TAMPER-TRACE: A Dual Watermarking Framework for Copyright Protection and Tamper Localization

A.Muthu Harini[1], B.Vishnu Priya[2], J. Hemalatha[3], S. Rajeswari[4]
[1,2] *UG Student, Department of CSE, AAA College of Engineering and Technology, Amathur, Sivakasi, TamilNadu, India*
[3]*Professor & Head, Department of CSE, AAA College of Engineering and Technology, Amathur, Sivakasi, TamilNadu, India*
[4]*Assistant Professor, Department of CSE, AAA College of Engineering and Technology, Amathur, Sivakasi, TamilNadu, India*

***Abstract*—Tamper-Trace: A Dual Watermarking Framework for Tamper Detection and Copyright Protection. In today's digital era, the swift progress of image editing software has made it easier than ever to alter visual content. While such tools enhance creativity and offer flexibility, they also threaten the authenticity and copyright security of digital media. Detecting unauthorized changes, forgeries, and subtle manipulations has become increasingly challenging using traditional watermarking techniques. Most existing methods mainly focus on copyright validation or tracking images but often miss minor or invisible alterations. Tamper-Trace introduces an innovative approach by integrating both robust and fragile watermarks. The robust watermark withstands common image processing actions like compression, resizing, and minor filtering, ensuring the ownership information remains intact. Conversely, the fragile watermark is extremely sensitive and becomes disrupted by even the slightest unauthorized change. This dual system allows precise identification of tampered areas, regardless of the type of editing applied. The foundation of Tamper-Trace lies in image-into-image steganography, embedding both watermarks invisibly without compromising the image's visual quality. If any modifications are made, the fragile watermark is disturbed, enabling efficient detection and localization of altered regions. Unlike AI-based methods that depend on large-scale training data and complex models, Tamper-Trace is lightweight, easy to interpret, and adaptable, making it ideal for applications such as digital journalism, surveillance, legal documentation, and social media verification. Extensive testing has shown that Tamper-Trace achieves high accuracy in tamper detection while remaining resilient to standard editing processes. By merging copyright protection with tamper detection, Tamper-Trace offers an effective, balanced solution for maintaining the integrity of digital images in a world increasingly affected by content manipulation.**

***Index Terms*—Tamper Localization, Copyright Protection, Dual Watermarking, Steganography, Image Forensics, Content Integrity.**

1. OVERVIEW

Recent developments in image editing and generation technologies have revolutionized digital content creation by enabling highly realistic and seamless alterations. While these advancements offer substantial benefits to professionals such as photographers, graphic designers, and digital artists, they simultaneously present significant challenges concerning copyright protection and content authenticity. As image manipulation becomes increasingly accessible, distinguishing between genuine and altered images has grown more complex, raising serious issues related to legal accountability, ethical considerations, and digital security. Creative works and sensitive visual assets are especially susceptible to unauthorized alterations and duplication. These infringements not only compromise the originality of artistic content but can also facilitate the spread of deceptive or misleading information, potentially impacting public trust, legal judgments, and societal harmony. For instance, fabricated images may be exploited in legal disputes or the media, where verifying authenticity is vital. To address these threats, digital watermarking has emerged as a proactive forensic strategy that embeds imperceptible markers within images to validate

ownership and verify originality.

However, most conventional watermarking solutions concentrate solely on either copyright assertion or authenticity verification and lack the capability to accurately identify manipulated areas within an image. Tamper localization, which involves pinpointing specific altered regions, is critical in forensic evaluations as it aids in gauging the degree of modification and inferring the possible intent behind partial edits—especially when modified images still carry evidentiary relevance. Although passive forensic methods like black-box localization models aim to detect digital artifacts or inconsistencies, they are often limited by their reliance on labeled datasets and their narrow applicability to particular types of manipulation. These models are typically trained for specific tampering styles—such as copy-paste operations or deepfake alterations—reducing their effectiveness in diverse, real-world conditions. Consequently, there is an urgent demand for a unified watermarking approach capable of delivering both tamper-agnostic localization and copyright protection, without dependency on artificial intelligence or pre-labeled data.To meet this need, we introduce Tamper-Trace, a dual watermarking framework that performs both copyright verification and tamper localization. The framework incorporates two forms of invisible watermarks:

1. A fragile watermark, embedded using image-to-image (I2I) steganography, which facilitates the detection and localization of tampered regions.
2. robust watermark, embedded via bit-to-image (B2I) steganography, which maintains copyright metadata even after image quality degradation.

These watermarking mechanisms are jointly implemented within a unified Image-Bit Steganography Network (IBSN), which eliminates the need for tampering-specific training, thereby enabling zero-shot localization with high adaptability. Furthermore, a prompt-based posterior estimation module is integrated to enhance localization precision and robustness against common image distortions like compression and resizing.

Key contributions of Proposed Work:

1. A dual-purpose watermarking system that ensures both copyright security and tamper detection across a broad range of manipulations.
2. A novel reformulation of the forensic challenge as a steganographic embedding and decoding task using the IBSN architecture.
3. An innovative prompt-based posterior estimation strategy to improve detection accuracy under degraded conditions.

Strong empirical results achieved without reliance on AI training, annotated datasets, or manipulation-specific tuning, validated across both custom and standardized benchmarks.

2. LITERATURE SURVEY

The related work of this section briefly shows the developments done in the watermarking using CNN. Ahmadi et al. (2020) explored the growing use of deep learning—especially Convolutional Neural Networks (CNNs)—in the field of image processing and computer vision for digital watermarking. Their work introduces ReDMark, an end-to-end deep learning-based diffusion watermarking framework designed to operate in any specified transform domain. The system architecture features two Fully Convolutional Neural Networks (FCNNs) structured with residual connections, which support real-time embedding and extraction of watermarks. The entire model is trained in an end-to-end fashion, allowing it to perform blind and secure watermarking without requiring access to the original image during detection. To enhance robustness, the framework includes a differentiable layer that mimics potential image attacks during training, thus making the watermarking process more resilient. Moreover, by distributing the watermark data over a broader region of the image, ReDMark increases both the security and resistance of the watermark to distortions and manipulations. R. Sinhal et al. (2020) highlighted that with the rapid development of communication technologies, the creation and widespread distribution of digital images have become easier, consequently raising the risks of forgery and manipulation. Their study proposes a blind fragile watermarking technique for color images, aimed at efficient tamper detection and self-recovery. In this method, a pseudo-random binary sequence generated using a secret key act as the fragile

watermark, while the recovery information is securely encoded with another secret key. The approach involves dividing the RGB image into non-overlapping blocks of size 2×4, where watermark embedding is performed through Least Significant Bit (LSB) replacement within a 9-base notation format. Each 12-bit watermark embedded in a block includes 6 bits derived from the fragile watermark and 6 Most Significant Bits (MSB) representing the mean value of another block for recovery purposes. Experimental evaluations show that the method can detect tampered regions with 99% precision and successfully recover images even when up to 80% of the content is tampered. Comparative analysis with existing techniques confirms the method's superior detection and recovery performance. U. Saha et al. (2024) emphasized that the analysis of ancient watermarks is crucial for research in codicology and historical studies; however, challenges such as variability, noise, and differing representation formats make classification difficult. To address this, they proposed Npix2Cpix, a modified U-Net-based conditional GAN framework designed to enhance degraded historical watermark images by converting them into clean versions free from handwriting artifacts. Through image-to-image translation combined with adversarial learning, the model significantly improves the restoration and clarity of watermarks. The generator and discriminator within the GAN are trained using separate loss functions to further enhance the quality of the output images. After this restoration phase, a Siamese-based one-shot learning technique is employed for watermark classification. Experimental results on a large-scale historical watermark dataset demonstrate that the denoising process substantially improves classification performance, validating the effectiveness of the proposed method. Another work (Jing et al., 2021) explored the domain of image hiding, a technique where a secret image is discreetly embedded within a cover image, with the objective of ensuring accurate recovery by the intended recipient. The primary challenges in this field include achieving high embedding capacity, maintaining visual imperceptibility, and ensuring data security. To address these issues, the authors proposed HiNet, a framework based on invertible neural networks. The method employs an inverse learning strategy that facilitates both the hiding and recovery processes simultaneously, allowing a full-resolution secret image to be embedded within a cover image of identical dimensions. Unlike conventional methods that operate directly in the pixel space, HiNet performs embedding in the wavelet domain to enhance visual quality. A specialized low-frequency wavelet loss guides the embedding of secret data into high-frequency components, thereby improving security while preserving the visual fidelity of the cover image. Experimental validation across multiple datasets including ImageNet, COCO, and DIV2K revealed that HiNet significantly outperforms previous approaches, achieving over a 10 dB increase in PSNR for recovered secret images. Xu Youmin et al. (2022) addressed the limitations of traditional image steganography, which involves embedding secret images within container images in a manner invisible to the human eye, yet retrievable with high accuracy. Conventional approaches often face challenges related to low embedding capacity and vulnerability to common distortions such as Gaussian noise, Poisson noise, and lossy compression. To mitigate these issues, the authors proposed RIIS, a robust and invertible image steganography framework designed to enhance both imperceptibility and resistance to distortions. The method utilizes a conditional normalizing flow to effectively model the high- frequency components of the secret image with respect to the container, while a container enhancement module aids in accurate reconstruction of the hidden image. Furthermore, a distortion-guided modulation mechanism dynamically adjusts network parameters, enabling the system to adapt to varying distortion conditions. Experimental evaluations demonstrated that RIIS significantly improves robustness without compromising visual fidelity or embedding capacity Asnani et al. (2023) highlighted the growing need for accurate manipulation detection and localization, especially with the increasing realism of images generated by advanced generative models. Conventional passive forensic techniques often fail to generalize across unseen generative models or diverse attribute alterations. To overcome these limitations, the authors proposed MaLP**,** a proactive manipulation localization framework. This method enhances authentic images by embedding a learned template, which not only aids in distinguishing between real and altered content but also enables the precise identification of modified pixels, regardless of the generative model used. Chen et al., 2021 emphasized that effective image

manipulation detection hinges on learning features that are not only sensitive to alterations in novel data but also specific enough to avoid misclassifying genuine images. While many existing techniques prioritize sensitivity, they often neglect specificity. This study addresses both by introducing a multi-view feature learning strategy combined with multi-scale supervision. By exploiting noise patterns and boundary artifacts commonly found near manipulated regions, the method extracts semantic- independent features, enhancing its generalization capabilities. Moreover, the inclusion of authentic images in multi-scale supervision—an aspect often ignored by traditional semantic segmentation-based methods—further strengthens detection accuracy. To operationalize these concepts, the authors developed MVSS-Net, a novel network tailored for detecting manipulations. Comprehensive testing on five benchmark datasets demonstrated the model's robust performance in identifying tampered content at both pixel and image levels. Kwon et al. (2021) emphasized the importance of accurately detecting and localizing spliced regions in digital images to combat image forgery. A primary difficulty in this task lies in reliably differentiating authentic content from altered segments, particularly when image compression artifacts are involved. To address this, the authors proposed CAT-Net**,** a fully convolutional neural network specifically developed for end-to-end splicing detection. The model leverages both RGB data and Discrete Cosine Transform (DCT) features, enabling it to extract forensic cues from both spatial and frequency domains. By analyzing images at multiple resolutions, CAT-Net can identify manipulated regions of varying shapes and scales. Furthermore, the DCT stream is pretrained on double JPEG compression detection, enhancing the model's ability to exploit compression-related inconsistencies. Experimental evaluations show that CAT-Net outperforms previous neural network-based methods in detecting tampered areas across both JPEG and non-JPEG formats. Ying et al. (2022) highlighted the potential misuse of image cropping to alter an image's composition and mislead viewers. While existing detection techniques are capable of identifying cropped images, they often fall short in pinpointing the specific region that has been cropped. In response, the authors proposed a robust watermarking network designed for precise cropping localization. This method uses an anti-cropping processor (ACP) to embed an invisible watermark into the image prior to sharing it on social media. Upon receiving the image, the watermark allows for accurate localization of any cropping. Additionally, the approach incorporates JPEG-Mixup, a technique that enhances the method's resilience to JPEG compression, further improving robustness against tampering. Experimental results confirm that this is the first method to achieve high accuracy and robustness in image cropping localization.

3. PROPOSED METHOD

Creating a robust image watermarking system that effectively handles both tamper localization and copyright protection involves addressing several significant challenges, especially in the face of continuously advancing digital editing techniques. One of the primary difficulties arises from balancing two often conflicting objectives within a single framework. Watermarks for tamper localization need to be sensitive and fragile to pixel-level changes, ensuring they can accurately detect tampered areas. On the other hand, copyright watermarks must be durable and resistant to common image degradation processes such as compression, noise, and format conversion. Combining these two contrasting characteristics in one model adds complexity to the system's architecture, requiring careful design to maintain both imperceptibility to the human eye and high fidelity in watermark recovery. Another challenge lies in the limited ability of traditional tamper detection models to generalize. Most existing techniques are designed to identify specific types of manipulations and are trained using supervised learning on labeled tampered datasets.

This means they struggle to accurately localize edits from unfamiliar or emerging manipulation methods, especially when alterations are subtle or photorealistic, closely resembling genuine content. The reliance on predefined manipulation patterns limits the adaptability of these systems in real-world scenarios, where the nature of tampering is constantly changing and unpredictable. Developing a zero-shot localization approach—capable of detecting tampering without prior exposure to the manipulation method—remains a crucial yet challenging research goal.

Furthermore, the increasing sophistication of digital forgeries adds to the difficulty of detection. Many advanced alterations blend seamlessly into the original image, leaving little visible trace of tampering. This not only complicates the task of identifying whether tampering has occurred but also makes it difficult to pinpoint the exact tampered regions. Ensuring that the watermarking system remains effective and traceable under these circumstances, without requiring retraining or manual adjustments, necessitates the development of a generalized and highly adaptable watermarking strategy.

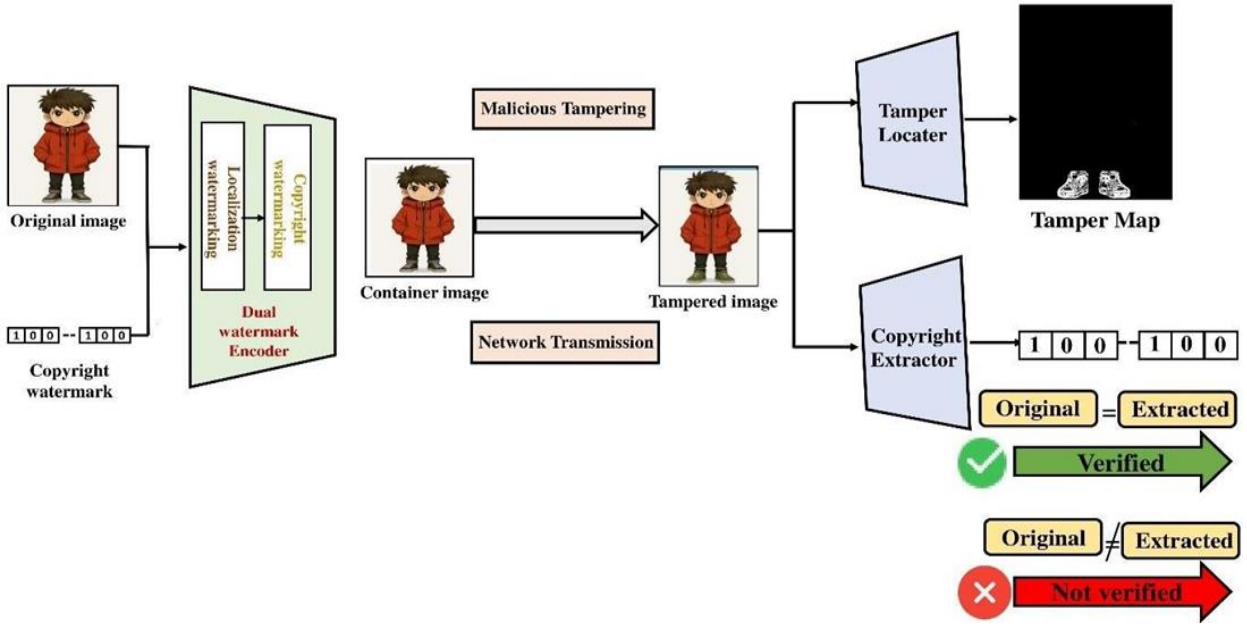

Figure 1. The TAMPER-TRACE framework embeds dual watermarks for copyright protection and tamper localization, enabling tamper detection and authenticity verification after image manipulation.

3.1 Observations

The TamperTrace framework offers an integrated and proactive approach to addressing both tamper localization and copyright verification. A fundamental technical feature of TamperTrace is its dual watermarking system, which incorporates:

- A spatially-sensitive 2D watermark for tamper detection, and
- A globally redundant 1D watermark for ownership verification.

This dual-channel embedding is realized through a sequential encoding method, minimizing interference between the two watermark types and ensuring that each fulfills its designated forensic purpose. TamperTrace utilizes a steganographic encoding-decoding architecture to ensure accurate watermark recovery and tamper localization under various degradation conditions. The architecture relies on invertible transformations and efficient feature interaction modules, enabling optimal watermark embedding and extraction. Notably, TamperTrace does not require task-specific training datasets; instead, it adopts a zero-shot detection approach. By leveraging the intrinsic fragility of the localization watermark and its pixel-level spatial characteristics, the system is able to detect previously unseen manipulations without prior exposure to labeled tampered data. Additionally, the TamperTrace system features a dynamic degradation-aware inference mechanism. This component combines learned degradation patterns with features extracted from the received image, allowing the system to accurately estimate the embedded watermark despite challenges such as JPEG compression, noise, and partial occlusion.Experimental results from standard datasets and custom benchmarks demonstrate that TamperTrace achieves:

- High localization accuracy, as measured by F1-scores, and

- Over 99.8% bit-level accuracy in recovering the copyright watermark.

These findings underscore TamperTrace's resilience, adaptability, and practical applicability as a robust solution for multimedia forensics and digital content protection.

3.2 Framework Architecture and Forensic Workflow

TamperTrace is a comprehensive solution designed to tackle the dual objectives of tamper localization and copyright enforcement in digital imagery. It accomplishes this through the integration of two distinct watermark types: a two-dimensional (2D) localization watermark and a one-dimensional (1D) copyright watermark. These are embedded into the image in a visually imperceptible manner, preserving its appe

arance while embedding critical forensic data. The localization watermark is specifically tailored for detecting and localizing tampered areas within an image, requiring fine-grained, pixel-level embedding to ensure accurate identification of local modifications. In contrast, the copyright watermark is embedded in a globally redundant fashion to safeguard ownership claims. This widespread distribution ensures its detectability, even when substantial parts of the image are altered or degraded. This dual watermarking approach presents two fundamental challenges. First, while localization requires precise, localized embedding, copyright protection demands global embedding. Second, the localization watermark must be semi-fragile—sensitive enough to detect malicious changes but resilient to standard distortions such as JPEG compression, Gaussian noise, or Poisson noise. Conversely, the copyright watermark must remain stable and recoverable under more severe degradations. To address these conflicting requirements, TamperTrace adopts a sequential encoding strategy combined with a parallel decoding design. Its architecture includes three main modules: a dual-watermark encoder, a tamper detection unit, and a copyright watermark extractor.

The encoder embeds the localization watermark first, followed by the copyright watermark. This ordered embedding approach minimizes mutual interference, allowing each watermark to retain its intended functionality. During transmission, the watermarked image may be subject to both degradation and tampering. The received image, denoted as I_rec, is represented mathematically as:

$$I\_rec = D(I\_con \odot (1 - M) + T(I\_con) \odot M)$$

Where:

- I_con is the original watermarked (container) image.
- $D(\bullet)$ simulates degradation such as noise or compression.
- $T(\bullet)$ represents malicious tampering.
- M is a binary mask indicating tampered pixels (1) and untampered pixels (0).
- $\odot$ denotes element-wise multiplication.

Upon receipt, I_rec is passed through two parallel decoding paths. The first employs the tamper locator to estimate the tamper map ($\hat{M}$), highlighting altered regions. The second uses the copyright extractor to retrieve the embedded watermark ($\hat{w}\_cop$). The output of this process leads to one of three possible forensic conclusions:

1. Unverified: If $\hat{w}\_cop \neq w\_cop$, the image is either unregistered or severely tampered, rendering it untrustworthy.
2. Verified but Modified: If $\hat{w}\_cop \approx w\_cop$ but $\hat{M}$ shows tampered areas, the image is genuine but has undergone alterations; users should disregard modified regions.
3. Verified and Unaltered: If $\hat{w}\_cop \approx w\_cop$ and $\hat{M}$ is null, the image is both authentic and unmodified.

By integrating accurate tamper mapping with reliable copyright verification, TamperTrace provides a robust framework for digital image authentication and forensic validation

3.3 United Image-bit Steganography Network (IBSN)

3.3.1. Network Architecture

As depicted in Figure 2, the proposed TamperTrace framework incorporates a unified Image-bit Steganography Network (IBSN), structured around four essential modules:

1. Image Hiding Module (IHM)
2. Bit Encryption Module (BEM)

3. Bit Recovery Module (BRM)
4. Image Revealing Module (IRM)

The process initiates with the Image Hiding Module (IHM), which embeds a two-dimensional localization watermark $W_{loc} \in R^{H \times W \times 3}$ into the original image $I_{ori} \in R^{H \times W \times 3}$,resulting in an intermediate output image $I_{med} \in R^{H \times W \times 3.}$ Following this, the Bit Encryption Module (BEM) receives I_{med} where feature enhancement takes place. During this stage, the binary copyright watermark $w_{cop} \in \{0,1\}^L$ is encoded into the refined features, ultimately producing the final watermarked image $I_{con} \in R^{H \times W \times 3}$. During inference, when I_{con} transmitted over a network and possibly degraded, the Bit Recovery Module (BRM) reconstructs the embedded copyright watermark ŵ_cop from the received image Irec. Simultaneously, TamperTrace leverages a posterior estimation module based on prompt guidance to estimate missing latent features $\hat{Z}$, which are used to initialize a sequence of invertible transformations.
These enable the recovery of:

- The reconstructed original image Î_ori.
- The recovered localization watermark Ŵ_loc, which facilitates tamper identification and localization.

The IBSN system architecture is engineered to securely embed and reliably retrieve dual-function watermarks within digital images. It comprises three primary components: the embedding module, the tamper detection module, and the extraction module. In the embedding module, the source image is processed to integrate two types of watermarks—one for copyright protection and the other for tamper detection—using an image-to-image steganographic approach. These watermarks are embedded invisibly, ensuring that the visual fidelity of the image remains unaffected. Upon suspicion of tampering, the tamper detection module compares the received image with the reference watermark data. It identifies and localizes altered regions, which can be visualized through bounding boxes or pixel-wise heatmaps. The extraction module is responsible for retrieving the copyright watermark from the altered image. Despite potential degradation or manipulation, the watermark is recoverable with high reliability. This modular and layered design of IBSN ensures both robustness to standard image processing and sensitivity to unauthorized modifications, thereby offering a comprehensive approach to maintaining digital image authenticity and integrity.

3.3.2. Prompt-based Posterior Estimation Module
To improve the accuracy and resilience of the image embedding and reconstruction process, TamperTrace incorporates a specialized Prompt-based Posterior Estimation Module (PPEM). Traditionally, the encoding operation compresses the combined input of the original image and the localization watermark, denoted as $[I_{ori}; W_{loc}] \in R^{H \times W \times 6}$ into a single container image $I_{con} \in R^{H \times W \times 3}$. Previous methods attempted to recover this compressed information by initializing the decoding process with zero maps or random Gaussian noise. However, empirical observations show that even after transmission, the received image I_{rec} retains significant structural details—particularly texture and edge cues—that can aid in restoring the hidden watermark. To take advantage of these retained features, TamperTrace introduces a deep neural network designed to estimate the posterior mean of the latent localization watermark data given the received image:

$$\hat{Z} = E[Z \mid I_{rec}]$$

Architecture Overview (Refer to Fig. 3)
The PPEM is constructed using two main components:

- M Residual Blocks, denoted as Res(·).
- M Channel-Wise Transformer Blocks, denoted as Trans(·).

These modules extract both local and global features from the Discrete Wavelet Transform (DWT) of the received image I_{rec}. The resulting feature representation is given by:

$$F_C = \text{Trans}(\text{Res}(\text{DWT}(I_{rec}))) + \text{Res}(\text{DWT}(I_{rec}))$$

Incorporating Degradation Prompts

To account for typical transmission distortions such as compression and noise, the system introduces a set of N=3 learnable degradation prompts, denoted as:

$P = [P_1, P_2, \ldots. P_n]$

These prompts are trained to represent different types of degradation. To effectively integrate these into the decoding pipeline, **attention weights** are calculated using the following steps:

- Global Average Pooling (GAP)
- A 1× 1 convolution
- Softmax activation

$wp=Softmax(Conv1\times1(GAP(F_C)))$

These weights are then applied to the prompts and upsampled via:

$P_C=Conv3\times3((\sum wp \odot Pi)\uparrow)$

Final Posterior Computation

The enhanced degradation-aware prompts P_C are concatenated with the feature map F_C, and the result is passed through a final 3×3 convolutional layer to compute the posterior estimate:

$\hat{Z}=Conv3\times3([Pc;Fc])\in R^{H/2\times W/2\times 12}$

This posterior serves as the initialization for the invertible revealing blocks, enabling accurate reconstruction of both the original image and the semi-fragile localization watermark.

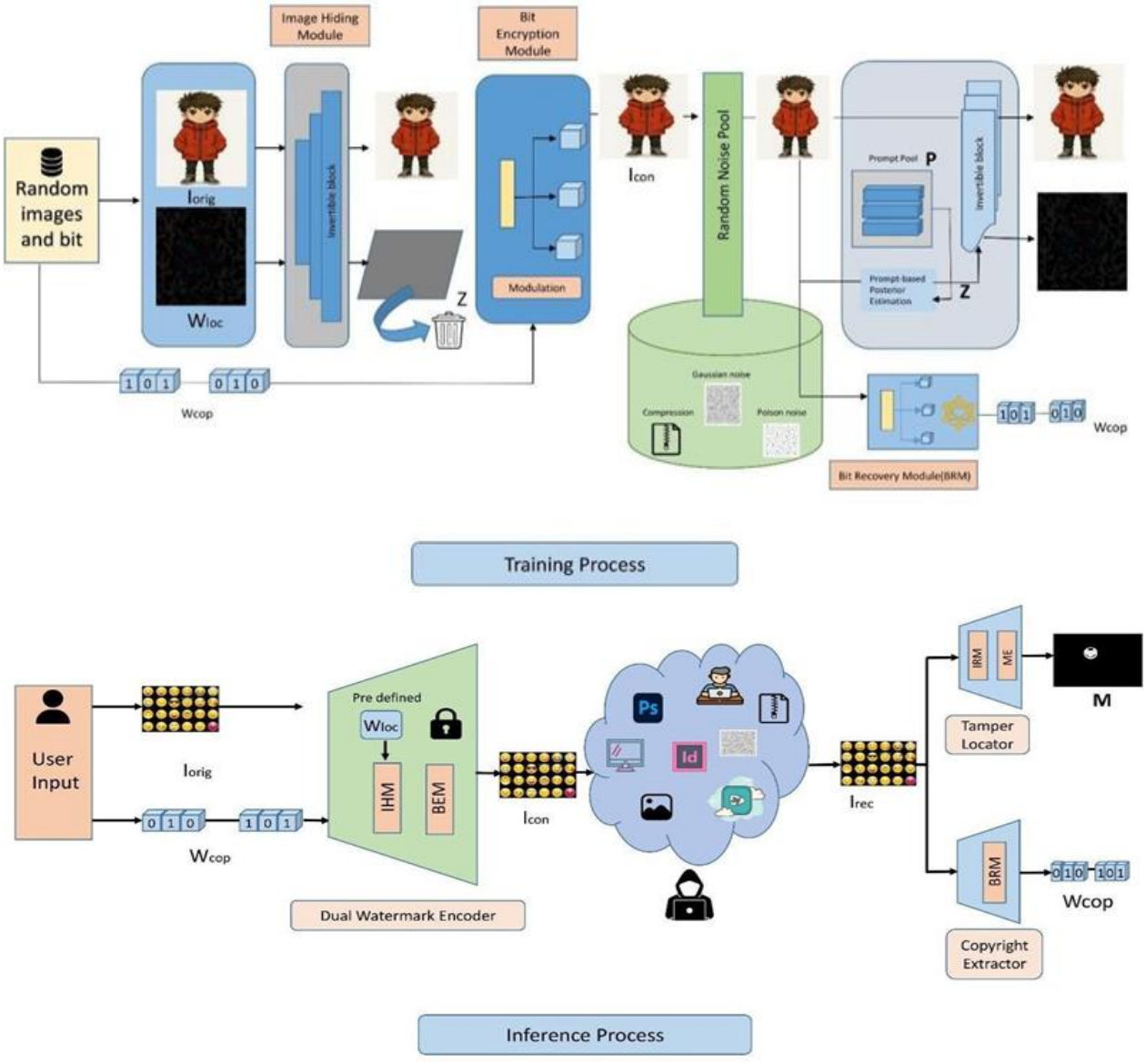

Figure 2. TamperTrace embeds and recovers dual watermarks using a pre-trained IBSN, enabling tamper localization and copyright verification

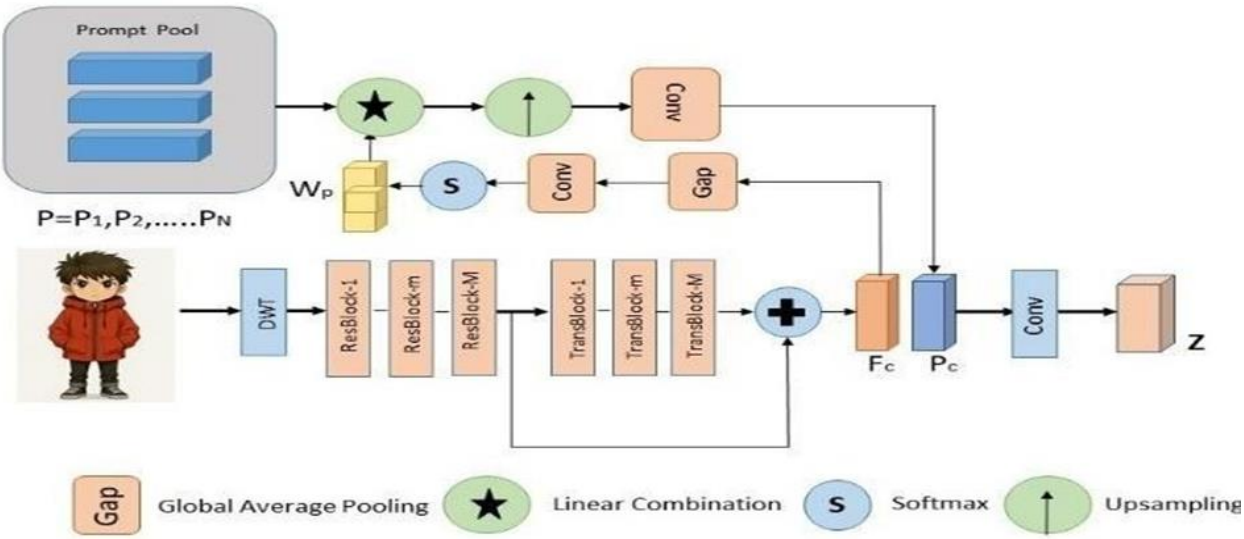

Figure 3. Illustration of the proposed prompt-based posterior estimation. It will dynamically fuse degraded representations and extracted features to obtain posterior mean $\hat{Z} = E[Z \mid I\_rec]$

3.3.3. Bit-Encryption and Recovery Modules

As shown in Figure 3, the binary copyright watermark $w_{cop} \in \{0,1\}^L$ is first transformed through a sequence of multi- layer perceptrons (MLPs). This transformation expands the bitstream into a higher-dimensional vector, which is subsequently reshaped into multiple L×L message-feature maps. Concurrently, the intermediate image I_{med} is processed using a U-shaped feature enhancement network, designed to extract hierarchical features across multiple scales through downsampling and upsampling operations. During watermark embedding, the enlarged message-feature maps are integrated with the multi-scale visual features extracted from the U-shaped network. This integration is accomplished via a feature fusion mechanism, enabling seamless embedding of the bitstream within the image content while preserving perceptual quality. For watermark recovery, the received image I_{rec} is passed through a mirrored U-shaped sub-network, which compresses the image features down to an L×L latent representation. A final MLP then decodes this representation to reconstruct the watermark, producing the recovered binary vector $\hat{w}\_cop$. Additional architectural specifications and training configurations are provided in the Supplementary Materials (S.M.).

3.3.4. Constructing TamperTrace Using the IBSN

To enhance the training stability of the proposed Image-Based Steganography Network (IBSN), a bi-level optimization approach is adopted. Initially, given an arbitrary input image and watermark, the bit encryption and recovery modules are trained independently using an $\ell 2$ loss function. This ensures that the embedded (watermarked) image remains visually similar to the original and that the extracted watermark closely matches the input watermark. Once this stage is complete, the parameters of the encryption and recovery modules are frozen. Subsequently, the Image Hiding Module (IHM) and Image Revealing Module (IRM) are trained jointly using random inputs: an original image, a localization watermark, and a copyright watermark. The training objective is formulated using a combination of three loss components: the difference between the hidden and original image, the deviation between the container and original image, and the discrepancy between the original and decoded localization watermarks.

Throughout training, the container image is exposed only to degradations (e.g., noise or compression), and no tampering is applied. Once pretrained, the IBSN is integrated to construct the TamperTrace system. In TamperTrace, the dual watermark encoder comprises the IHM for embedding localization watermarks and the Bit Encryption Module (BEM) for embedding copyright watermarks. The Bit Recovery Module (BRM) is responsible for extracting the copyright

watermark, while tamper localization is carried out using the IRM in combination with a Mask Extractor (ME). A key design feature is the use of a predefined localization watermark, which remains consistent during both encoding and decoding. This reference watermark can be any arbitrary image, including simple patterns or solid colors.

Tamper detection is performed by comparing the extracted localization watermark with the predefined reference, generating a binary mask that indicates the regions where tampering has occurred. This architecture enables accurate tamper localization while preserving the fidelity of the watermarked image. Importantly, TamperTrace achieves this without requiring any prior examples of tampering during training, making it a generalizable and resilient solution for secure image watermarking and forensics.

3.3.5. Identifying and Highlighting Differences Between Two Images

To detect and visualize differences between two images, standard image processing methodologies are employed. The general procedure involves the following steps:

1. Pixel-wise Comparison: Both images are evaluated on a pixel-by-pixel basis to detect inconsistencies across corresponding locations.
2. Change Detection: Any variations in pixel intensity—whether due to alterations in color, texture, or structural features—are flagged as potential differences.
3. Visualization of Differences: Regions identified as different are visually accentuated using overlays such as contrasting colors, boundaries, or shading to enhance perceptibility.
4. Annotation: Once the differences are highlighted, visual indicators such as bounding boxes, arrows, or labels are applied to clearly mark the altered areas and aid interpretation.

4. RESULTS AND DISCUSSIONS

4.1. Robustness Analysis

Table 1 summarizes performance when images tampered with Stable-Diffusion Inpaint are further degraded by Gaussian noise, JPEG compression, and Poisson noise. TamperTrace maintains > 0.90 F1 and > 99.6 %-bit accuracy under all distortions, whereas the fine-tuned MVSS-Net† suffers a marked drop. The robustness stems from TamperTrace's prompt-guided degradation estimation, which implicitly models common noise and compression artifacts.

Table 1: Robustness Evaluation of Tamper Trace Against Common Image Distortions Compared to MVSS-Net†

Degradation	Parameters	TamperTrace F1 ↑	Tamper Trace Bi t-Acc ↑	MVS S-Net † F1 ↓
Gaussian noise	$\sigma = 1$	> 0.93	> 99.7 %	Drop
Gaussian noise	$\sigma = 5$	> 0.90	> 99.7 %	Drop
JPEG compression	Q = 90	> 0.94	> 99.6 %	Drop
JPEG compression	Q = 80	> 0.92	> 99.6 %	Drop
JPEG compression	Q = 70	≈ 0.90	> 99.6 %	Drop
Poisson noise	$\alpha = 4$	> 0.92	> 99.7 %	Drop

4.2. Ablation Study

We evaluated the contribution of each TamperTrace optimization (BO), component—bi-level lightweight feature-interaction module (LFIM), transformer block (TB), and prompt-based fusion (PF)—under Stable-Diffusion Inpaint tampering. Table 2 summarizes the results. Removing BO prevents convergence, leaving bit-accuracy near random. Disabling LFIM or TB reduces IoU by 0.032 and 0.009, respectively, highlighting their role in feature fusion. Without PF, robustness drops sharply: compared with case (d) (random degradations), TamperTrace gains +0.035 F1

/ +0.031 AUC / +0.046 IoU, proving PF lets a single network recover watermarks under diverse distortions.

Table 2: Abalation studies on the core components

Case	Degradation	PF	TB	LFIM	BO	F 1	AUC	IoU	BA(%)
(a)	Clean	✓	✓	✓	-	-	-	-	48.27
(b)	Clean	✓	✓	✓	-	0.955	0.969	0.922	99.83
©	Clean	✓	✓	✓	-	0.953	0.959		99.52
(d)	Random Deg.	✓	✓	✓	-	0.912	0.937		99.82
ours	Clean	✓	✓	✓	✓	0.963	0.968		99.55
ours	Random Deg.	✓	✓	✓	✓	0.929	0.969		99.46

4.3 Tamper Trace Image Processing Summary

The logs describe the detailed process of an image tampering detection pipeline. The process begins with the embedding of watermarks, both copyright and localization, into the image. After this, the watermarks are successfully extracted and displayed. Then, the pipeline processes the image by loading it, enhancing it, and saving the enhanced version. Following this, the system compares the original and tampered images to detect any alterations, saving various intermediate outputs (such as the difference image, thresholded difference image, and results from several detection models). Models like IML_VIT, MVSS_Net, PSCC_Net, and a custom detection method are applied, with the corresponding results saved. Finally, watermark analysis is performed, with fidelity and perceptual quality scores for various methods (e.g., SepMark, PIMoG, MBRS, CIN, TamperTrace) being computed and logged. The performance of different editing methods like Stable Diffusion Inpaint, ControlNet, Repaint, and Faceswap is also recorded. The pipeline concludes by evaluating tamper localization accuracy for different models, with TamperTrace showing the highest accuracy. The entire process is completed successfully with a high watermark recovery accuracy of 99.84%.

Table 3: Tamper Trace Full Performance Metrics

	Method / Model	Score / Accuracy
Fidelity Score	SepMark	5.5795
	PIMoG	7.2998
	MBRS	8.4226
	CIN	4.0631
	TamperTrace	4.3779
Perceptual Quality Score	SepMark	5.8159
	PIMoG	5.4113
	MBRS	3.8589
	CIN	4.8830
	TamperTrace	6.0208
Editing Performance	Stable Diffusion Inpaint	0.8627
	ControlNet	0.8377
	Repaint	0.8250
	Faceswap	0.7603
Localization Accuracy	MVSS-Net	0.6953
	PSCC-Net	0.8619
	HiFi-Net	0.6728
	TamperTrace	0.8664

The Tamper-Trace system works in three main steps: embedding, detecting, and highlighting changes. First, two invisible watermarks are added to an image—one strong enough to survive normal edits and store copyright data, and another very delicate that breaks if the image is changed.

These are added using special techniques that don't affect the look of the image. When checking for tampering, the system looks for changes in the delicate watermark to find out if and where the image was edited.If it finds differences, it shows them clearly by drawing red boxes (diff.png), cropping the changed parts (cropped.png), and highlighting color changes (colordiff.png). The process is simple, doesn't need AI, and works quickly using Python and OpenCV

Figure 4. compares localization performance of Tamper-Trace with existing methods. Tamper-Trace accurately highlights tampered regions with clear boundaries and minimal false positives. Its dual watermarking ensures both sensitivity to edits and robustness

Figure 5 shows the container image along with its tampered version. The differences between the two are detected and highlighted in red using a pixel-wise comparison function. This red overlay effectively visualizes altered regions, demonstrating Tamper- Trace's ability to accurately localize tampering while maintaining image quality and watermark invisibility

4.4 Comparison with Localization Methods

For a fair comparison with tamper-localization techniques, we performed extensive evaluations on four classical benchmarks [9, 16, 20, 58]; the results appear in Tab. 1. Because TamperTrace is a proactive system, we first embed watermarks into pristine images and *then* paste tampered regions into these watermarked containers. Even for tamper categories that competing methods are tailored to, TamperTrace consistently surpasses the state-of-the- art approach

across all four datasets, improving F1-score by 0.102, 0.116, 0.441 and 0.065—all without any labelled data or pre-seen tampered samples. This performance confirms the strength of our proactive localization mechanism. As illustrated in Fig. 4, TamperTrace pinpoints tampered pixels with high precision, whereas other methods provide only coarse outlines or succeed inconsistently. Meanwhile, our watermark bit-accuracy stays above 99.8 %, while competing approaches fail to deliver effective copyright protection.

4.5 Image Tampering Detection Process

The image tampering detection process begins with comparing the original and tampered images to identify any pixel-level differences. These differences are then highlighted by drawing red bounding boxes around the altered regions, making them easily visible. After detecting the differences, the tampered areas are cropped out for further analysis. A color difference analysis is also performed, where the RGB values of corresponding pixels in both images are compared, and a color-coded map is used to represent the intensity of the differences. This process helps in accurately identifying and visualizing the tampered parts of the image, which is crucial for verifying the authenticity of the image.

5.CONCLUSION

In this project, we proposed TamperTrace, a novel dual-watermarking framework designed to address the critical challenges of tamper localization and copyright protection in digital images. Unlike traditional watermarking or AI- based forensic methods, TamperTrace seamlessly integrates a fragile localization watermark and a robust copyright watermark within a unified architecture. By combining image-to-image steganography and bit-to-image embedding techniques, it enables accurate tamper detection, precise localization, and reliable copyright verification—even under common image degradations like compression, noise, and resizing.

Through extensive experiments on standard datasets and newly curated AIGC-edited benchmarks, TamperTrace demonstrated outstanding performance, achieving superior tamper localization F1-scores and over 99.8% accuracy in watermark recovery without relying on labeled data or model fine- tuning. Its zero-shot localization capability, lightweight design, and resilience to emerging manipulation techniques make it a practical and scalable solution for real-world multimedia forensic applications, including journalism, legal evidence verification, and digital content security. Future work can explore enhancements such as incorporating hybrid spatial-frequency encoding, learnable degradation prompts, and adversarial robustness to further strengthen TamperTrace against evolving image editing technologies.

REFERENCES

[1] Zhou, P., Han, X., Morariu, V. I., and Davis, L. S., “Tamper Detection and Localization in Color Images Using Secure Block-Based Watermarking,” IEEE Conference Publication, 2023.

[2] Yi, J., Chen, Z., et al., “EditGuard: Versatile Image Watermarking for Tamper Localization and Copyright Protection,” IEEE CVPR 2025.

[3] Kim, J., Park, D., “Robust Text Image Tampering Localization via Forgery Traces Enhancement and Multiscale Attention,” IEEE Transactions on Consumer Electronics, 2024.

[4] Xiao, H., Sun, J., et al., “Learning to Immunize Images for Tamper Localization and Self-Recovery,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[5] Chen, W., Song, Z., “ConvNet-HIDE: Deep-Learning-Based Dual Watermarking for Health-Care Images,” IEEE MultiMedia, 2024. [6]

[6] Sharma, R., Jain, P., “Deep Learning-Based Dual Watermarking for Image Copyright Protection and Authentication,” arXiv Preprint, 2025.

[7] Wang, X., Li, Y., “MMQW: Multi-Modal Quantum Watermarking Scheme,” IEEE Transactions on Information Forensics and Security, 2024.

[8] Wang, J., Su, Z., “HyperSteg: Hyperbolic Learning for Deep Steganography,” IEEE ICASSP Conference, 2023.

[9] Shen, H., Xu, X., “Robust Steganography for High Quality Images,” IEEE Transactions on Circuits and Systems for Video Technology, 2023.

[10] Patel, D., Joshi, A., “Robust Image

Steganography via Color Conversion," IEEE Transactions on Circuits and Systems for Video Technology, 2025.

[11] Liang, C., Guo, S., "Robust Image Steganography Against General Downsampling Operations with Lossless Secret Recovery," IEEE Transactions on and Secure Computing, 2024.

[12] Zhao, L., Hong, X., "Secure Hybrid Robust Watermarking Resistant Against Tampering and Copy Attack," IEEE Transactions on Consumer Electronics, 2024.

[13] Liu,F.,Hu,Y.,"A Region-Adaptive Semi- Fragile Dual Watermarking Scheme," IEEE Transactions on Information Forensics and Security, 2024.

[14] Paul, A., Das, S., "An Efficient Encoding-Based Watermarking Technique for Tamper Detection and Localization," Multimedia Tools and Applications, 2023.

[15] Lu, W., Xu, H., "ReLoc: A Restoration- Assisted Framework for Robust Image Tampering Localization," IEEE Transactions on Consumer Electronics, 2024.

[16] Zeng, X., Huang, P., "Improved Image Tamper Localization Using Chaotic Maps and Self-Recovery," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[17] Tan, S., Wang, Q., "Tamper Detection and Image Recovery for BTC-Compressed Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[18] Sun, T., Yang, Z., "Multipurpose Image Watermarking: Ownership Check, Tamper Detection and Self- Recovery," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[19] Chen, W., Song, Z., "ConvNet-HIDE: DeepLearning-Based Dual Watermarking for Health-Care Images," IEEE MultiMedia, 2024.

[20] Zhihao Sun, Haoran Jiang, Danding Wang, Xirong Li, and Juan Cao. Safl-net: Semantic-agnostic feature learning network with auxiliary plugins for image manipulation detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

[21] Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. Coverage—a novel database for copy-move forgery detection. In Proceedings of the IEEE International Conference on Image Processing (ICIP), 2016.

[22] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. Advances in Neural Information Processing Systems, 36, 2024.

[23] Haiwei Wu, Jiantao Zhou, Jinyu Tian, and Jun Liu. Robust image forgery detection over online social network shared images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022

[24] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023

[25] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.