	Quantization inconsistencies	Effective in compression-based forgeries	Limited for RAW images
Noise pattern analysis	PRNU (sensor noise)	Device-level authentication	Sensitive to compression

2.2 Deep Learning-Based Approaches

Architecture	Application	Advantages	Limitations
CNN	Spatial forgery detection	High accuracy	Domain overfitting
RNN/LSTM	Temporal video analysis	Motion consistency modeling	Training complexity
Vision Transformer (ViT)	Global dependency modeling	Better generalization	Large data requirement
GAN-discriminator reuse	Deepfake detection	Good synthetic artifact capture	Vulnerable to adversarial attacks
Multimodal (Audio+Video)	Talking head detection	Higher reliability	Dataset scarcity

3. RELATED WORK

Digital image and video forgery detection has evolved significantly over the past two decades. Research in this field can be broadly categorized into three major phases: (1) traditional forensic approaches, (2) deep learning-based detection, and (3) multimodal and robust detection frameworks.

3.1 Traditional Image Forgery Detection

Early research focused on detecting statistical inconsistencies introduced during image manipulation. These approaches relied heavily on handcrafted features.

Copy–Move Forgery Detection

Copy–move forgery involves duplicating a region within the same image to conceal or replicate objects. Early block-based methods used Discrete Cosine Transform (DCT) and Principal Component Analysis (PCA) for matching similar regions. Keypoint-based methods using SIFT and SURF improved robustness to scaling and rotation.

However, these methods struggle under heavy post-processing operations such as compression, noise addition, or geometric transformations.

Image Splicing Detection

Splicing involves merging content from multiple images. Techniques based on color filter array (CFA) artifacts, illumination inconsistency, and JPEG compression signatures were proposed. Sensor Pattern Noise (PRNU) analysis was also introduced to verify camera source authenticity.

While effective in controlled settings, these approaches require high-resolution images and often fail against modern GAN-based synthetic images.

3.2 Deep Learning-Based Detection

The introduction of convolutional neural networks (CNNs) significantly improved forgery detection accuracy.

CNN-Based Spatial Detectors

CNN models automatically learn hierarchical spatial features. Architectures such as ResNet, Xception, and EfficientNet achieved high performance on benchmark datasets like FaceForensics++ and Celeb-DF.

However, studies reported a 15–20% performance drop when models were evaluated on unseen datasets, highlighting generalization limitations.

Temporal Modeling for Video Forgery

For video-based deepfake detection, temporal inconsistencies such as unnatural blinking, lip-sync mismatches, and motion artifacts were exploited using RNNs and LSTMs. These methods analyze frame-to-frame coherence to detect manipulation.

Though effective, temporal models are computationally expensive and sensitive to video compression.

Transformer-Based Approaches

Vision Transformers (ViT) have recently gained popularity due to their ability to model long-range dependencies. These models show improved robustness compared to CNNs but require large-scale training data and high computational resources.

3.3 GAN-Based Forgery and Detection Arms Race

Generative Adversarial Networks (GANs) revolutionized media synthesis. Models such as StyleGAN and diffusion-based architectures generate highly realistic fake faces and videos.

To counter this, researchers leveraged GAN discriminator features for detection. However, adversarial training strategies allow forgeries to bypass detectors, creating an ongoing arms race between generation and detection models.

3.4 Multimodal Detection Approaches

Recent studies explore multimodal detection by combining visual, audio, and physiological signals.

Examples include:

- Audio-visual synchronization analysis
- Heart rate signal estimation from facial videos
- Speech-lip movement consistency

Multimodal systems demonstrate improved robustness but suffer from dataset scarcity and synchronization challenges.

3.5 Explainable and Robust Detection

With increasing legal and forensic use, explainability has become critical. Techniques such as Grad-CAM and attention visualization provide interpretability by highlighting manipulated regions.

Robust detection approaches incorporate:

- Adversarial training
- Domain adaptation
- Physics-based modeling (lighting, shadows, reflections)

Despite progress, fully robust, explainable, and lightweight systems remain an open research challenge.

3.6 Summary of Research Trends

Research Phase	Key Techniques	Strength	Limitation
Traditional Methods	DCT, SIFT, PRNU	Interpretable	Weak against GANs
CNN-based	Xception, ResNet	High benchmark accuracy	Poor cross-dataset generalization
Temporal Models	LSTM, 3D CNN	Motion analysis	High computational cost
Transformer Models	ViT, Swin	Global feature modeling	Data-intensive
Multimodal	Audio-Visual fusion	Improved robustness	Dataset limitations

4. BENCHMARK DATASETS

4.1 Image Forgery Datasets

Dataset	Type	Samples	Strength	Limitation
CASIA v2	Splicing/Copy–Move	12,614	Widely used	Limited resolution
CoMoFoD	Copy–Move	260	Post-processing variants	Small size
Columbia	Splicing	363	High quality	Very small dataset

4.2 Deepfake / Video Datasets

Dataset	Year	Size	Type	Notes
FaceForensics++	2019	1,000+ videos	Manipulated faces	Multiple compression levels
Celeb-DF	2020	5,600+	Realistic deepfakes	High quality
DFDC	2020	100,000+	Diverse deepfakes	Industry-scale

5. EVALUATION METRICS

Metric	Formula	Purpose
Accuracy	(TP+TN)/(Total)	Overall correctness
Precision	TP/(TP+FP)	False alarm control
Recall	TP/(TP+FN)	Detection completeness
F1-Score	2PR/(P+R)	Balanced performance
AUC	ROC area	Threshold-independent evaluation

6. PERFORMANCE TRENDS

Table: Summary of Key Findings

Aspect	Current State	Challenges	Future Direction
Detection Accuracy	>95% on benchmarks	15–20% generalization drop	Cross-domain learning
Robustness	Good on known attacks	Adversarial vulnerability	Physics-based features
Efficiency	100+ GFLOPs	Mobile deployment	Lightweight models
Interpretability	Limited	Legal requirements	Explainable AI
Multimodal	Emerging	Dataset scarcity	Self-supervised learning

7. GAN and Diffusion Model Forgery

The emergence of generative models has fundamentally transformed digital media synthesis. Advanced generative architectures can now produce hyper-realistic images and videos that are visually indistinguishable from authentic media. Among these, Generative Adversarial Networks (GANs) and diffusion models represent the two dominant paradigms in synthetic media generation. While these technologies have enabled creative and commercial applications, they have simultaneously intensified challenges in digital forgery detection.

7.1 GAN-Based Face Generation

Generative Adversarial Networks (GANs), introduced by Ian Goodfellow in 2014, consist of two neural networks:

Generator – produces synthetic sample

Discriminator – distinguishes real from fake samples.

These networks compete in a minimax game, resulting in progressively realistic outputs.

Applications in Face Generation

GANs have been extensively used for:

- Face swapping
- Identity morphing
- Expression transfer
- Attribute manipulation (age, gender, emotion)

Popular GAN architectures include:

- NVIDIA's **StyleGAN** series
- ProGAN
- CycleGAN

Forensic Artifacts in GAN Images

- Although highly realistic, GAN-generated faces often exhibit
- Frequency spectrum inconsistencies
- Checkerboard artifacts
- Abnormal color distributions
- Irregular eye reflections
- Inconsistent background texture
- Detection methods exploit these artifacts using frequency-domain analysis and CNN-based classifiers.

However, newer GANs significantly reduce such detectable traces, making detection increasingly challenging.

7.2 StyleGAN-Based Manipulation

The **StyleGAN2** and **StyleGAN3** architectures improved realism through:

- Style-based modulation
- Adaptive instance normalization
- Improved alias-free synthesis

Key Characteristics

- High-resolution face synthesis (1024×1024 and above)
- Disentangled latent space manipulation
- Fine-grained control over facial attributes

Forensic Implications

StyleGAN-generated images show:

- Weak or missing camera sensor noise (PRNU)
- Synthetic frequency fingerprints
- Unrealistic corneal specular highlight

Recent detection research focuses on:

- GAN fingerprint analysis
- Spectral residual learning
- Patch-based anomaly detection

Despite progress, StyleGAN3 significantly reduces aliasing artifacts, weakening frequency-based detection strategies.

7.3 Diffusion Model-Based Synthesis

Diffusion models represent a newer class of generative models that outperform GANs in image fidelity.

Prominent diffusion-based systems include:

- Stable Diffusion
- DALL·E
- Midjourney

Working Principle

Diffusion models operate in two stages:

- **Forward process:** Gradually add noise to real images
- **Reverse process:** Learn to reconstruct images from noise

Unlike GANs, diffusion models optimize likelihood-based objectives, resulting in

- Higher diversity
- Fewer visual artifact
- More stable training

Forensic Challenges

Diffusion-generated images:

- Do not show classical GAN frequency artifact
- Exhibit natural noise distributions
- Better mimic real-world lighting

Detection approaches now rely on:

- Noise residual analysis
- Latent space trace detection
- Patch-level statistical modeling
- Synthetic texture anomaly detection

Because diffusion models better simulate physical image formation processes, traditional forensic cues become less reliable.

7.4 Adversarial Examples in Deepfake Detection

Adversarial attacks pose a major threat to deepfake detection systems.

An adversarial example is created by adding small perturbations to an image such that:

- Human perception remains unchanged
- Deep learning classifiers misclassify the input

Attack Types

- FGSM (Fast Gradient Sign Method)
- PGD (Projected Gradient Descent)
- Black-box adversarial attacks
- Physical-world attacks (printed perturbations)

Impact on Forgery Detection

Deepfake detectors can experience:

- Significant accuracy degradation
- False negatives under minimal perturbation
- Model collapse under adaptive attacks

To counter adversarial threats, researchers propose

- Adversarial training
- Ensemble detection model
- Certified robustness techniques
- Frequency-based detection
- Physics-informed modeling

Comparative Summary

Model Type	Realism	Artifact Visibility	Detection Difficulty	Computational Cost
Early GANs	Moderate	High	Low	Moderate
StyleGAN 2	High	Medium	Moderate	High
StyleGAN 3	Very High	Low	High	High
Diffusion Models	Extremely High	Very Low	Very High	Very High

8. MAJOR CHALLENGES

8.1 Cross-Domain Generalization

Models trained on one dataset drop 15–20% accuracy when tested on unseen datasets.

8.2 Adversarial Attacks

Small perturbations can bypass CNN detectors.

8.3 Dataset Bias

Most datasets focus only on face manipulation.

8.4 Edge Deployment

High GFLOPs models unsuitable for mobile surveillance systems.

8.5 Legal & Ethical Compliance

Need explainable decisions in court evidence systems.

9. EMERGING RESEARCH DIRECTIONS

9.1 Cross-Domain Learning

Domain adaptation and meta-learning approaches.

9.2 Physics-Based Features

Camera sensor modeling, light reflection consistency.

9.3 Lightweight Architectures

MobileNet, EfficientNet, pruning, quantization.

9.4 Explainable AI

Grad-CAM visualization, attention heatmaps.

9.5 Self-Supervised Multimodal Learning

Contrastive learning using audio-visual synchronization.

10. FIGURES

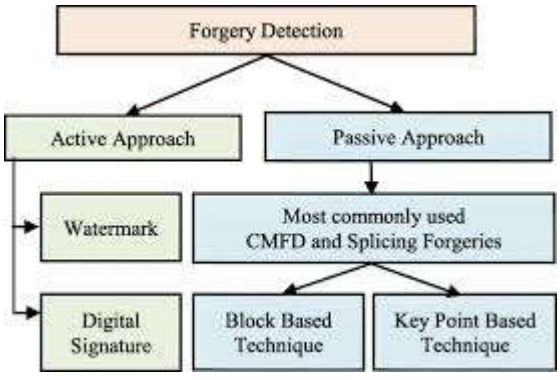

Figure 2: Taxonomy of Forgery Detection

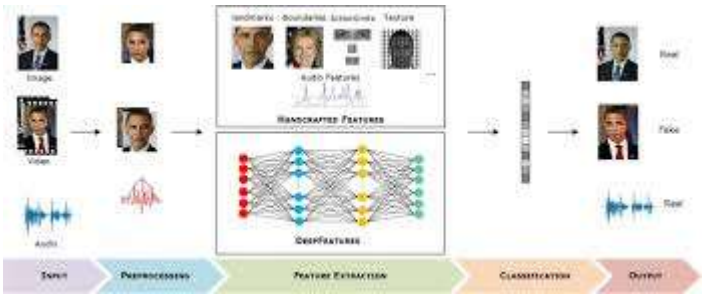

Figure 3: Deepfake Detection Pipeline

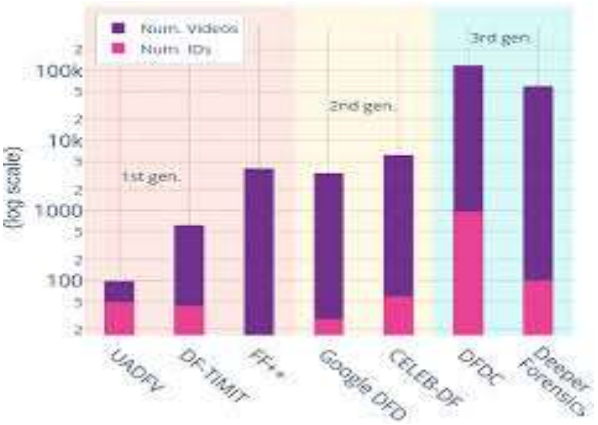

Figure 4: Dataset Distribution Graph

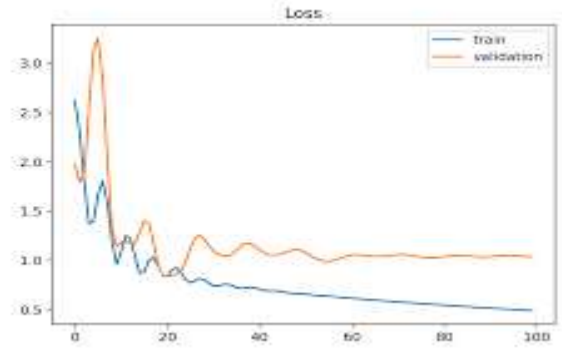

Figure 5: Cross-Domain Performance Drop

11. RESEARCH GAP ANALYSIS

Gap	Explanation	Research Opportunity
Real-world robustness	Models fail outside lab datasets	Domain adaptation
Explainability	Black-box predictions	XAI integration
Lightweight deployment	High computation	Edge AI optimization

Gap	Explanation	Research Opportunity
Multimodal scarcity	Few labeled datasets	Self-supervised learning

12. CONCLUSION

Digital forgery detection has progressed from handcrafted statistical methods to sophisticated deep learning systems achieving >95% benchmark accuracy. However, real-world deployment reveals significant challenges including cross-domain degradation, adversarial vulnerability, computational cost, and interpretability concerns. Future systems must integrate physics-aware modeling, multimodal self-supervised learning, lightweight optimization, and explainable AI frameworks to build robust and legally compliant forensic systems.

ACKNOWLEDGMENT

We thank Dr. Pratibha V. Kashid for her mentorship and SVIT Nashik for computational resources. We also acknowledge local authorities for their collaboration.

REFERENCES

1. Liang, E., Zhang, K., Hua, Z., & Jia, X. (2025). Frequency-driven deep learning network for image splicing forgery detection. Knowledge-Based Systems, 330(Part A), 114365.

2. Kadha, V., Bakshi, S., & Das, S. K. (2025). Unravelling Digital Forgeries: A Systematic Survey on Image Manipulation Detection and Localization. ACM Computing Surveys.

3. Chen, Y., Mi, C., Wei, J., Zhu, Y., & Zhou, J. (2025). M2-Net: multi-view learning multi-scale fusion network for image tampering detection. Knowledge-Based Systems, 327, 114122.

4. Nguyen-Le, H-H., Tran, V-T., Nguyen, D-T., & Le-Khac, N-A. (2025). Deepfake Detection Across Image, Video, and Audio: A Comprehensive Survey with Empirical Evaluation of Generalization and Robustness. eLife.

5. Taneja, N., Mishra, G. S., & Bhardwaj, D. (2025). Unmasking anti-forensic techniques: A DCNN-driven approach to uncover contrast enhancement and median filtering detection. Journal of Forensic Sciences, 70(6), 2324-2337.

6. A review of deep learning based multimodal forgery detection for video and audio. (2025). Discover Applied Sciences, 7, 987.

7. Primbetov, A. et al. (2025). HUMAN PERCEPTION VERSUS ARTIFICIAL INTELLIGENCE IN DEEPFAKE VIDEO DETECTION: AN EMPIRICAL STUDY. Zenodo.

8. An Overview of Passive Digital Image Tampering Detection Methods. (2025). IEEE Xplore.

9. A systematic literature review of video forgery detection techniques. (2025). Multimedia Tools and Applications, 84, 41277–41327.

10. Agarwal, S., Sharma, D., Girdhar, N., Kim, C., & Jung, K. H. (2025). A Survey of Image Forensics: Exploring Forgery Detection in Image Colorization. Computers, Materials and Continua, 84(3), 4195-4221.