International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

Deep Generative And Discriminative Modeling For Image Forgery Detection And Localization

G. Mary Pushpa
Dept. of Electronics and Communication Engineering
BESTIU,Anantapur,AP, India
marypushpa.50@gmail.com

K. Sravan Abhilash
Dept. of Electronics and Communication Engineering
Malla Reddy University, Hyderabad, India
sravan.abhi789@gmail.com

***Abstract*—Digital images are widely used in domains such as journalism, healthcare, law enforcement, and social media, where authenticity is crucial [1], [2]. However, the rapid evolution of image editing tools and AI-driven generative models has enabled highly realistic manipulations, including copy–move, splicing, and deepfake generation [3], [10]–[12]. These sophisticated forgeries are increasingly difficult to detect, making traditional image forensic methods inadequate for modern scenarios [1], [2], [11]. Despite advancements in deep learning-based forensics, significant challenges remain [6]–[8]. Existing approaches typically rely on either generative models [3]–[5], which reconstruct image distributions but lack strong classification and localization capabilities, or discriminative models [6]–[8], [13], [14], which perform well in classification but struggle with precise localization. Furthermore, current datasets are limited in diversity and often fail to represent real-world manipulations, affecting model generalization [10], [17], [18].**

This research proposes a unified deep learning framework that integrates generative and discriminative models for robust image forgery detection and localization. A GAN-based generative module performs reconstruction-driven anomaly detection by learning authentic image patterns and producing residual maps that highlight manipulated regions [3]–[5]. In parallel, a Vision Transformer (ViT)-based discriminative module utilizes self-attention mechanisms to capture global dependencies, enabling accurate classification and pixel-level localization through attention heatmaps [13], [14], [19], [20]. By combining reconstruction-based anomaly detection with attention-driven analysis, the framework enables simultaneous detection and localization while providing explainable outputs [8], [9], [16].

This work contributes to advanced digital image forensics by delivering a scalable, robust, and interpretable solution for detecting AI-generated manipulations, with applications in media verification, cybersecurity, and law enforcement [11], [12], [16], [20].

***Index Terms*—Image Forgery Detection, Generative Adversarial Networks (GAN), Vision Transformers (ViT), Deepfake Detection, Image Localization, Residual Mapping, Attention Mechanisms, Explainable AI, Multi-forgery Dataset, Digital Image Forensics**

I. Introduction

The rapid growth of digital imaging technologies, coupled with the exponential rise of social media platforms such as Instagram, Facebook, and X (Twitter), has significantly increased the creation, sharing, and dissemination of multimedia content. Images are now a dominant medium for communication, influencing public perception, journalism, legal evidence, and decision-making processes. However, alongside these advancements, sophisticated image editing tools like Adobe Photoshop and GIMP have made it increasingly easy to manipulate images with high realism. This has raised serious concerns regarding the authenticity and integrity of digital visual content.

Image forgery refers to the intentional alteration of an image to misrepresent its original content. Such manipulations can be used for malicious purposes, including spreading misinformation, creating fake evidence, and manipulating public opinion. In critical domains such as journalism, healthcare, law enforcement, and national security, the consequences of forged images can be severe and far-reaching. Therefore, developing reliable and automated techniques for detecting image forgery has become an essential task in the field of digital forensics.

Image forgeries can be broadly categorized into several types based on the manipulation technique used. The most common types include copy-move forgery, splicing, and retouching. In copy-move forgery, a region of an image is duplicated and pasted within the same image to conceal or replicate objects. Splicing involves combining regions from multiple images to create a composite image, often with the intention of deceiving the viewer. Retouching includes subtle modifications such as smoothing, enhancement, or removal of details to alter the perception of the image. The unprecedented





(a) Original Image (b) Forged Image

Fig. 1: Illustration of one of the forgery technique (copy-move) where an object is duplicated within the same image.

growth of digital technologies has transformed the way visual information is created, stored, transmitted, and consumed. Digital images have become one of the most influential forms of communication in modern society, serving as essen- tial sources of information across diverse domains including

journalism, healthcare, law enforcement, scientific research, military intelligence, insurance investigations, e-commerce, and social media platforms [1], [2]. The widespread availability of smartphones equipped with high-resolution cameras, combined with the rapid expansion of cloud storage and social networking services, has resulted in the generation and sharing of billions of images every day. Consequently, digital images are frequently perceived as reliable representations of reality and are often used as supporting evidence in critical decision-making processes.

The integrity and authenticity of digital images are therefore of paramount importance. In many situations, images serve as legal evidence, medical documentation, journalistic proof, or forensic records. Any unauthorized modification to such images can lead to misinformation, false accusations, fraudulent activities, reputational damage, and significant societal consequences. For example, manipulated images distributed through social media platforms can influence public opinion, affect political outcomes, and contribute to the spread of misinformation campaigns. Similarly, forged medical images may lead to incorrect diagnoses, while tampered surveillance images can compromise criminal investigations and judicial proceedings. These concerns have elevated image authenticity verification into a major research area within the broader field of digital image forensics [1], [11].

Historically, image manipulation required specialized expertise and sophisticated software tools, limiting its accessibility to a relatively small group of users. However, recent advances in image editing technologies have dramatically changed this landscape. Modern software applications enable users to perform complex manipulations through intuitive interfaces and automated editing functionalities. Common image forgery techniques include copy–move forgery, image splicing, image retouching, object insertion, object removal, and content replacement. In copy–move forgery, a region of an image is duplicated and pasted elsewhere within the same image to conceal or replicate objects. Image splicing involves combining content from multiple images into a single composite image. Retouching operations modify image appearance through enhancement, smoothing, or alteration of visual details. These manipulations can be executed with remarkable precision, making them increasingly difficult to detect through visual inspection alone [2], [6].

The emergence of artificial intelligence has further accelerated the evolution of image manipulation techniques. In particular, deep learning-based generative models have demonstrated extraordinary capabilities in synthesizing highly realistic visual content. Generative Adversarial Networks (GANs) introduced a revolutionary paradigm in image generation by employing an adversarial learning process between a generator and a discriminator [3]. Subsequent developments such as Deep Convolutional GANs, StyleGAN, and related architectures have significantly improved image synthesis quality, enabling the generation of photorealistic faces, objects, and scenes [4]. These advancements have facilitated the creation of deepfakes and other forms of synthetic media that are often indistinguish-

able from authentic content. Deepfake technology, originally developed for beneficial applications such as entertainment, virtual reality, and digital content creation, has also been exploited for malicious purposes including identity fraud, mis- information dissemination, impersonation attacks, and social engineering campaigns [10]–[12].

The increasing sophistication of AI-generated content presents substantial challenges for traditional image foren- sic methodologies. Conventional forgery detection approaches rely primarily on handcrafted features and statistical incon- sistencies introduced during the manipulation process. Tech- niques based on sensor pattern noise analysis, compression artifacts, color inconsistencies, frequency-domain characteris- tics, and local feature descriptors have demonstrated effec- tiveness in identifying certain classes of image manipulations [1], [2]. Nevertheless, these methods often assume the pres- ence of detectable artifacts that may not exist in modern AI-generated images. Furthermore, advanced post-processing operations such as compression, scaling, filtering, rotation, blurring, and noise addition can obscure forensic traces, re-ducing the effectiveness of traditional detection techniques. As a result, many existing forensic methods exhibit limited robustness when confronted with sophisticated manipulations encountered in real-world scenarios.

To overcome these limitations, researchers have increasingly adopted deep learning approaches for image forgery detection. Convolutional Neural Networks (CNNs) have emerged as powerful tools for automatically learning hierarchical fea- ture representations from image data. Unlike handcrafted approaches, CNNs can identify complex visual patterns and subtle inconsistencies without requiring manually designed feature extraction procedures [7], [8]. Numerous CNN-based forensic frameworks have demonstrated superior performance in detecting image manipulations across various datasets and forgery types. However, despite their success, CNN-based approaches possess certain limitations. Their receptive fields are inherently local, making it difficult to capture long-range dependencies and global contextual relationships within an image. Moreover, CNN models often require large volumes of annotated training data and may struggle to generalize effectively across unseen manipulation techniques.

More recently, transformer-based architectures have gained considerable attention within the computer vision community. Vision Transformers (ViTs) extend the success of transformers from natural language processing to image analysis tasks by employing self-attention mechanisms that model relationships among image patches [13]. Unlike CNNs, Vision Transformers can capture both local and global contextual information, enabling them to identify subtle semantic inconsistencies that may indicate image manipulation. Furthermore, attention mechanisms provide a degree of interpretability by high- lighting image regions that contribute most significantly to classification decisions. Subsequent developments, including hierarchical transformer architectures such as Swin Transform- ers, have further improved representation learning capabilities for visual tasks [14]. These characteristics make transformer-

based approaches highly attractive for advanced image forensic applications.

Parallel to discriminative learning approaches, generative models have also been investigated for anomaly detection and forensic analysis. Generative frameworks learn the underlying distribution of authentic images and attempt to reconstruct them. When a manipulated image is presented, discrepancies between the original image and its reconstruction can reveal suspicious regions indicative of tampering [5]. This reconstruction-based anomaly detection paradigm offers an effective mechanism for identifying previously unseen manipulation types and reducing dependency on extensive labeled datasets. However, generative models often lack strong classification capabilities and may produce coarse localization results. Consequently, relying solely on generative modeling can limit overall forensic performance.

Despite substantial progress in both generative and discriminative learning paradigms, significant research challenges remain unresolved. Existing approaches frequently emphasize either anomaly reconstruction or classification performance while neglecting the complementary strengths offered by the alternative paradigm. Furthermore, accurate localization of manipulated regions remains a difficult task, particularly when dealing with small, subtle, or semantically complex forgeries. Another challenge arises from the limited diversity of publicly available datasets, which often fail to adequately represent the wide variety of manipulations encountered in real-world environments. These limitations can negatively impact model generalization and reduce effectiveness when deployed in practical forensic applications [10], [17], [18].

The growing prevalence of AI-generated manipulations highlights the urgent need for more comprehensive forensic frameworks capable of simultaneously detecting, localizing, and explaining image manipulations. Such systems must exhibit robustness against diverse forgery techniques, adaptability to emerging synthetic media technologies, and transparency in decision-making processes. Motivated by these requirements, this research explores the integration of generative and discriminative deep learning paradigms within a unified forensic framework. By combining GAN-based reconstruction-driven anomaly detection with Vision Transformer-based attention-guided classification and localization, the proposed approach seeks to leverage the complementary strengths of both methodologies. The ultimate objective is to develop a scalable, robust, and interpretable solution for image forgery detection and localization that can effectively address the challenges posed by modern AI-driven image manipulation technologies.

II. Evolution of Image Forgery Detection

The field of image forgery detection has evolved significantly over the past two decades in response to the increasing sophistication of image manipulation techniques. Early forensic methods relied primarily on handcrafted features and statistical analysis, while recent advances have introduced machine learning and deep learning approaches capable of automatically learning complex representations from image data. Despite these developments, challenges related to generalization, localization accuracy, interpretability, and robustness continue to motivate ongoing research. Understanding the evolution of image forgery detection provides valuable insights into the strengths and limitations of existing methodologies and highlights the need for more comprehensive forensic frameworks.

A. Traditional Approaches

Traditional image forgery detection methods are primarily based on the observation that image manipulations often introduce statistical inconsistencies and artifacts that differ from those present in authentic images. These approaches typically rely on handcrafted features extracted from either the spatial domain or the frequency domain of an image [1], [2].

One of the earliest and most extensively studied forgery types is copy–move forgery, where a region of an image is copied and pasted elsewhere within the same image. To detect such manipulations, researchers developed block-based matching techniques that divide an image into overlapping blocks and compare their similarity using various feature representations. Methods based on Discrete Cosine Transform (DCT), Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and wavelet transforms became popular due to their ability to capture distinctive image characteristics while reducing computational complexity [1].

Another important category of traditional approaches is keypoint-based forgery detection. Algorithms such as Scale-Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF) identify salient image points and generate descriptors that are invariant to scaling, rotation, and moderate illumination changes. These descriptors can be matched across different image regions to identify duplicated or manipulated content [6]. Keypoint-based approaches generally offer improved robustness against geometric transformations compared to block-based methods and have been widely applied in copy–move forgery detection.

Traditional forensic methods have also exploited sensor-related artifacts, including sensor pattern noise, color filter array inconsistencies, chromatic aberrations, and JPEG compression signatures. These techniques analyze intrinsic image characteristics generated during image acquisition and storage to identify anomalies indicative of tampering [2]. While such methods have demonstrated effectiveness in controlled environments, their performance often deteriorates when images undergo extensive post-processing operations.

Although traditional approaches are computationally efficient and relatively interpretable, they suffer from several limitations. Their reliance on manually designed features restricts their ability to adapt to diverse manipulation types. Furthermore, modern image editing techniques can effectively conceal many of the artifacts targeted by traditional forensic methods, reducing detection reliability in real-world scenarios.

B. Machine Learning Approaches

To overcome the limitations of purely handcrafted techniques, researchers began integrating machine learning algorithms into image forgery detection systems. In these approaches, handcrafted features are first extracted from images and subsequently used as inputs to classification models capable of distinguishing authentic images from manipulated ones. Common machine learning classifiers employed in image forensics include Support Vector Machines (SVMs), Random Forests, Decision Trees, k-Nearest Neighbors (k-NN), and Artificial Neural Networks. These algorithms learn decision boundaries from labeled training data and often achieve superior classification performance compared to rule-based forensic techniques [7].

Machine learning approaches enabled the fusion of multiple feature types, allowing forensic systems to exploit complementary information from texture descriptors, statistical features, frequency-domain representations, and keypoint-based features. Such hybrid feature representations improved robustness against different manipulation strategies and enhanced overall detection accuracy.

Another important contribution of machine learning methods was their ability to generalize beyond specific forgery artifacts. Rather than relying solely on predefined statistical inconsistencies, classifiers could learn complex relationships among extracted features, enabling the detection of a broader range of manipulations. As larger datasets became available, machine learning-based forensic systems demonstrated promising results across multiple forgery detection benchmarks.

Despite these improvements, machine learning approaches still depend heavily on handcrafted feature engineering. The effectiveness of the classifier is largely determined by the quality and representativeness of the extracted features. Consequently, these methods often struggle when confronted with previously unseen manipulations or highly sophisticated forgeries that do not exhibit the characteristics captured by manually designed features.

C. Deep Learning Approaches

The emergence of deep learning has transformed image forgery detection by enabling automatic feature learning directly from image data. Unlike traditional machine learning approaches, deep neural networks eliminate the need for manual feature engineering and instead learn hierarchical representations through multiple layers of nonlinear processing [7], [8].

Convolutional Neural Networks (CNNs) represent the most widely adopted deep learning architecture in image forensics. CNNs are capable of learning low-level features such as edges and textures, intermediate representations capturing local structures, and high-level semantic features that distinguish authentic and manipulated content. Numerous CNN-based forensic frameworks have demonstrated remarkable success in detecting copy–move forgeries, image splicing, retouching operations, and AI-generated manipulations [6]–[8].

Subsequent developments introduced fully convolutional networks, encoder–decoder architectures, and segmentation-based models capable of performing pixel-level localization of tampered regions. These architectures significantly improved the ability of forensic systems to not only detect manipulations but also identify their precise spatial locations within an image [9], [16].

Recent years have witnessed the growing adoption of transformer-based architectures in image forensics. Vision Transformers (ViTs) employ self-attention mechanisms to model relationships among image patches and capture long-range dependencies that may be overlooked by conventional convolutional networks [13]. By incorporating global contextual information, transformer-based models have demonstrated strong performance in detecting subtle and semantically complex manipulations. Hierarchical transformer architectures such as Swin Transformers have further enhanced feature representation and localization capabilities [14].

Generative models have also emerged as important tools in modern image forensics. Generative Adversarial Networks (GANs) can learn the distribution of authentic images and identify anomalies through reconstruction-based analysis. Such approaches are particularly attractive because they offer the potential to detect previously unseen manipulation types without requiring extensive labeled datasets [3]–[5]. Moreover, generative models have been used to improve forensic robustness through synthetic data augmentation and adversarial training strategies.

The integration of discriminative and generative learning paradigms has become an increasingly active area of research, as it offers opportunities to combine classification accuracy with anomaly detection and localization capabilities.

D. Current Limitations

Despite substantial advances in image forgery detection, several significant challenges remain unresolved. One of the primary limitations of existing approaches is their difficulty in generalizing across diverse manipulation techniques and datasets. Many forensic models perform well on specific benchmark datasets but experience significant performance degradation when evaluated on images originating from different sources or manipulated using previously unseen methods [10], [17], [18].

Another challenge involves the detection of highly realistic AI-generated content. Modern generative models can synthesize images with remarkable visual quality, reducing the presence of detectable artifacts and making discrimination increasingly difficult [11], [12]. As generative technologies continue to evolve, forensic systems must continuously adapt to emerging manipulation techniques.

Accurate localization of manipulated regions also remains a complex problem. While many models achieve high image-level classification accuracy, they often struggle to precisely identify the boundaries of forged regions, particularly when manipulations are small, semantically consistent, or blended seamlessly into the surrounding image content [9], [16].

Interpretability represents another important concern. Many deep learning-based forensic systems operate as black-box models, making it difficult to understand the reasoning behind their decisions. This limitation is particularly problematic in legal, forensic, and security-sensitive applications where transparent and explainable evidence is essential.

Finally, existing datasets frequently fail to capture the diversity and complexity of real-world manipulations. Many benchmark datasets contain synthetic or simplified forgeries that do not fully represent contemporary editing workflows and AI-generated content. Consequently, models trained on these datasets may exhibit limited robustness when deployed in practical environments [10], [17], [18].

These limitations highlight the need for next-generation image forensic frameworks capable of integrating anomaly detection, classification, localization, and explainability within a unified architecture. Such requirements motivate the exploration of hybrid generative and discriminative approaches for robust image forgery detection and localization.

III. Theoretical Foundations

Modern image forgery detection systems increasingly rely on deep learning architectures capable of learning complex visual representations from large-scale datasets. Among these architectures, generative and discriminative models have emerged as two dominant paradigms for forensic analysis. While generative models focus on learning the underlying distribution of authentic images and identifying deviations from expected patterns, discriminative models directly learn decision boundaries that distinguish authentic images from manipulated ones. More recently, explainable artificial intelligence (XAI) techniques have been incorporated into forensic systems to enhance transparency and interpretability. The theoretical foundations of these approaches provide the basis for developing robust and reliable image forgery detection frameworks.

A. Generative Modeling

Generative modeling aims to learn the probability distribution of authentic images and generate samples that resemble the training data. Unlike discriminative approaches that focus solely on classification, generative models attempt to understand the intrinsic structure and characteristics of image data. This capability makes them particularly useful for anomaly detection and forensic analysis, where manipulated regions often deviate from the distribution of authentic content.

1) Generative Adversarial Networks (GANs): Generative Adversarial Networks (GANs), introduced by Goodfellow et al. [3], represent one of the most influential developments in generative modeling. A GAN consists of two neural networks: a generator and a discriminator. The generator attempts to produce realistic images from random noise, while the discriminator seeks to distinguish generated images from real images. Through this adversarial learning process, both networks continuously improve until the generated samples become highly realistic.

Mathematically, the GAN optimization objective can be expressed as:

$$\min_G \max_D V(D,G)$$

$$E^*x \sim p * data(x)[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

where G denotes the generator, D denotes the discriminator, x represents real image samples, and z is a random latent vector [3].

Subsequent GAN architectures, including Deep Convolutional GANs (DCGANs) and StyleGANs, have significantly improved image generation quality and representation learning capabilities [4]. In image forensics, GANs have been widely employed for anomaly detection, synthetic image generation, data augmentation, and forgery localization.

2) Reconstruction Learning: Reconstruction learning is based on the principle that a model trained exclusively on authentic images should be capable of reconstructing genuine content accurately while struggling to reconstruct manipulated regions. The reconstruction process can be implemented using GANs, autoencoders, variational autoencoders, or other generative frameworks.

Given an input image I, the reconstruction model generates a reconstructed image $\hat{I}$:

$$\hat{I} = G(I) \quad (2)$$

where $G(\cdot)$ represents the reconstruction network. The reconstruction error is then computed as:

$$R = |I - \hat{I}| \quad (3)$$

where R denotes the residual map.

Regions containing manipulations often produce larger reconstruction errors because they deviate from the authentic image distribution learned during training. Consequently, reconstruction learning provides an effective mechanism for highlighting suspicious regions within an image [5].

3) Anomaly Detection: Anomaly detection is a critical application of generative modeling in digital image forensics. Instead of explicitly learning every possible manipulation type, anomaly detection models learn the characteristics of authentic images and identify deviations as potential forgeries.

This approach offers several advantages. First, it reduces dependence on large annotated forgery datasets. Second, it enables detection of previously unseen manipulation techniques. Third, it provides pixel-level evidence through reconstruction discrepancies. GAN-based anomaly detection frameworks have demonstrated promising performance in identifying image manipulations that are difficult to detect using purely discriminative methods [5], [19].

Despite these strengths, generative approaches often exhibit limitations in classification accuracy and localization precision. These shortcomings motivate the integration of discriminative models capable of learning highly discriminative decision boundaries.

B. Discriminative Modeling

Discriminative models directly learn mappings between image representations and class labels. Rather than modeling the entire data distribution, they focus on maximizing classification performance by learning features that distinguish authentic images from manipulated ones.

1) Convolutional Neural Networks (CNNs): Convolutional Neural Networks (CNNs) have become a fundamental component of modern image analysis systems. CNNs employ convolutional filters to automatically learn hierarchical feature representations from image data [7], [8].

A convolution operation can be expressed as:

$$Y(i,j) = \sum_m \sum_n X(i-m, j-n) K(m,n) \quad (4)$$

where X represents the input image, K denotes the convolution kernel, and Y is the resulting feature map.

CNNs learn low-level features such as edges and textures in early layers, while deeper layers capture high-level semantic information. This hierarchical feature extraction capability enables CNNs to identify subtle artifacts and inconsistencies associated with image manipulation.

Several CNN-based architectures have been successfully applied to image forgery detection, including ResNet, DenseNet, EfficientNet, and Fully Convolutional Networks (FCNs) [16], [20]. These models have demonstrated strong performance in both image-level classification and pixel-level localization tasks.

However, CNNs inherently possess limited receptive fields and primarily focus on local spatial information. Consequently, they may struggle to capture long-range dependencies and global contextual relationships within complex scenes.

2) Vision Transformers: Vision Transformers (ViTs) represent a significant advancement in computer vision by adapting transformer architectures originally developed for natural language processing to image analysis tasks [13].

Instead of processing images through convolutional operations, ViTs divide an image into a sequence of fixed-size patches:

$$I = P_1, P_2, \ldots, P_N \quad (5)$$

where P_i denotes an image patch.

These patches are transformed into embeddings and processed using self-attention mechanisms. Unlike CNNs, Vision Transformers can directly model relationships among all image regions, enabling the capture of global contextual information.

The ability to understand long-range dependencies makes Vision Transformers particularly suitable for image forgery detection, where manipulated regions may exhibit subtle semantic inconsistencies relative to distant image content [14], [20].

3) Attention Mechanisms: Attention mechanisms constitute the core component of transformer architectures. They allow the model to dynamically focus on the most relevant image regions during feature extraction and decision making.

The self-attention operation is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (6)$$

where Q, K, and V represent query, key, and value matrices, respectively [13].

Through self-attention, each image patch interacts with every other patch, enabling the model to capture global semantic relationships. In image forensics, attention mechanisms help identify suspicious regions that contribute most significantly to forgery predictions. This property not only improves detection accuracy but also enhances interpretability through visual attention maps.

C. Explainable Artificial Intelligence

As deep learning models become increasingly complex, explainability has emerged as a critical requirement for forensic applications. In domains such as law enforcement, cybersecurity, and legal investigations, forensic conclusions must be supported by interpretable evidence rather than opaque predictions.

1) Residual Maps: Residual maps provide explanations by visualizing reconstruction errors generated by generative models. The residual image highlights discrepancies between the original image and its reconstructed version:

$$R(x,y) = |I(x,y) - \hat{I}(x,y)| \quad (7)$$

where (x, y) denotes pixel coordinates.

Regions exhibiting high residual values are more likely to contain manipulations. Consequently, residual maps serve as intuitive visual indicators of anomalous content and support forensic decision-making [5], [19].

2) Attention Maps: Attention maps are generated by transformer-based models and visualize the importance assigned to different image regions during classification. By examining attention distributions, investigators can identify which areas influenced the model's decision most strongly.

Unlike traditional saliency methods, attention maps are directly derived from the internal self-attention mechanisms of transformer architectures, providing a natural explanation of the decision-making process [13], [14], [20].

3) Forensic Interpretability: Forensic interpretability refers to the ability of a forensic system to provide understandable and verifiable evidence supporting its predictions. In practical applications, a detection result alone is often insufficient; investigators require explanations indicating why a particular image has been classified as manipulated.

The combination of residual maps from generative models and attention maps from discriminative models offers complementary forms of interpretability. Residual maps reveal anomaly locations, while attention maps illustrate the reasoning process of the classification model. Together, these outputs improve transparency, increase user trust, and facilitate

adoption of deep learning-based forensic systems in high-stakes environments [11], [12], [16].

Therefore, explainable artificial intelligence serves as a critical bridge between advanced deep learning models and practical forensic applications, enabling more reliable and accountable image forgery detection systems.

IV. Proposed Deep Generative and Discriminative Framework

The increasing sophistication of image manipulation techniques, particularly those enabled by artificial intelligence, has highlighted the limitations of existing forensic systems that rely exclusively on either generative or discriminative learning paradigms. Generative models are effective in identifying anomalies by learning the distribution of authentic images, whereas discriminative models excel at classification by learning decision boundaries between authentic and manipulated content. However, each approach exhibits inherent limitations when employed independently. Generative models often lack strong classification capabilities, while discriminative models may struggle to accurately localize manipulated regions and explain their predictions. To address these challenges, this research proposes a unified deep generative and discriminative framework that integrates the complementary strengths of both paradigms for robust image forgery detection and localization [5], [8], [13], [20].

“‘latex “‘

The proposed framework consists of four major components: a generative branch, a discriminative branch, a fusion strategy, and a localization mechanism. Together, these components enable the system to simultaneously identify, classify, and localize manipulated regions while providing interpretable evidence supporting its decisions.

A. Generative Branch

The generative branch is designed to learn the intrinsic distribution of authentic images and identify deviations that may indicate image manipulation. This branch is implemented using a Generative Adversarial Network (GAN)-based reconstruction framework, where the model is trained primarily on authentic images to capture natural image statistics and structural characteristics [3]–[5].

During training, the generative model learns to reconstruct authentic images with high fidelity. When a manipulated image is presented during inference, the reconstruction process attempts to generate a version of the image that conforms to the learned distribution of authentic content. Since forged regions often differ from genuine image structures, these regions tend to be reconstructed less accurately than authentic areas.

The discrepancy between the input image and its reconstructed counterpart produces a residual map that highlights suspicious regions. Areas exhibiting significant reconstruction errors are considered potential manipulation candidates. This reconstruction-driven anomaly detection process enables the framework to identify both known and previously unseen forgery types without relying exclusively on labeled examples [5], [19].

An important advantage of the generative branch is its ability to capture subtle inconsistencies that may not be directly observable through conventional classification approaches. By focusing on deviations from authentic image distributions, the generative component provides valuable forensic evidence regarding the presence and location of manipulated content.

B. Discriminative Branch

While the generative branch focuses on anomaly detection, the discriminative branch is responsible for learning highly discriminative representations capable of distinguishing authentic images from manipulated ones. This branch employs a Vision Transformer (ViT)-based architecture due to its ability to model both local and global image relationships through self-attention mechanisms [13], [14].

Unlike conventional Convolutional Neural Networks that primarily capture local spatial patterns, Vision Transformers process images as sequences of image patches and establish interactions among all image regions. This capability allows the model to analyze semantic consistency across the entire image and identify manipulations that may appear visually plausible at the local level but inconsistent within the broader scene context.

The discriminative branch learns forensic representations from both authentic and manipulated images. Through supervised training, the model develops the ability to recognize manipulation artifacts, semantic inconsistencies, texture irregularities, and contextual anomalies associated with various forgery techniques, including copy–move forgery, image splicing, object removal, and AI-generated manipulations [8], [11], [20].

Another significant advantage of the discriminative branch is its ability to generate attention weights that indicate which image regions contribute most strongly to the classification decision. These attention patterns provide useful information for subsequent localization and explainability processes.

C. Fusion Strategy

The core novelty of the proposed framework lies in the integration of generative and discriminative information through a unified fusion strategy. Rather than treating anomaly detection and classification as independent tasks, the proposed framework combines outputs from both branches to create a comprehensive forensic representation.

The generative branch provides residual-based anomaly information that highlights deviations from authentic image distributions, while the discriminative branch provides semantic and contextual information learned through supervised classification. These complementary representations capture different aspects of image manipulation and therefore offer greater forensic reliability when combined.

The fusion process integrates reconstruction residuals, discriminative feature embeddings, and attention information into

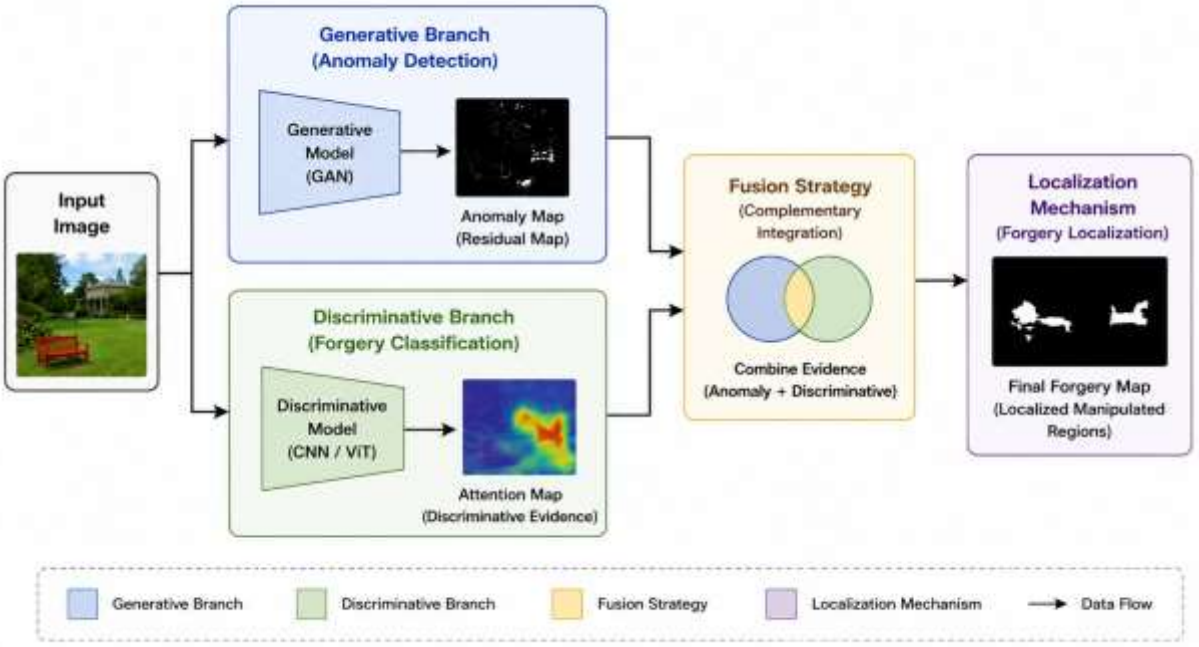


Fig. 2: Overview of the proposed deep generative and discriminative framework for image forgery detection and localization. The generative branch performs anomaly detection, while the discriminative branch performs forgery classification. Their complementary outputs are fused to generate the final localization map.

a unified decision-making framework. By combining anomaly-based evidence with classification-based evidence, the system can reduce false positives and improve robustness against diverse manipulation techniques [5], [8], [20].

From a forensic perspective, this hybrid strategy offers several advantages. First, it improves generalization by leveraging both distribution learning and discriminative learning. Second, it increases robustness against previously unseen manipulations. Third, it enhances detection reliability by requiring consistency between anomaly-based and classification-based evidence before generating final predictions.

Consequently, the fusion strategy serves as a bridge between generative and discriminative learning paradigms, enabling the framework to benefit from the strengths of both approaches while mitigating their individual weaknesses.

D. Localization Mechanism

Beyond image-level detection, forensic systems must accurately identify the spatial location of manipulated regions. Therefore, the proposed framework incorporates a dedicated localization mechanism that combines information from residual maps generated by the generative branch and attention maps produced by the discriminative branch.

The residual maps identify regions that deviate from authentic image distributions, while attention maps indicate regions that contribute significantly to the classification decision. By integrating these complementary sources of information, the localization mechanism produces refined manipulation maps that highlight suspicious image regions with greater precision [9], [16], [20].

The localization process is designed to operate at a fine-grained level, enabling the identification of small and subtle manipulations that may otherwise remain undetected. This capability is particularly important for modern image forgeries, where manipulated regions are often carefully blended into surrounding content to avoid visual detection.

Furthermore, the localization mechanism contributes to explainability by providing visual evidence supporting the forensic decision. Investigators can examine both anomaly maps and attention maps to understand why specific image regions were classified as manipulated. Such interpretability is essential in applications involving legal investigations, digital evidence verification, cybersecurity analysis, and media authentication [11], [12], [16].

Overall, the proposed localization mechanism transforms the framework from a simple image classification system into a comprehensive forensic analysis tool capable of simultaneously detecting, localizing, and explaining image manipulations.

E. Framework Significance

The proposed deep generative and discriminative framework represents a unified approach to image forgery detection and localization that combines reconstruction-based anomaly detection, attention-guided classification, and explainable localization within a single architecture. By integrating the complementary strengths of GAN-based generative modeling

and Vision Transformer-based discriminative learning, the framework addresses several limitations of existing forensic systems, including poor generalization, limited localization capability, and lack of interpretability.

The framework is designed to support a wide range of image manipulation scenarios, including copy–move forgery, image splicing, object insertion, object removal, image retouching, and AI-generated synthetic content. As image manipulation technologies continue to evolve, the integration of generative and discriminative learning paradigms provides a promising direction for developing scalable, robust, and explainable digital image forensic systems capable of addressing emerging challenges in media authentication and cybersecurity [11], [19], [20].

V. Conclusion

The rapid advancement of image editing technologies and artificial intelligence-based generative models has significantly increased the complexity and realism of digital image manipulations. Traditional image forensic techniques, which primarily rely on handcrafted features and statistical inconsistencies, often struggle to detect sophisticated forgeries such as image splicing, copy–move manipulation, and AI-generated synthetic content [1], [2], [11]. Although recent deep learning-based approaches have achieved substantial improvements in detection performance, existing methods typically focus on either generative or discriminative learning paradigms independently, limiting their effectiveness in terms of localization accuracy, generalization capability, and interpretability [5], [8], [13].

This paper presented a conceptual framework for deep generative and discriminative modeling for image forgery detection and localization. The proposed framework integrates GAN-based generative modeling with Vision Transformer-based discriminative learning to exploit the complementary strengths of both approaches. The generative branch focuses on reconstruction-driven anomaly detection by learning authentic image distributions and identifying suspicious regions through residual analysis, while the discriminative branch leverages self-attention mechanisms to capture global contextual relationships and perform robust forgery classification [3], [5], [13], [14]. Furthermore, the integration of residual maps and attention maps provides a unified mechanism for manipulation localization and enhances the explainability of forensic decisions [9], [16], [20].

By combining anomaly-based reasoning with attention-guided classification, the proposed framework addresses several key challenges in modern image forensics, including the detection of previously unseen manipulations, improved localization of forged regions, and enhanced forensic interpretability. The framework is designed to support diverse manipulation scenarios, including copy–move forgery, image splicing, image retouching, object insertion and removal, as well as emerging deepfake and AI-generated content [10]–[12]. In addition, the incorporation of explainable artificial intelligence principles strengthens the transparency and reliability of forensic analysis, making the framework more suitable for high-stakes applications such as law enforcement, cybersecurity, judicial investigations, and media verification.

As digital media generation technologies continue to evolve, the need for robust, scalable, and trustworthy forensic solutions will become increasingly important. The proposed deep generative and discriminative framework provides a promising direction for next-generation image forgery detection systems by integrating reconstruction learning, discriminative classification, localization, and explainability within a unified architecture. Future research may explore multimodal forensic analysis, foundation models for image authentication, self-supervised learning strategies, and real-time deployment frameworks to further enhance the effectiveness and practical applicability of digital image forensic systems [11], [14], [20].

References

[1] H. Farid, "Image Forgery Detection," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 16–25, 2009.

[2] J. Fridrich and M. Goljan, "Digital Image Forensics," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 26–37, 2009.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.

[4] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *International Conference on Learning Representations (ICLR)*, 2016.

[5] T. Schlegl, P. Seebock, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks," in *Information Processing in Medical Imaging*, 2017, pp. 146–157.

[6] Y. Rao and J. Ni, "A Deep Learning Approach to Detection of Splicing and Copy-Move Forgeries in Images," in *IEEE International Workshop on Information Forensics and Security*, 2016.

[7] B. Bayar and M. C. Stamm, "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer," in *ACM Workshop on Information Hiding and Multimedia Security*, 2016, pp. 5–10.

[8] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning Rich Features for Image Manipulation Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1053–1061.

[9] D. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, and A. K. Roy-Chowdhury, "Exploiting Spatial Structure for Localizing Manipulated Image Regions," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[10] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[11] L. Verdoliva, "Media Forensics and Deepfakes: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[12] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations (ICLR)*, 2021.

[14] Z. Liu, Y. Lin, Y. Cao et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 10012–10022.

[15] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[16] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "A Full-Image Full-Resolution End-to-End Trainable CNN Framework for Image Forgery Detection," *IEEE Access*, vol. 8, pp. 133488–133502, 2020.

[17] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
[18] B. Dolhansky, R. Howes, B. Pflaum et al., "The DeepFake Detection Challenge Dataset," arXiv:2006.07397, 2020.
[19] M. P. Gangan, A. Kadan, and L. V. L., "A Robust Image Forensic Framework Utilizing Multi-Colorspace Enriched Vision Transformer for Distinguishing Natural and Computer-Generated Images," arXiv:2308.07279, 2023.
[20] H. Liu, Z. Tan, C. Tan, Y. Wei, Y. Zhao, and J. Wang, "Forgery-Aware Adaptive Transformer for Generalizable Synthetic Image Detection," arXiv:2312.16649, 2023.