GenPTW: In-Generation Image Watermarking for Provenance Tracing and Tamper Localization

Zhenliang Gan, Chunya Liu, Yichao Tang, Binghao Wang, Weiqiang Wang, Xinpeng Zhang
School of Computer Science, Fudan University
Shanghai, China
zlgan23@m.fudan.edu.cn,yichao_tang@fudan.edu.cn,zhangxinpeng@fudan.edu.cn

Abstract

The rapid development of generative image models has brought tremendous opportunities to AI-generated content (AIGC) creation, while also introducing critical challenges in ensuring content authenticity and copyright ownership. Existing image watermarking methods, though partially effective, often rely on post-processing or reference images, and struggle to balance fidelity, robustness, and tamper localization. To address these limitations, we propose **GenPTW**, an In-**Gen**eration image watermarking framework for latent diffusion models (LDMs), which integrates **P**rovenance Tracing and **T**amper Localization into a unified **W**atermark-based design. It embeds structured watermark signals during the image generation phase, enabling unified provenance tracing and tamper localization. For extraction, we construct a frequency-coordinated decoder to improve robustness and localization precision in complex editing scenarios. Additionally, a distortion layer that simulates AIGC editing is introduced to enhance robustness. Extensive experiments demonstrate that GenPTW outperforms existing methods in image fidelity, watermark extraction accuracy, and tamper localization performance, offering an efficient and practical solution for trustworthy AIGC image generation.

CCS Concepts

• **Security and privacy** → **Database and storage security**.

Keywords

Latent diffusion model, image generation, responsible ai, image watermarking, security

1 Introduction

Generative models are evolving at an unprecedented pace, particularly text-to-image (T2I) diffusion models such as Stable Diffusion, DALL·E 3, and Imagen. These models are capable of synthesizing highly realistic and visually compelling images, while also supporting flexible editing, thereby reshaping the landscape of visual content creation. However, this impressive generative capability is a double-edged sword, introducing a range of security risks including content misuse, ambiguous copyright ownership, and difficulties in tamper detection. In recent years, incidents involving AI-generated images being stolen, maliciously edited, or even forged as fabricated evidence have become increasingly common, threatening both public discourse and the credibility of legal systems. These issues fundamentally highlight two critical challenges: verifying content authenticity and tracing generative responsibility.

Image watermarking is a widely adopted technique for copyright protection and provenance tracing. However, most existing methods focus primarily on authenticity verification and ownership identification, falling short in terms of accurately localizing tampered regions. Tamper localization plays a crucial role in delineating the boundary between generated and modified content, clarifying the responsibility of generative models. Moreover, it enables the assessment of tampering severity and reveals potential malicious intent, making it a key component for achieving comprehensive traceability in AIGC-generated images.

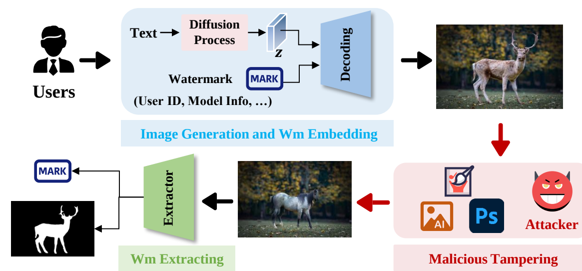

Figure 1: The process of embedding and extracting “GenPTW” for dual forensic objectives.

Several recent studies have begun to explore the integration of copyright identification and tamper detection. For instance, SepMark [50] introduces a separable watermarking structure to improve robustness against attacks, while EditGuard [59] leverages local vulnerabilities in image steganography to enable tamper region localization. However, these methods follow a post-generation paradigm, where watermarks are embedded after the image has been generated. This leads to a disconnect from the generation process, increased deployment complexity, and reduced overall efficiency.

Therefore, recent studies have shifted towards embedding watermarks directly within the diffusion-based generation process, known as In-Generation watermarking. For example, Stable Signature [14] injects watermarks during generation by fine-tuning the VAE decoder, but requires training a separate model for each watermark, making it unsuitable for large-scale deployment. Moreover, existing In-Generation watermarking methods are vulnerable to AIGC edits and aggressive degradations (e.g., composite attacks, JPEG compression), often resulting in complete watermark loss. Most of these methods also lack the capability to localize tampered regions, restricting their forensic effectiveness.

To mitigate these risks, it is imperative to develop a verifiable and traceable watermarking mechanism for AIGC-generated images, along with enhanced capabilities for tamper localization and responsibility attribution in the presence of malicious modifications. To clarify the task boundaries, we redefine the dual forensic objectives as illustrated in Fig. 1: **(1) Provenance tracing**, which is used to track source information, including the model, user, time, and event details. ; and **(2) Tamper localization**, which accurately identifies and highlights pixel-level manipulated regions for visual evidence.

To this end, we propose GenPTW, a watermarking framework tailored to latent-space diffusion models, which unifies watermark embedding, extraction, and tamper localization within a single architecture. In the embedding stage, watermark information is injected into multi-scale latent features during the image generation process, without disrupting the structure of the diffusion pipeline. In the extraction stage, we design a frequency-coordinated decoder, which extracts the embedded copyright watermark from the low-frequency components of the generated image, and localizes tampered regions from the high-frequency components. Additionally, the watermark feature map obtained from the low-frequency branch is used as an auxiliary cue to enhance tamper localization accuracy. To improve robustness, we design a distortion layer that simulates AIGC editing operations, including inpainting operations and VAE reconstructions, enabling the model to better withstand various types of manipulations and degradations. Furthermore, a gradient-guided encoder is employed to embed the watermark under Just Noticeable Difference (JND) constraints, using a modification cost map, and is regularized across multiple latent-space scales to ensure both invisibility and fidelity. Our contributions are summarized as follows:

(1) We propose GenPTW, a unified proactive defense framework that integrates provenance tracing and tamper localization tasks, achieving tight coupling between the encoding and decoding processes.

(2) We construct a frequency-coordinated decoder for watermark extraction and tamper localization, which significantly improves extraction accuracy and localization robustness under various degradation attacks.

(3) We introduce a distortion layer that simulates AIGC edits to enhance robustness, and use multi-scale loss in spatial and latent domains to improve visual quality.

(4) Extensive experiments show that GenPTW achieves superior performance over existing watermarking and forensic baselines in terms of visual fidelity, flexibility, and robustness.

2 Related Work

2.1 Image Tamper Detection and Localization

Passive Methods.

Passive image analysis methods examine intrinsic attributes such as statistical features, lighting conditions, color distribution, noise discrepancies, and DCT correlations [7, 10, 18, 26, 34] to identify tampering without external information. Traditional hand-crafted methods were limited by poor generalization and insufficient robustness, leading to the adoption of deep learning-based approaches [5, 42, 51, 52, 65].

For instance, Zhuang et al. [66] developed an encoder-decoder framework incorporating dense connections and dilated convolutions, while DOA-GAN [22] introduced a dual-order attention GAN to improve localization accuracy. Wu et al. [49] utilized noise modules to enhanced robustness against social media distortions. HiFi-Net [17] proposes hierarchical feature analysis and refinement strategies. TruFor [16] enhances sensitivity to tampering traces by combining RGB images with noise fingerprints. Additionally, Diff-Forensics [55] employs diffusion models as feature extractors for tamper localization. Despite these advancements, passive methods still require domain-specific training data for optimal performance.

Proactive Methods.

Proactive methods involve embedding imperceptible markers or watermarks into images, which are easily destoryed or altered when tampering occurs. Traditional fragile watermarking method, such as block-wise hash verification or pixel-level grayscale analysis [6, 21, 27, 28, 30, 38], have limited localization accuracy and flexibility. To address these limitations, deep learning-based approaches have been developed. For instance, FakeTagger [46] leverages recoverable one-hot encoding messages for tamper verification by embedding messages and recovering them after manipulation. Similarly, MaLP [2] adds a learned template to encrypted real images to enhance tamper detection and localization.

More recently, methods like EditGuard [59], V2AMark [61], and OmniGuard [60] have employed two-stage embedding for pixel-level localization and copyright protection, through combining steganography and watermarking technology. However, they still require a preset steganographic template to ensure precise pixel-level tamper localization.

2.2 Image Watermarking

Post-hoc Watermarking.

Digital watermarking plays a crucial role in traceability, content authentication, and copyright protection. Traditional watermarking methods, such as DwtDct [37] and DwtDctSvd [37], manually design embedding mechanisms to insert watermark into imperceptible spatial or frequency domains. DNN methods optimize the trade-offs between invisibility and robustness more effectively than manual designs. For example, Hidden [64] was a pioneering end-to-end (END) framework for watermark embedding and extraction. Building on this, De-END [13] enhances information interaction between the encoder and decoder, while CIN [35] uses a flow-based reversible network to ensure coupling of embedding and extraction processes. To improve robustness, differentiable noise layers are used to simulate real-world distortions such as JPEG compression, screenshot capture, or photographic degradation [MBRS [23], StegaStamp [44], Pimog [12], LFM [48], and DeNol [11]] during training. Despite these advancements, they remain vulnerable to new attacks such as AIGC-inpainting.

Recent work, such as Robust-Wide [19] designed denoise sampling guidance module and OmniGuard [60] proposed a lightweight AIGC editing simulation layer to enhance robustness against AIGC in-painting. However, post-hoc watermarking techniques are relatively easy to remove, and users can evade the watermark.

In-Generation Watermarking.

In-generation watermarking involves embedding watermark during the image creation process itself, focusing on three main strategies: Firstly, *Initial Noise Modulation*. Methods like Tree-Ring [47] embed watermark features by altering the Fourier spectrum of initial Gaussian noise vectors, while Gaussian Shading [54] encodes watermarks as encrypted Gaussian-distributed patterns injected into initial noise. These approaches avoid the need for model fine-tuning but may compromise the quality and diversity of generated images due to random noise distributions change. condly, *Dataset Attribution*. Techniques such as WatermarkDM [63], ProMark [1], and Diffusion-Shield [8] embed copyright info by retraining diffusion models on watermarked datasets. While these methods allow for watermark extraction from generated images, they require extensive computational resources and risk degrading model performance after the retraining process. Thirdly, *Latent Space Adaptation*. Stable Signature [14] fine-tunes the VAE decoder to imprint watermarks but requires separate copies of the decoder for each watermark, which hinders scalability. RoSteALS [4] exploits latent space redundancy to embed watermarks without modifying the decoder. Similarly, WOUAF [24] maps fingerprints into the latent space and embeds watermark information by fine-tuning the decoder parameters through weight modulation, thereby achieving high attribution accuracy while maintaining output quality. In contrast, LaWa [40] integrates watermark features into latent variables via auxiliary networks while keeping decoder parameters frozen, thus ensuring scalability and efficiency.

3 Method

3.1 Overall Framework of GenPTW

As illustrated in Fig. 2, we present GenPTW, a unified watermarking framework tailored for latent diffusion models. It supports joint forensic objectives of provenance attribution and tamper localization within a single architecture. Unlike prior methods that separate watermark extraction and tamper detection into two independent modules, which often require redundant embedding of both ownership and localization watermarks, GenPTW integrates the two tasks within a unified design.

In the embedding phase, a latent representation is first generated by the diffusion process. Given a watermark message (e.g., user ID), GenPTW enables the pre-trained latent decoder to simultaneously embed the watermark into the latent space and decode it into a watermarked image. In the extraction phase, we design a frequency-coordinated decoder that leverages the robustness of low-frequency components to extract the watermark, while exploiting the tamper sensitivity of high-frequency details to detect manipulated regions. Moreover, watermark features from the low-frequency branch serve as auxiliary cues to guide the high-frequency localization stream, thereby improving accuracy. To improve resilience under real-world AIGC manipulations, we introduce a distortion simulation layer that simulates AIGC edits. Additionally, a JND-constrained perceptual loss is applied in the embedding phase, using a pixel-wise cost map to control perturbation strength and location, ensuring watermark imperceptibility while preserving image quality.

This unified design allows GenPTW to achieve robust watermark extraction and accurate tamper localization under diverse AIGC distortion scenarios. The following sections elaborate on each component of the framework.

3.2 Multi-scale Latent Space Embedding

We follow the latent diffusion model (LDM) paradigm [41], where the image I_{source} is encoded into a compact latent representation $z = \mathcal{E}(I_{source})$ by a factor of f and decoded by $\mathcal{D}(z)$ in a multi-stage manner. During generation, the diffusion process synthesizes z, which is progressively upsampled to reconstruct the final image.

To embed watermark information, we adopt a coarse-to-fine strategy that injects the message into latent features at multiple decoder stages. Given a k-bit binary watermark message $m \in \{0,1\}^k$, a message processor W_{Pro} generates the initial watermark embedding w_0, which is added to z before decoding. At each subsequent decoder stage $i \in \left\{1,\dots,\frac{f}{2}\right\}$, a watermark feature encoder W_{Emb_i} takes the previous watermark feature w_{i-1} as input and outputs a spatial watermark feature w_i that matches the shape of the corresponding latent feature z_i. The watermarked latent z_{m_i} is then computed and passed to the next decoding stage:

$$w_i = W_{Emb_i}(w_{i-1}), \quad z_{m_i} = z_i + w_i \tag{1}$$

The modified decoder $\mathcal{D}_w$ replaces $\mathcal{D}$ to reconstruct the watermarked image $I_w = \mathcal{D}_w(z)$.

3.3 Frequency-Coordinated Decoder

We design a frequency-coordinated decoder that performs tamper localization using high-frequency features and watermark extraction using low-frequency features. Prior studies have shown that high-frequency components are more sensitive to local manipulations [33, 45, 53], while low-frequency information remains stable under various distortions [39, 56]. As illustrated in Fig. 3, tampered regions often exhibit more noticeable artifacts in the high-frequency domain, whereas low-frequency representations demonstrate stronger robustness. To improve reliability under severe degradations, we incorporate the low-frequency watermark feature map as an auxiliary cue to enhance the robustness and accuracy of tamper localization.

As shown in Fig. 2, the generated watermarked image I_w is first passed through a distortion simulation layer to obtain the degraded image I_d. We apply the Discrete Cosine Transform (DCT) [15] to extract its high- and low-frequency components. The low-frequency component I_l is fed into the watermark decoder W_{Dec} to produce a spatial watermark feature map W_{map}, which is further processed by an MLP and a Sigmoid activation to yield the final predicted message $\hat{m}$.

$$\mathbf{W}_{map} = \mathbf{W}_{Dec}(\mathbf{I}_l), \quad \hat{\mathbf{m}} = \mathrm{Sigmoid}(\mathrm{MLP}(\mathbf{W}_{map})) \tag{2}$$

The watermark feature map $\mathbf{W}_{map}$ is concatenated with the high-frequency feature $\mathbf{I}_h$ and fed into a ConvNeXt [32]-based global feature encoder $\mathbf{CN}_{Enc}$ to extract multi-scale features:

$$\mathbf{I}_{all} = \{\mathbf{I}_h, \mathbf{W}_{map}\} \tag{3}$$

$$\begin{gathered}\{F_{s_1}, F_{s_2}, F_{s_3}, F_{s_4}\} = \mathbf{CN}_{Enc}(\mathbf{I}_{all}),\\ F_{s_i} \in \mathbb{R}^{\frac{H}{2^{(i+1)}} \times \frac{W}{2^{(i+1)}} \times C_i}\end{gathered} \tag{4}$$

Here, C_i denote the total number of output channels at each scale i.

Each feature map G_{s_i} is then processed by the multi-scale decoder to generate the corresponding tamper prediction mask:

$$P_{f_i} = \text{Multi-Scale Decoder}(F_{s_i}) \tag{5}$$

To enhance multi-scale fusion, we introduce a weighting network *Gated* that takes $\mathbf{I}_{all}$ as input and outputs a normalized weight

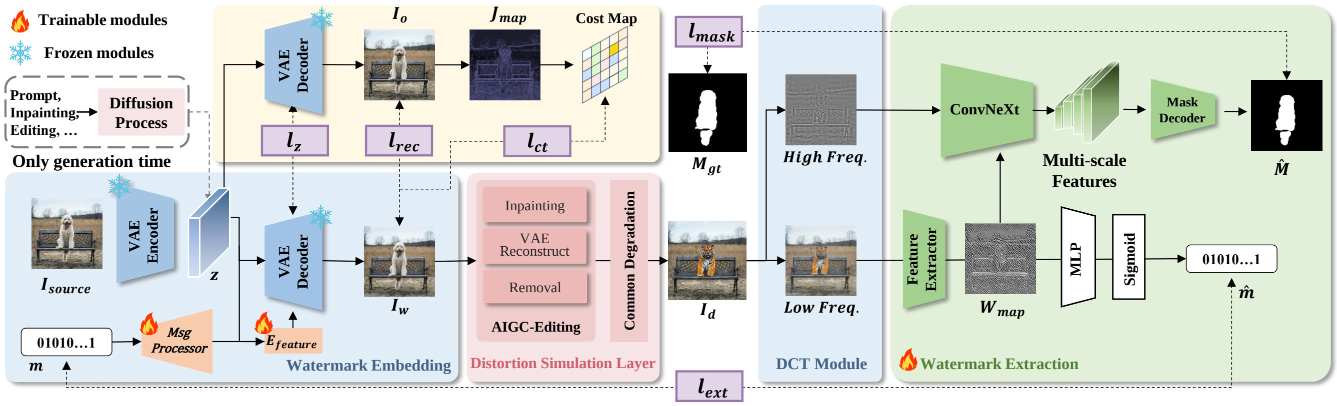

Figure 2: The Framework of Our Method.

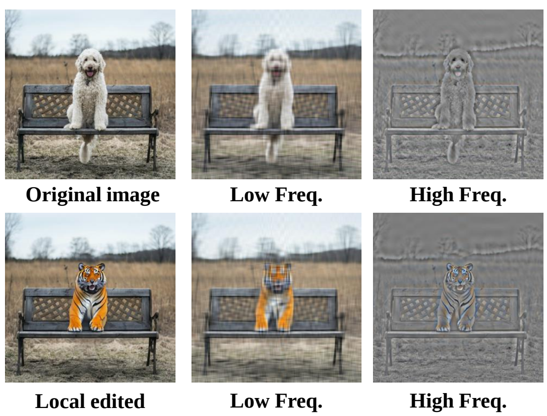

Figure 3: High- and low-frequency feature visualization before and after local editing.

tensor W, where each channel corresponds to a specific scale:

$$W = \text{Gated}(\mathbf{I}_{all}), \quad W \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 4} \tag{6}$$

The final tamper prediction is obtained by performing a weighted fusion of all scale-specific masks and resizing it to the original image size:

$$\hat{M} = \text{Resize}\left(\sum_{i=1}^{4} W_i \cdot P_{g_i}, H, W\right) \tag{7}$$

The performance of watermark extraction is measured using binary cross-entropy loss between the predicted watermark $\hat{\mathbf{m}}$ and the ground-truth message $\mathbf{m}$:

$$\ell_{\text{ext}} = \lambda_k \ell_{\text{bce}}(\hat{\mathbf{m}}, \mathbf{m}) \tag{8}$$

For tamper localization, we compute a combination of pixel-wise loss using mean squared error (MSE) and edge-aware loss [3] between the predicted mask $\hat{\mathbf{M}}$ and the ground-truth mask $\mathbf{M}_{\text{gt}}$:

$$\ell_{\text{mask}} = \lambda_m \cdot \ell_{\text{mse}}(\hat{\mathbf{M}}, \mathbf{M}_{\text{gt}}) + \gamma \cdot \ell_{\text{edge}}(\hat{\mathbf{M}}, \mathbf{M}_{\text{gt}}) \tag{9}$$

where γ is set to 20.

3.4 Distortion Layer

To improve robustness against real-world distortions, we introduce a distortion simulation layer between watermark embedding and extraction. This layer processes the watermarked image I_w and produces a degraded version I_d to simulate realistic editing conditions. It is only used during training and removed during inference.

The distortion layer includes two categories: AIGC editing and common degradations. AIGC editing covers inpainting, VAE reconstruction, and content removal, while common degradation involves typical image perturbations such as JPEG compression and brightness adjustment. During training, each image is randomly passed through one AIGC editing and one degradation operation to simulate practical distortion pipelines. Further implementation details are provided in the appendix.

AIGC-Editing Simulation. We categorize AIGC editing operations into three types, each designed to improve either tamper localization or watermark robustness under different scenarios:

1) Real inpainting editing: We adopt inpainting operations based on real diffusion models to simulate localized AIGC-style content regeneration. The editing strength is randomly sampled between 0.3 and 1.0. For samples from the UltraEdit dataset, we use the provided masks and prompts; otherwise, masks are randomly generated and prompts are set to `None`. This operation enables the model to learn tamper localization under realistic partial editing.

2) VAE reconstruction editing: This operation encodes and decodes the image using a frozen VAE from Stable Diffusion to simulate global semantic rewriting. Recent findings [60] show that watermark corruption after editing is primarily caused by VAE compression. We therefore use this strategy to enhance the model's ability to retain watermarks under global modifications.

3) Watermark-region removal: We simulate aggressive local tampering by replacing the masked watermark region with the corresponding area from the original image. This operation mimics targeted watermark removal attacks and improves the model's robustness against intentional deletion.

In summary, the inpainting and removal operations represent realistic and simulated local edits, respectively, and are used to train the model for watermark-guided tamper localization. In contrast, the VAE reconstruction serves as a global editing surrogate, ensuring that the watermark remains extractable even under significant content shifts.

3.5 Ensuring Visual Quality

Compared to single-task watermarking methods focused solely on copyright protection, our approach inevitably embeds more information, which may introduce noticeable visual artifacts. To mitigate this quality degradation, we apply constraints both during and after image generation.

First, during the decoding process, we impose multi-scale constraints on latent features to preserve spatial consistency between the clean and watermarked representations. Then, after image synthesis, we incorporate a Just-Noticeable-Difference (JND)–guided loss to control the visibility of watermark perturbations. The JND map is a hand-crafted model that estimates the minimum distortion perceivable by the human visual system at each pixel, allowing us to selectively constrain residuals where artifacts are more likely to be noticed.

Specifically, during the latent decoding process, the original decoder $\mathcal{D}$ and the modified decoder $\mathcal{D}_w$ perform simultaneous decoding at each stage $i \in \left\{1, \ldots, \frac{f}{2}\right\}$, producing the intermediate latent features z_i and z_{m_i} respectively. To ensure that the injected watermark does not significantly distort the latent representations, we apply a multi-scale MSE constraint over all decoder stages:

$$l_z = \sum_{i=1}^{f/2} \|z_i - z_{m_i}\|_2^2 \tag{10}$$

This loss encourages the preservation of spatial structure in the latent space during watermark embedding, thus mitigating visual degradation in the final output.

After image generation, we obtain two outputs: the clean image I_o and the watermarked image I_w. To minimize the perceptual visibility of watermark residuals, we introduce a JND-guided modulation strategy.

For the clean image I_o, we compute its JND map $\mathrm{JND}(I_o) \in \mathbb{R}^{3\times H\times W}$. This map is used to estimate the perceptual tolerance for pixel-level changes. We then construct a cost matrix as:

$$\text{Cost Map} = 1 - \alpha_{\text{JND}} \cdot \mathrm{JND}(I_o) \tag{11}$$

and define the JND-weighted residual loss as:

$$l_{ct} = \text{Cost Map} \odot I_w \tag{12}$$

To ensure the perceptual similarity between the watermarked image I_w and the original image I_o, we employ a combination of pixel-wise distortion and perceptual loss functions. The pixel-wise distortion is measured by MSE, defined as $l_I = \|I_w - I_o\|_2^2$. For perceptual similarity, we adopt the LPIPS loss [58], which aligns better with human perception.

$$l_{rec} = \lambda_I l_I + \lambda_{LPIPS} l_{LPIPS}(I_w, I_o) \tag{13}$$

Finally, the overall visual quality loss is defined as:

$$l_{quality} = l_{rec} + \lambda_z l_z + \lambda_{ct} l_{ct} \tag{14}$$

where λ_I, λ_{LPIPS}, λ_z, and λ_{ct} are the corresponding loss weights.

3.6 Training Details

The entire training procedure is conducted in an end-to-end manner. We initialize the loss weights as follows: $\lambda_k = 5$, $\lambda_m = 1.5$, $\lambda_I = 0.1$, $\lambda_{LPIPS} = 1$, $\lambda_z = 0.001$, and $\lambda_{ct} = 10$. To further improve the visual quality of the generated watermarked images, we adopt a dynamic loss weighting strategy. Specifically, once the extraction loss ℓ_{ext} falls below 0.05 and the tamper localization loss ℓ_{mask} is less than 0.1, we increase the emphasis on visual quality by adjusting the weights to $\lambda_I = 0.5$, $\lambda_{LPIPS} = 5$, $\lambda_z = 0.005$, and $\lambda_{ct} = 50$. During the initial 10,000 training steps, no distortion is applied. Thereafter, the distortion simulation layer is progressively introduced to enhance robustness against realistic degradations.

The architecture of the message processor W_{Pro} comprises three fully connected layers, followed by two Conv-BN-SELU blocks and a final 2D convolution layer. Each watermark embedding module W_{Emb_i} consists of a Conv-BN-SELU block and an upsampling layer. The watermark decoder W_{Dec} is built using stacked Conv-BN-SELU blocks and gated convolution modules to support structured feature decoding.

4 Experiments

4.1 Experimental Setup

Our training data consists of the MS COCO dataset [29] and 20,000 edited image pairs curated from the UltraEdit dataset [62](including original images, edited images, corresponding masks, and editing instructions). For samples from UltraEdit, the editing masks are provided, while for other datasets, masks are randomly generated using a mixed-shape strategy. All images are resized to a resolution of 512×512. The model is trained using the AdamW optimizer with an initial learning rate of 1×10^{-5} and a batch size of 2. We adopt a cosine annealing learning rate schedule. All experiments are conducted on an NVIDIA A100 GPU server.

4.2 Comparision with Localization Methods

To evaluate the tamper localization performance of our proposed *GenPTW*, we compare it against several state-of-the-art passive localization methods, including PSCC-Net [31], MVSS-Net [9], CAT-Net [25] , and IML-ViT [36], as well as the proactive watermark-based method EditGuard [59]. OmniGuard [60] is not included in the comparison as the method has not yet been publicly released. We adopt the F1-score and AUC as evaluation metrics. The evaluation is conducted on 1,000 testing images, comprising 500 samples from the publicly available AGE-Set-C dataset and 500 additional samples curated by us. Each sample consists of a manipulated image, its corresponding ground-truth mask, and the original clean image. For manipulation types, we employ advanced generative editing models, including Stable Diffusion Inpaint [41] and ControlNet Inpaint [57] with prompts set to “None”, as well as the unconditional inpainting method Lama [43]. Classical image splicing operations are also incorporated to cover non-AIGC editing scenarios. To assess robustness under real-world conditions, we randomly apply one type of common degradation to the manipulated images. The degradation types include Gaussian noise ($\sigma = 1$–10), JPEG compression (quality factor $Q = 60$–80), brightness adjustment, and contrast adjustment.

Figure 4: Qualitative examples of generated images using GenPTW.

Table 1: Localization performance of the proposed GenPTW and other SOTA proactive or passive manipulation localization methods. “Clean” and “Degraded” denote detection under the clean condition, and under the condition of randomly selecting JPEG, Gaussian noise, brightness adjustment, and contrast adjustment.

	Stable Diffusion Inpaint				Controlnet Inpaint				Splicing				Lama			
	Clean		Degraded		Clean		Degraded		Clean		Degraded		Clean		Degraded	
Method	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
MVSS-Net [9]	0.178	0.488	0.165	0.634	0.178	0.492	0.236	0.697	0.423	0.798	0.327	0.749	0.024	0.505	0.044	0.477
CAT-Net [25]	0.145	0.679	0.127	0.674	0.167	0.711	0.143	0.681	0.196	0.718	0.187	0.704	0.151	0.724	0.147	0.713
PSCC-Net [31]	0.166	0.501	0.104	0.472	0.177	0.565	0.145	0.563	0.189	0.693	0.181	0.601	0.132	0.329	0.129	0.314
IML-ViT [36]	0.213	0.596	0.217	0.604	0.200	0.576	0.204	0.578	0.473	0.754	0.452	0.747	0.105	0.456	0.111	0.442
EditGuard [59]	0.966	0.971	0.724	0.913	**0.968**	0.987	0.735	0.927	0.934	0.991	0.757	0.921	**0.965**	0.969	0.718	0.917
GenPTW (Ours)	**0.971**	**0.998**	**0.957**	**0.995**	0.963	**0.998**	**0.941**	**0.973**	**0.937**	**0.993**	**0.908**	**0.991**	0.961	**0.971**	**0.919**	**0.989**

As shown in Table 1, *GenPTW* consistently demonstrates strong localization performance across a range of manipulation tasks. Under clean conditions, it achieves F1 scores above 0.96 and AUC approaching 1.0. Even under common degradations such as JPEG compression, color jitter, and Gaussian noise, GenPTW maintains

Table 2: Fidelity and bit recovery accuracy comparison between the proposed GenPTW and other SOTA watermarking methods. Note that "SD Inpaint*" denotes the regeneration from the image via an inpainting model, while "SD Inpaint" ensures that the non-edited regions remain entirely consistent with the original image.

					Bit Accuracy (%)						
					Global Edit		Local Edit		Common Degradation		
	Method	Capacity	PSNR	SSIM	Instructp2p	SD Inpaint*	SD Inpaint	Random Dropout	JPEG	Combined	Gaussian Noise
Post	PIMoG [12]	30 bits	36.73	0.917	0.683	0.654	0.928	0.966	0.958	0.955	0.767
	SepMark [50]	30 bits	33.45	0.903	0.909	0.943	0.966	0.979	**0.987**	0.958	**0.969**
	EditGuard [59]	64 bits + $\mathbf{W}_{\text{loc}}$	36.78	0.928	0.572	0.632	0.966	**0.980**	**0.987**	0.960	0.765
	Robust-Wide [20]	64 bits	**39.18**	**0.980**	**0.976**	**0.956**	**0.997**	0.968	0.981	**0.976**	0.747
In-Gen	Stable Signature [14]	48 bits	31.43	0.834	0.561	0.626	0.805	0.864	0.921	0.914	0.803
	WOUAF [24]	64 bits	30.71	0.847	0.587	0.601	0.824	0.882	0.991	0.935	0.947
	Lawa [40]	48 bits	35.14	0.821	0.591	0.629	0.832	0.889	**0.998**	**0.953**	0.960
	GenPTW (Ours)	64 bits	**37.12**	**0.908**	**0.963**	**0.999**	**0.982**	**0.978**	0.991	0.942	**0.961**

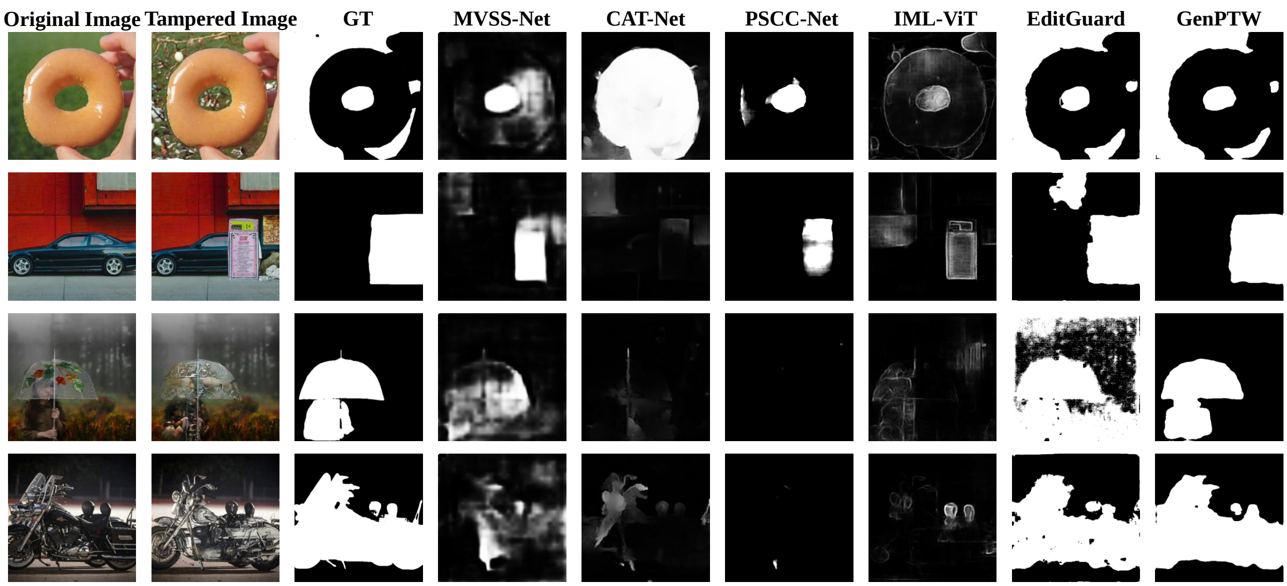

Figure 5: Visualized Comparison between our GenPTW and other methods.

high accuracy and stable performance, indicating strong robustness and generalization across tasks. Compared to existing methods, GenPTW delivers superior performance under degraded settings. For example, in the Splicing and Lama tasks, it achieves F1 scores of 0.908 and 0.919, respectively, significantly outperforming both passive detection methods and existing watermark-based approaches. In contrast, EditGuard exhibits noticeable drops in mask quality under degradation and is more sensitive to threshold settings, leading to instability in challenging conditions.

Figure 5 further compares the visual localization results across different methods. Passive methods such as PSCC-Net and IML-ViT tend to miss tampered regions under complex edits or degradations. Meanwhile, proactive methods like EditGuard often produce noisy or incomplete masks, with results highly dependent on hyperparameter tuning. In comparison, GenPTW consistently generates accurate and well-aligned masks across various types of manipulations, without requiring extensive post-processing or parameter adjustments. It is worth noting that for full-image semantic rewriting tasks such as InstructP2P, GenPTW is still able to reliably extract embedded identity and detect tampering. However, as such manipulations fundamentally alter the global content structure of the image, the model tends to classify the entire image as a tampered region. Rather than a misclassification, this reflects our design perspective—prioritizing the protection of the original visual structure over the accommodation of broad semantic transformation.

4.3 Comparison with Deep Watermarking

We comprehensively compare the performance of GenPTW with existing in-generation watermarking methods and post-generation

Table 3: Ablation study on different input combinations for W_{Dec} and CN_{Enc}.

W_{Dec}	CN_{Enc}	ACC	PSNR	SSIM	F1	AUC
Low Freq.	High Freq.+W_{map}	0.992	**37.41**	0.873	**0.970**	**0.998**
Image	Image	**0.998**	36.56	**0.879**	0.963	0.992
High Freq.	Low Freq.+W_{map}	0.953	32.22	0.801	0.952	0.980
Low Freq.	Low Freq. +W_{map}	0.989	37.36	0.866	0.908	0.974
High Freq.	High Freq. +W_{map}	0.938	31.98	0.789	0.892	0.961

Table 4: Ablation study on the effect of multi-scale loss in spatial and latent domains.

l_{ct}	l_z	ACC	PSNR	SSIM	F1	AUC
✓	✓	0.997	**37.48**	**0.876**	**0.968**	**0.998**
×	×	0.996	28.62	0.724	0.958	0.991
×	✓	0.997	33.47	0.873	0.964	0.990
✓	×	**0.999**	36.63	0.823	0.966	0.993

watermarking techniques. The in-generation methods include Stable Signature, WOUAF, and LaWa, while the post-generation baselines consist of PIMoG [12], SepMark [50], EditGuard [59], and Robust-Wide [20]. We test all the results on 1000 512×512 images with paired prompt in the dataset of UltraEdit [62]. The degradation settings are configured as follows: Gaussian noise with $\sigma = 25$, JPEG compression with quality Q = 70, and brightness perturbation with ±30% adjustment. The Combined Attack includes 40% center cropping, brightness scaling of 2.0, and JPEG compression at quality 80.

As shown in Table 2, *GenPTW* achieves the highest bit recovery accuracy under most degradation conditions, while maintaining excellent visual fidelity with a PSNR of **37.12** dB. This performance surpasses all in-generation watermarking baselines and is comparable to or even better than several post-processing watermarking techniques. Specifically, under both local and global AIGC editing, GenPTW substantially outperforms existing in-generation methods. Thanks to the joint embedding of both copyright and tamper-localizable watermarks, GenPTW improves upon EditGuard by **0.34** dB in PSNR, along with a significant boost in bit-level accuracy across all tested scenarios. In the InstructP2P full-image editing task, GenPTW achieves a bit recovery accuracy of **0.963**, only **0.013** lower than Robust-Wide, which is explicitly trained for AIGC editing scenarios. Meanwhile, GenPTW provides better trade-offs in SSIM and robustness under diverse transformations. As illustrated in Fig. 4, we visualize several samples generated using Stable Diffusion v2, followed by full-image semantic rewriting with InstructP2P. Even when the overall style and structure of the image are significantly altered, GenPTW can still accurately extract the embedded watermark. This demonstrates the strong resilience and generalization capability of our method under both global and local edits, as well as under typical real-world degradation.

4.4 Ablation Study

4.4.1 Effect of frequency-guided inputs for W_{Dec} and CN_{Enc}. To investigate the impact of input design for the watermark decoder W_{Dec} and the tamper localization encoder CN_{Enc}, we conduct an ablation study across various input combinations, as summarized in Table 2. Specifically, we explore using original images, low-frequency and high-frequency components, and an auxiliary watermark guidance map W_{map} as inputs to the two modules.

As shown in Table 3, the configuration using low-frequency input for W_{Dec} and high-frequency input combined with W_{map} for CN_{Enc} achieves the best overall performance, with a PSNR of **37.41 dB**, an SSIM of 0.873, and a near-perfect AUC of **0.998**. This setup effectively balances visual fidelity and forensic accuracy. In contrast, directly embedding the watermark into high-frequency components leads to noticeable quality degradation, with PSNR dropping to around 32 dB and SSIM significantly reduced—indicating the presence of perceptible artifacts. While these configurations may still yield competitive detection metrics, they suffer from compromised perceptual quality. Using the original image as input preserves fidelity and achieves high SSIM, but lacks explicit frequency-level guidance and underperforms in terms of overall consistency compared to our proposed design.

4.4.2 Effect of multi-scale loss in spatial and latent domains. We conduct an ablation study to investigate the impact of incorporating loss terms in the spatial and latent domains. Specifically, we analyze the contribution of the contrastive texture-aware loss l_{ct}, designed based on the JND, and the latent consistency loss l_z computed over the multi-scale latent features.

As shown in Table 4, introducing l_z alone leads to a notable improvement in SSIM (from 0.724 to 0.873), suggesting that encouraging consistency in the latent space substantially enhances perceptual similarity. Meanwhile, incorporating l_{ct} leads to overall gains in both PSNR and SSIM, indicating its effectiveness in guiding spatial fidelity preservation under visually sensitive regions. When both loss terms are applied jointly, the model achieves the best trade-off across all metrics, with PSNR reaching 37.48 and SSIM improving to 0.876. These results validate the complementary benefits of combining spatial and latent-domain supervision, and highlight the importance of perceptual-aware regularization for high-fidelity watermark recovery.

5 Conclusion

In this paper, we propose *GenPTW*, a unified in-generation framework for proactive provenance tracing and tamper localization. To the best of our knowledge, it is the first in-generation image watermarking solution that simultaneously supports both provenance tracing and tamper localization. To improve extraction accuracy, we design a frequency-coordinated decoder that disentangles low-frequency watermark recovery from high-frequency tamper detection. To enhance robustness against AIGC editing and common degradations, we introduce a distortion simulation layer that mimics realistic generative manipulations. Additionally, to preserve visual quality, we incorporate a JND-constrained perceptual loss guided by a pixel-wise modification cost map. Extensive experiments demonstrate that GenPTW consistently outperforms existing watermarking and forensic baselines in terms of fidelity, localization precision, and robustness under diverse tampering scenarios.

References

[1] Vishal Asnani, John Collomosse, Tu Bui, Xiaoming Liu, and Shruti Agarwal. 2024. ProMark: Proactive diffusion watermarking for causal attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10802–10811.

[2] Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. 2023. Malp: Manipulation localization using a proactive scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Tu Bui, Shruti Agarwal, and John Collomosse. 2023. TrustMark: Universal Watermarking for Arbitrary Resolution Images. *arXiv preprint arXiv:2311.18297* (2023).

[4] Tu Bui, Shruti Agarwal, Ning Yu, and John Collomosse. 2023. RoSteALS: Robust Steganography using Autoencoder Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 933–942.

[5] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. 2021. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[6] Baotian Cheng, Rongrong Ni, and Yao Zhao. 2012. A refining localization watermarking for image tamper detection and recovery. In *2012 IEEE 11th International Conference on Signal Processing*, Vol. 2. 984–988.

[7] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. 2015. Efficient dense-field copy–move forgery detection. *IEEE Transactions on Information Forensics and Security* 10, 11 (2015), 2284–2297.

[8] Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, and Jiliang Tang. 2023. DiffusionShield: A Watermark for Copyright Protection against Generative Diffusion Models. *arXiv preprint arXiv:2306.04642* (2023).

[9] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. 2022. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 3539–3553.

[10] Yu Fan, Philippe Carré, and Christine Fernandez-Maloigne. 2015. Image splicing detection with local illumination estimation. In *2015 IEEE international conference on Image processing (ICIP)*. IEEE, 2940–2944.

[11] Han Fang, Kejiang Chen, Yupeng Qiu, Jiayang Liu, Ke Xu, Chengfang Fang, Weiming Zhang, and Ee-Chien Chang. 2023. Denol: a few-shot-sample-based decoupling noise layer for cross-channel watermarking robustness. In *Proceedings of the 31st ACM international conference on multimedia*. 7345–7353.

[12] Han Fang, Zhaoyang Jia, Zehua Ma, Ee-Chien Chang, and Weiming Zhang. 2022. PIMoG: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network. In *Proceedings of the 30th ACM International Conference on Multimedia (MM)*.

[13] Han Fang, Zhaoyang Jia, Yupeng Qiu, Jiyi Zhang, Weiming Zhang, and Ee-Chien Chang. 2022. De-END: Decoder-driven watermarking network. *IEEE Transactions on Multimedia* 25 (2022), 7571–7581.

[14] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. 2023. The stable signature: Rooting watermarks in latent diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2023).

[15] Rafael C. Gonzalez and Richard E. Woods. 2018. *Digital Image Processing* (4th ed.). Pearson. 262 pages.

[16] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. 2023. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[17] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. 2023. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Jong Goo Han, Tae Hee Park, Yong Ho Moon, and Il Kyu Eom. 2016. Efficient Markov feature extraction method for image splicing detection using maximization and threshold expansion. *Journal of Electronic Imaging* 25, 2 (2016), 023031–023031.

[19] Runyi Hu, Jie Zhang, Ting Xu, Jiwei Li, and Tianwei Zhang. 2024. Robust-wide: Robust watermarking against instruction-driven image editing. In *European Conference on Computer Vision*. Springer, 20–37.

[20] Runyi Hu, Jie Zhang, Ting Xu, Jiwei Li, and Tianwei Zhang. 2025. Robust-wide: Robust watermarking against instruction-driven image editing. In *European Conference on Computer Vision*. Springer, 20–37.

[21] Nasir N Hurrah, Shabir A Parah, Nazir A Loan, Javaid A Sheikh, Mohammad Elhoseny, and Khan Muhammad. 2019. Dual watermarking framework for privacy protection and content authentication of multimedia. *Future generation computer Systems* 94 (2019), 654–673.

[22] Ashraful Islam, Chengjiang Long, Arslan Basharat, and Anthony Hoogs. 2020. Doa-gan: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[23] Zhaoyang Jia, Han Fang, and Weiming Zhang. 2021. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In *Proceedings of the 29th ACM International Conference on Multimedia (MM)*.

[24] Changhoon Kim, Kyle Min, Maitreya Patel, Sheng Cheng, and Yezhou Yang. 2024. Wouaf: Weight modulation for user attribution and fingerprinting in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8974–8983.

[25] Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. 2021. CAT-Net: Compression artifact tracing network for detection and localization of image splicing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

[26] Ce Li, Qiang Ma, Limei Xiao, Ming Li, and Aihua Zhang. 2017. Image splicing detection based on Markov features in QDCT domain. *Neurocomputing* 228 (2017), 29–36.

[27] Siau-Chuin Liew, Siau-Way Liew, and Jasni Mohd Zain. 2013. Tamper localization and lossless recovery watermarking scheme with ROI segmentation and multilevel authentication. *Journal of digital imaging* 26 (2013), 316–325.

[28] Chia-Chen Lin, Ting-Lin Lee, Ya-Fen Chang, Pei-Feng Shiu, and Bohan Zhang. 2023. Fragile Watermarking for Tamper Localization and Self-Recovery Based on AMBTC and VQ. *Electronics* 12, 2 (2023), 415.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[30] Shao-Hui Liu, Hong-Xun Yao, Wen Gao, and Yong-Liang Liu. 2007. An image fragile watermark scheme based on chaotic image pattern and pixel-pairs. *Appl. Math. Comput.* 185, 2 (2007), 869–882.

[31] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. 2022. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 11 (2022), 7505–7517.

[32] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11976–11986.

[33] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. 2021. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16317–16326.

[34] Siwei Lyu, Xunyu Pan, and Xing Zhang. 2014. Exposing region splicing forgeries with blind local noise estimation. *International journal of computer vision* 110 (2014), 202–221.

[35] Rui Ma, Mengxi Guo, Yi Hou, Fan Yang, Yuan Li, Huizhu Jia, and Xiaodong Xie. 2022. Towards Blind Watermarking: Combining Invertible and Non-invertible Mechanisms. In *Proceedings of the ACM International Conference on Multimedia (MM)*.

[36] Xiaochen Ma, Bo Du, Xianggen Liu, Ahmed Y Al Hammadi, and Jizhe Zhou. 2023. IML-ViT: Image Manipulation Localization by Vision Transformer. *arXiv preprint arXiv:2307.14863* (2023).

[37] KA Navas, Mathews Cheriyan Ajay, M Lekshmi, Tampy S Archana, and M Sasikumar. 2008. Dwt-dct-svd based watermarking. In *2008 3rd International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE'08)*. IEEE, 271–274.

[38] Neena Raj NR and R Shreelekshmi. 2022. Fragile watermarking scheme for tamper localization in images using logistic map and singular value decomposition. *Journal of Visual Communication and Image Representation* 85 (2022), 103500.

[39] Mehul S Raval and Priti P Rege. 2003. Discrete wavelet transform based multiple watermarking scheme. In *TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region*, Vol. 3. IEEE, 935–938.

[40] Ahmad Rezaei, Mohammad Akbari, Saeed Ranjbar Alvar, Arezou Fatemi, and Yong Zhang. 2024. Lawa: Using latent space for in-generation image watermarking. In *European Conference on Computer Vision*. Springer, 118–136.

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[42] Jee-Young Sun, Seung-Wook Kim, Sang-Won Lee, and Sung-Jea Ko. 2018. A novel contrast enhancement forensics based on convolutional neural networks. *Signal Processing: Image Communication* 63 (2018), 149–160.

[43] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

[44] Matthew Tancik, Ben Mildenhall, and Ren Ng. 2020. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2117–2126.

[45] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. 2022. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition. 2364–2373.

[46] Run Wang, Felix Juefei-Xu, Meng Luo, Yang Liu, and Lina Wang. 2021. Faketagger: Robust safeguards against deepfake dissemination via provenance tracking. In *Proceedings of the 29th ACM international conference on multimedia*. 3546–3555.

[47] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. 2023. Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust. *arXiv preprint arXiv:2305.20030* (2023).

[48] Eric Wengrowski and Kristin Dana. 2019. Light field messaging with deep photographic steganography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1515–1524.

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

[50] Xiaoshuai Wu, Xin Liao, and Bo Ou. 2023. SepMark: Deep Separable Watermarking for Unified Source Tracing and Deepfake Detection. In *Proceedings of the ACM international conference on Multimedia (MM)*.

[51] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. 2018. Image copy-move forgery detection via an end-to-end deep neural network. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1907–1915.

[52] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. 2019. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[53] Yixin Wu, Feiran Zhang, Tianyuan Shi, Ruicheng Yin, Zhenghua Wang, Zhenliang Gan, Xiaohua Wang, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. 2025. Explainable Synthetic Image Detection through Diffusion Timestep Ensembling. *arXiv preprint arXiv:2503.06201* (2025).

[54] Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. 2024. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12162–12171.

[55] Zeqin Yu, Jiangqun Ni, Yuzhen Lin, Haoyi Deng, and Bin Li. 2024. DiffForensics: Leveraging Diffusion Prior to Image Forgery Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12765–12774.

[56] Zihan Yuan, Qingtang Su, Decheng Liu, and Xueting Zhang. 2021. A blind image watermarking scheme combining spatial domain and frequency domain. *The visual computer* 37 (2021), 1867–1881.

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

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[59] Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. 2024. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11964–11974.

[60] Xuanyu Zhang, Zecheng Tang, Zhipei Xu, Runyi Li, Youmin Xu, Bin Chen, Feng Gao, and Jian Zhang. 2024. OmniGuard: Hybrid Manipulation Localization via Augmented Versatile Deep Image Watermarking. *arXiv preprint arXiv:2412.01615* (2024).

[61] Xuanyu Zhang, Youmin Xu, Runyi Li, Jiwen Yu, Weiqi Li, Zhipei Xu, and Jian Zhang. 2024. V2a-mark: Versatile deep visual-audio watermarking for manipulation localization and copyright protection. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 9818–9827.

[62] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. 2024. UltraEdit: Instruction-based Fine-Grained Image Editing at Scale. arXiv:2407.05282 [cs.CV] https://arxiv.org/abs/2407.05282

[63] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. 2023. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137* (2023).

[64] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. 2018. Hidden: Hiding data with deep networks. In *European Conference on Computer Vision (ECCV)*.

[65] Xinshan Zhu, Yongjun Qian, Xianfeng Zhao, Biao Sun, and Ya Sun. 2018. A deep learning approach to patch-based image inpainting forensics. *Signal Processing: Image Communication* 67 (2018), 90–99.

[66] Peiyu Zhuang, Haodong Li, Shunquan Tan, Bin Li, and Jiwu Huang. 2021. Image tampering localization using a dense fully convolutional network. *IEEE Transactions on Information Forensics and Security* 16 (2021), 2986–2999.