This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

EditGuard: Versatile Image Watermarking for Tamper Localization and Copyright Protection

Xuanyu Zhang[1,2], Runyi Li[1], Jiwen Yu[1], Youmin Xu[1], Weiqi Li[1], Jian Zhang[1,2] ✉
[1] School of Electronic and Computer Engineering, Peking University
[2] Peking University Shenzhen Graduate School-Rabbitpre AIGC Joint Research Laboratory

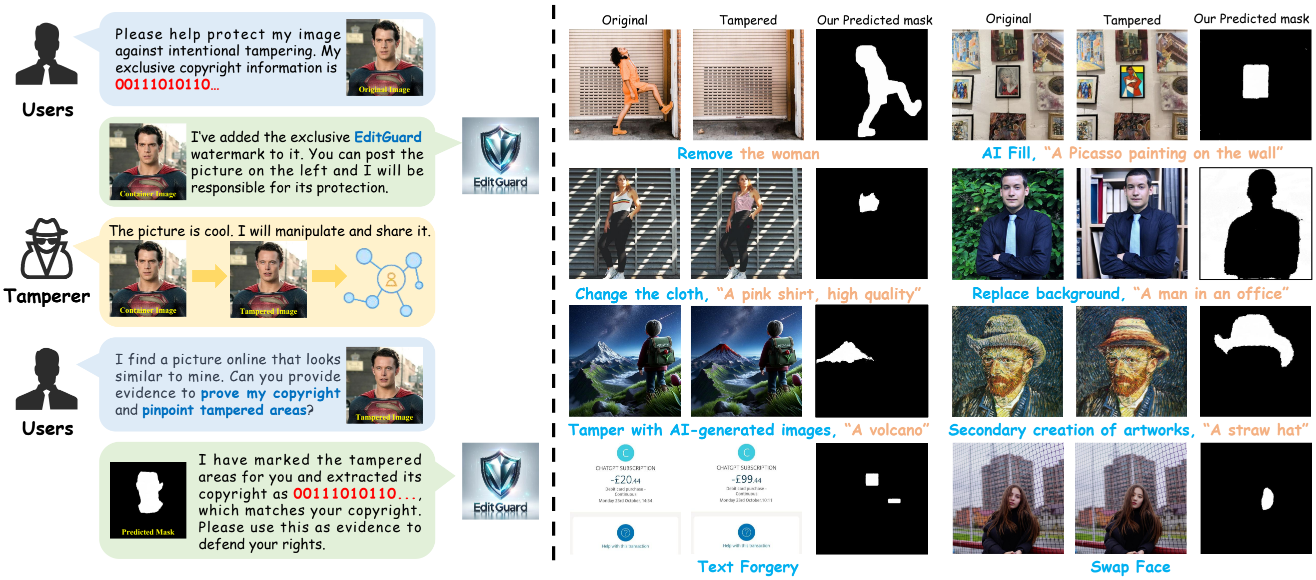

Figure 1. We propose a versatile proactive forensics framework **EditGuard**. The application scenario is shown on the left, wherein users embed invisible watermarks to their images via EditGuard in advance. If suffering tampering, users can defend their rights via the tampered areas and copyright information provided by EditGuard. Some supported tampering methods (marked in blue) and localization results of EditGuard are placed on the right. Our EditGuard can achieve over 95% localization precision and nearly 100% copyright accuracy.

Abstract

In the era of AI-generated content (AIGC), malicious tampering poses imminent threats to copyright integrity and information security. Current deep image watermarking, while widely accepted for safeguarding visual content, can only protect copyright and ensure traceability. They fall short in localizing increasingly realistic image tampering, potentially leading to trust crises, privacy violations, and legal disputes. To solve this challenge, we propose an innovative proactive forensics framework ***EditGuard****, to unify copyright protection and tamper-agnostic localization, especially for AIGC-based editing methods. It can offer a meticulous embedding of imperceptible watermarks and precise decoding of tampered areas and copyright information. Leveraging our observed fragility and locality of image-into-image steganography, the realization of EditGuard can be converted into a united image-bit steganography issue, thus completely decoupling the training process from the tampering types. Extensive experiments verify that our EditGuard balances the tamper localization accuracy, copyright recovery precision, and generalizability to various AIGC-based tampering methods, especially for image forgery that is difficult for the naked eye to detect.*

This work was supported by National Natural Science Foundation of China under Grant 62372016. (✉ Corresponding author: Jian Zhang.)

1. Introduction

Owing to the advantageous properties of diffusion models and the bolstering of extensive datasets, AI-generated content (AIGC) models like DALL·E 3 [14], Imagen [52], and Stable Diffusion [51], can produce lifelike and wondrous images, which brings great convenience to photography enthusiasts and image editors. Nonetheless, the remarkable capabilities of these models come with a double-edged sword, presenting new challenges in copyright protection and information security. The efficiency of image manipulation [43, 47, 49, 51, 55, 68, 72] has blurred the line between fact and forgery, ushering in myriad security and legal concerns. For instance, artistic works are vulnerable to malicious tampering or unauthorized AI-facilitated recreations, making it challenging to protect their original creations [13]. Meanwhile, forged images may be spread online or used as court evidence, causing adverse effects on public opinion, ethical issues, and social stability.

Given the challenges of preventing image tampering from the source, image watermarking has become a consensus for proactive forensics [54]. However, most prevalent forensic image watermarking [12, 15, 44, 62] still focus on detecting image authenticity or protecting image copyrights, but fall short when it comes to advanced demands, such as localizing tampered areas. Tamper localization facilitates an evaluation of the severity of the image manipulation, and provides an understanding of the intent of attackers, potentially allowing for the partial reuse of the tampered images. However, most passive forensics methods such as previous black-box localization networks [8, 56, 61] tend to seek anomalies like artifacts or flickers in images but struggle to detect more realistic textures and more advanced AIGC models. Moreover, they inevitably need to introduce tampered data during the training and focus solely on specific "CheapFake" tampering like slicing and copy-and-paste [8, 45, 56], or on "DeepFake" targeting human faces [2, 50], restricted in generalizability. Thus, it is vital to develop an integrated watermarking framework that unites tamper-agnostic localization and copyright protection.

To clarify our task scope, we re-emphasize the definition of dual forensics tasks as illustrated in Fig. 1: **(1) Copyright protection:** "Who does this image belong to?" We aspire to accurately retrieve the original copyright of an image, even suffering various tampering and degradation. **(2) Tamper localization:** "Where was this image manipulated?" We aim to precisely pinpoint the tampered areas, unrestrained by specific tampering types. To the best of our knowledge, no existing method accomplishes these two tasks simultaneously, while maintaining a balance of high precision and extensive generalizability.

To address this urgent demand, we propose a novel proactive forensics framework, dubbed **EditGuard**, to protect copyrights and localize tamper areas for AIGC-based editing methods. Specifically, drawing inspiration from our observed locality and fragility of image-into-image (I2I) steganography and inherent robustness of bit-into-image steganography, we can transform the realization of EditGuard into a joint image-bit steganography issue, which allows the training of EditGuard to be entirely decoupled from tampering types, thereby endowing it with exceptional generalizability and locate tampering in a zero-shot manner. In a nutshell, our contributions are as follows:

❑ (1) We present the first attempt to design a deep versatile proactive forensics framework **EditGuard** for universal tamper localization and copyright protection. It embeds dual invisible watermarks into original images and accurately decodes tampered areas and copyright information.

❑ (2) We have observed the fragility and locality of I2I steganography and innovatively convert the solution of this dual forensics task into training a united Image-Bit Steganography Network (IBSN), and utilize the core components of IBSN to construct EditGuard.

❑ (3) We introduce a prompt-based posterior estimation module to enhance the localization accuracy and degradation robustness of the proposed framework.

❑ (4) The effectiveness of our method has been verified on our constructed dataset and classical benchmarks. Compared to other competitive methods, our approach has notable merits in localization precision, generalization abilities, and copyright accuracy without any labeled data or additional training required for specific tampering types.

2. Related works

2.1. Tamper Localization

Prevalent passive image forensic techniques have focused on localizing specific types of manipulations [25, 33, 34, 53, 61, 75]. Meanwhile, some universal tamper localization methods [5, 23, 31, 35, 63, 65–67] also tend to explore artifacts and anomalies in tampered images. For instance, MVSS-Net [8] employed multi-view feature learning and multi-scale supervision to jointly exploit boundary artifacts and the noise view of images. OSN [61] proposed a novel robust training scheme to address the challenges posed by lossy operations. Trufor [17] used a learned noise-sensitive fingerprint and extracted both high-level and low-level traces via transformer-based fusion. HiFi-Net [18] utilized multi-branch feature extractor and localization modules for both CNN-synthesized and edited images. SAFL-Net [56] constrained a feature extractor to learn semantic-agnostic features with specific modules and auxiliary tasks. However, the above-mentioned passive localization methods are often limited in terms of generalization and localization accuracy, which usually work on known tampering types that have been trained. Although MaLP [2] used template matching for proactive tamper localization, it still requires a large number of forgery images and cannot fully decouple the network training from the tamper types.

2.2. Image Watermarking

Image watermarking can be broadly used for the verification, authenticity, and traceability of images. Although traditional fragile watermarking [6, 24, 29, 36, 37, 48] can also achieve block-wise tamper localization, their localization accuracy and flexibility are unsatisfactory. How to realize joint pixel-level tamper localization and copyright protection remains largely unexplored. Owing to the development

Figure 2. Fragility and locality of I2I steganography. The first line shows that when $\mathbf{X}'_{con}$ is changed, $\hat{\mathbf{X}}_{sec}$ will also be fragilely demaged. The second line plots the attribution maps $\frac{1}{|\mathbb{S}|}\sum_{(i,j)\in\mathbb{S}}\frac{\partial\hat{\mathbf{X}}_{sec[i,j]}}{\partial\mathbf{X}'_{con}}$ of five point sets $\mathbb{S}$ (marked by ①-⑤) in $\hat{\mathbf{X}}_{sec}$. We observed that $\hat{\mathbf{X}}_{sec}$ almost only has a strong response at the corresponding positions of $\mathbf{X}'_{con}$ and its neighborhoods.

of deep learning, deep image watermarking has attracted increased attention. For instance, HiDDeN [74] firstly introduced a deep encoder-decoder network to hide and recover bitstream. Moreover, many distortion layers such as differentiable JPEG and screen-shooting [1, 10, 40, 62] were designed to enhance its robustness. Flow-based models [11, 44] were also utilized to further improve the fidelity of container images. Recently, researchers [7, 12, 27, 59, 73] have designed specialized watermarking mechanisms for large-scale image generation models [51], to merge watermarking into the generation process. However, these deep watermarking methods have a singular function and cannot accurately localize the tampered areas.

3. Overall Framework of EditGuard

3.1. Motivation

Challenges of existing methods: (1) How to equip existing watermark methods, which are solely for copyright protection, with the ability to localize tampering is the crux of EditGuard. We will solve it via the framework design in Sec. 3.2. (2) Most previous tamper localization methods inevitably introduce specific tampering data during network training but tend to raise generalization concerns in unknown tampering types, which will be addressed in Sec. 3.3.

Our observations: Fortunately, we observed that image-into-image (I2I) steganography exhibits distinct fragility and locality, possessing great potential to address these issues. Concretely, I2I steganography [3, 28, 42, 46, 64, 69] aims to hide a secret image $\mathbf{X}_{sec}$ into a cover image $\mathbf{X}_{cov}$ to produce a container image $\mathbf{X}_{con}$, and reveal $\hat{\mathbf{X}}_{sec}$ and $\hat{\mathbf{X}}_{cov}$ with minimal distortion from the received image $\mathbf{X}'_{con}$. We discover that when $\mathbf{X}'_{con}$ undergoes significant alteration compared to $\mathbf{X}_{con}$, $\hat{\mathbf{X}}_{sec}$ will also be damaged and generate artifacts (the first row of Fig. 2), which is called ***fragility***. Furthermore, we notice that the artifacts in $\hat{\mathbf{X}}_{sec}$ are almost pixel-level corresponding to the changes in $\mathbf{X}'_{con}$ relative to $\mathbf{X}_{con}$, which is called ***locality***. To demonstrate this locality, we select five 7×7 point sets on $\hat{\mathbf{X}}_{sec}$ and calculated their attribution maps with respect to $\mathbf{X}'_{con}$. As plotted in the second row of Fig. 2, $\hat{\mathbf{X}}_{sec}$ only exhibits a strong response at the corresponding locations of $\mathbf{X}'_{con}$ and their immediate vicinity, almost irrelevant to other pixels. These properties inspire us to treat $\mathbf{X}_{sec}$ as a special localization watermark and embed it within existing watermarking frameworks.

3.2. Framework Design and Forensics Process

To realize united tamper localization and copyright protection, EditGuard is envisioned to embed both a 2D localization watermark and a 1D copyright traceability watermark into the original image in an imperceptible manner, which allows the decoding end to obtain the copyright of the images and a binary mask reflecting tampered areas. However, designing such a framework needs to solve the compatibility issue of two types of watermarks.

(1) Local vs. Global: The localization watermark is required to be hidden in the corresponding pixel positions of the original image, while the copyright watermark should be unrelated to spatial location and embedded in the global area redundantly. **(2) Semi-fragile vs. Robust**: The desired attribute of the localization watermark is semi-fragile, which means it is fragile to tampering but robust against some common degradations (such as Gaussian noise, JPEG compression, and Poisson noise) during network transmission. However, the copyright should be extracted nearly losslessly, irrespective of tampering or degradation.

To address the two pivotal conflicts, EditGuard employs a **"sequential encoding and parallel decoding"** structure, which comprises a dual-watermark encoder, a tamper locator, and a copyright extractor. As shown in Fig. 3, the dual-watermark encoder will sequentially add pre-defined localization watermark and global copyright watermark $\mathbf{w}_{cop}$ provided by users to the original image $\mathbf{I}_{ori}$, forming the container image $\mathbf{I}_{con}$. Our experiments have proved that parallel encoding cannot effectively add dual watermarks into images (in supplementary materials **(S.M.)**). In contrast, sequential embedding effectively prevents cross-interference by hiding these two watermarks. Furthermore, we model the network transmission process in which the received (tampered) image $\mathbf{I}_{rec}$ is transformed from $\mathbf{I}_{con}$ as:

$$\mathbf{I}_{rec} = \mathcal{D}(\mathbf{I}_{con} \odot (\mathbf{1} - \mathbf{M}) + \mathcal{T}(\mathbf{I}_{con}) \odot \mathbf{M}), \quad (1)$$

where $\mathcal{T}(\cdot)$, $\mathcal{D}(\cdot)$ and $\mathbf{M}$ respectively denote the tamper function, degradation operation, and tempered mask. Moreover, the parallel decoding processes allow us to flexibly train each branch under different levels of robustness and obtain the predicted mask $\hat{\mathbf{M}}$ via the tamper locator and traceability watermark $\hat{\mathbf{w}}_{cop}$ via the copyright extractor. We can categorize the dual forensic process of EditGuard into the following scenarios.

❑ **Case 1:** If $\hat{\mathbf{w}}_{cop} \not\approx \mathbf{w}_{cop}$, suspicious $\mathbf{I}_{rec}$ is either not

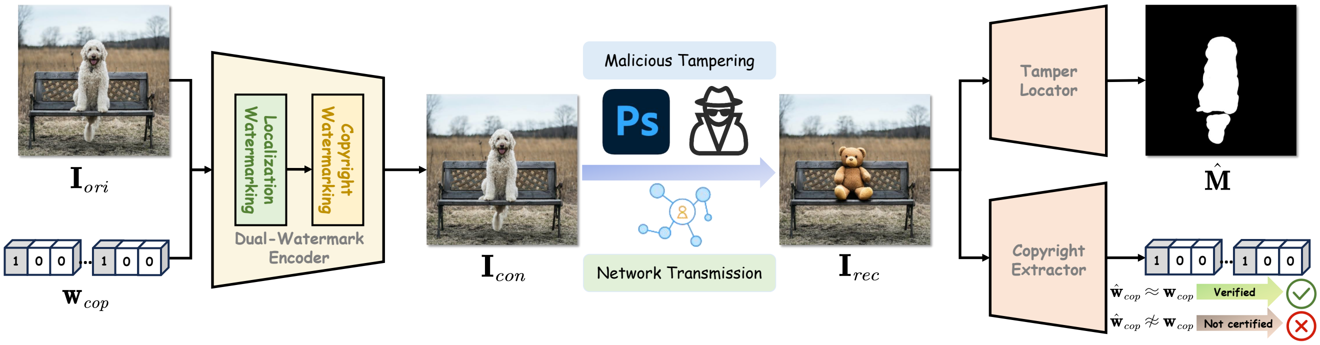

Figure 3. Illustration of the proposed proactive forensics framework EditGuard. The dual-watermark encoder sequentially embeds the pre-defined localization watermark and copyright watermark $\mathbf{w}_{cop}$ into the original image $\mathbf{I}_{ori}$, generating the container image $\mathbf{I}_{con}$. After encountering potential malicious tampering and degradation during network transmission, tampered mask $\hat{\mathbf{M}}$ and copyright information $\hat{\mathbf{w}}_{cop}$ are respectively extracted via the tamper locator and copyright extractor from the received image $\mathbf{I}_{rec}$.

registered in our EditGuard or underwent extremely severe global tampering, rendering it unreliable as evidence.

❑ **Case 2:** If $\hat{\mathbf{w}}_{cop} \approx \mathbf{w}_{cop}$ and $\hat{\mathbf{M}} \not\approx \mathbf{0}$, suspicious $\mathbf{I}_{rec}$ has undergone tampering, disqualifying it as valid evidence. Users may infer the intention of tamperers based on $\hat{\mathbf{M}}$ and decide whether to reuse parts of the image.

❑ **Case 3:** If $\hat{\mathbf{w}}_{cop} \approx \mathbf{w}_{cop}$ and $\hat{\mathbf{M}} \approx \mathbf{0}$, $\mathbf{I}_{rec}$ remains untampered and trustworthy under the shield of EditGuard.

3.3. Transform Dual Forensics into Steganography

To realize universal and tamper-agnostic localization, we resort to our observed locality and fragility of I2I steganography. As described in Sec. 3.1, localization watermarking and tamper locator in Fig. 3 can be effectively realized via image hiding and revealing. Meanwhile, combing with the robustness of current bit-into-image steganography, copyright watermarking and extractor in Fig. 3 are achieved via bit encryption and recovery. Thus, we can convert the realization of the dual forensics framework EditGuard into a united image-bit steganography network. **Our training objective is just a self-recovery mechanism, meaning it only needs to ensure the input and output of the steganography network maintain high fidelity under various robustness levels, with no need to introduce any labeled data or tampered samples.** During inference, it can naturally locate tampering via simple comparisons **in a zero-shot manner** and extract copyright exactly.

4. United Image-bit Steganography Network

4.1. Network Architecture

As plotted in Fig. 4, the proposed IBSN includes an image hiding module (IHM), a bit encryption module (BEM), a bit recovery module (BRM), and an image revealing module (IRM). First, the IHM aims to hide a localization watermark $\mathbf{W}_{loc}{\in}\mathbb{R}^{H\times W\times 3}$ into the original image $\mathbf{I}_{ori}{\in}\mathbb{R}^{H\times W\times 3}$, resulting in an intermediate output $\mathbf{I}_{med}{\in}\mathbb{R}^{H\times W\times 3}$. Subsequently, $\mathbf{I}_{med}$ is fed to the BEM for feature refinement, while the copyright watermark $\mathbf{w}_{cop}{\in}\{0,1\}^L$ is modulated into the BEM, forming the final container image $\mathbf{I}_{con}{\in}\mathbb{R}^{H\times W\times 3}$. After network transmission, the BRM will faithfully reconstruct the copyright watermark $\hat{\mathbf{w}}_{cop}$ from the received container image $\mathbf{I}_{rec}$. Meanwhile, $\mathbf{I}_{rec}$ predicts the missing information $\hat{\mathbf{Z}}$ via the prompt-based posterior estimation and uses it as the initialization for the invertible blocks, producing $\hat{\mathbf{I}}_{ori}$ and semi-fragile watermark $\hat{\mathbf{W}}_{loc}$.

4.2. Invertible Blocks in IHM and IRM

Given the inherent capacity of flow-based models to precisely recover multimedia information, we harness stacked invertible blocks to construct image hiding and revealing modules. The original image $\mathbf{I}_{ori}{\in}\mathbb{R}^{H\times W\times 3}$ and localization watermark $\mathbf{W}_{loc}{\in}\mathbb{R}^{H\times W\times 3}$ will undergo discrete wavelet transformations (DWT) to yield frequency-decoupled image features. We then employ enhanced additive affine coupling layers to project the original image and its corresponding localization watermark branches. The transformation parameters are generated from each other. The enhanced affine coupling layer is composed of a five-layer dense convolution block [46] and a lightweight feature interaction module (LFIM) [4]. The LFIM can enhance the non-linearity of transformations and capture the long-range dependencies with low computational cost. More details are presented in **S.M.**. Finally, the revealed features are then transformed to the image domain via the inverse DWT.

4.3. Prompt-based Posterior Estimation

To bolster the fidelity and robustness of the image hiding and revealing module, we introduce a degradation prompt-based posteriori estimation module (PPEM). Since the encoding network tends to compress $[\mathbf{I}_{ori};\mathbf{W}_{loc}]{\in}\mathbb{R}^{H\times W\times 6}$ into the container image $\mathbf{I}_{con}{\in}\mathbb{R}^{H\times W\times 3}$, previous methods [42, 64] typically utilized a random Gaussian initialization or an all-zero map at the decoding end to compensate for the lost high-frequency channels. Yet, our observations suggest that discarded information lurks within the

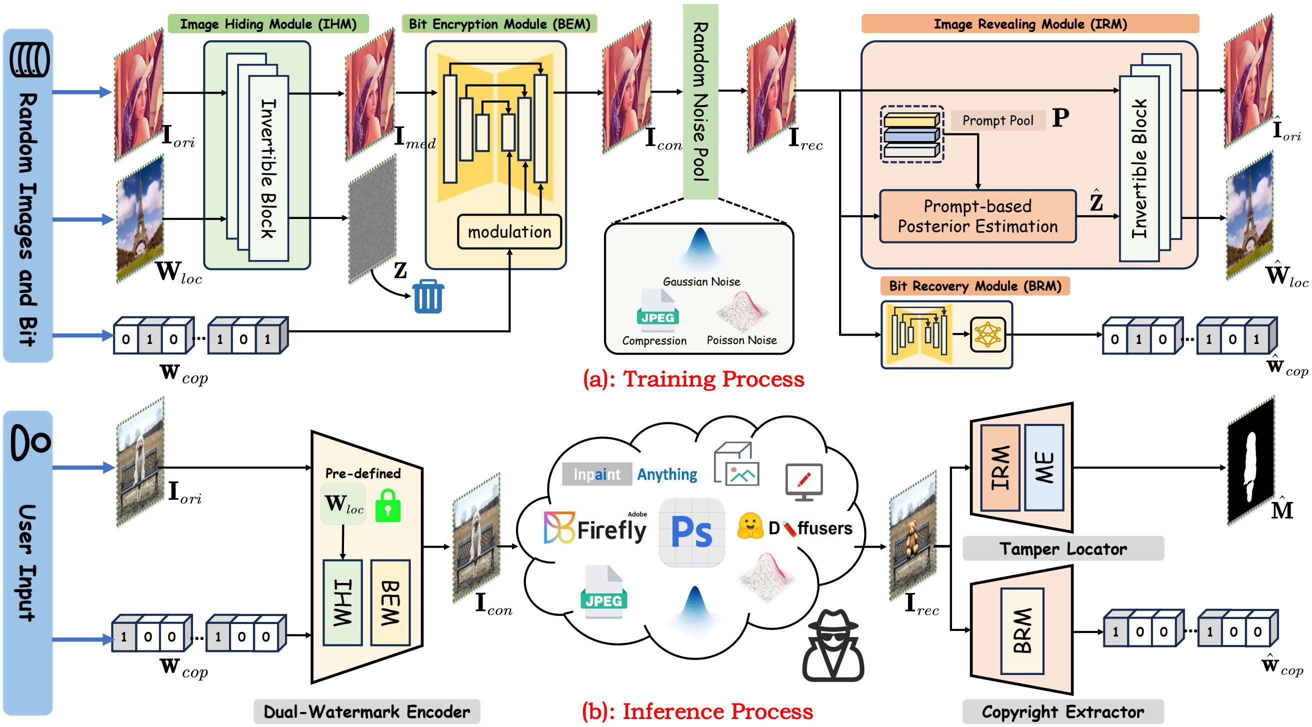

Figure 4. Illustration of the united Image-bit Steganography Network (IBSN). In the training process, we randomly sample original image $\mathbf{I}_{ori}$, localization watermark $\mathbf{W}_{loc}$ (a natural RGB image) and copyright watermark $\mathbf{w}_{cop}$ and expect the IBSN to recover $\hat{\mathbf{I}}_{ori}$, $\hat{\mathbf{W}}_{loc}$ and $\hat{\mathbf{w}}_{cop}$ with high fidelity. In the inference process, we use core components of the pre-trained IBSN with a mask extractor (ME) to construct our EditGuard, and pre-define a simple solid color image as a localization watermark.

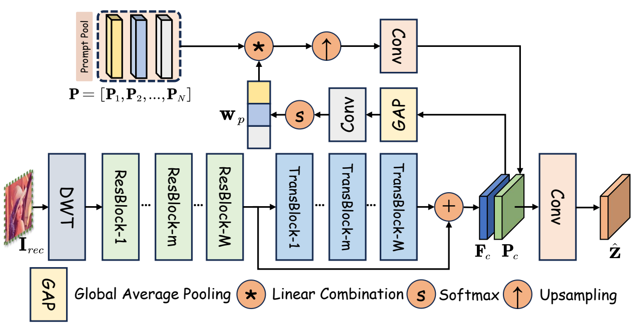

Figure 5. Illustration of the proposed prompt-based posterior estimation. It will dynamically fuse degraded representations and extracted features to obtain posterior mean $\hat{\mathbf{Z}}=\mathbb{E}[\mathbf{Z}|\mathbf{I}_{rec}]$.

edges and textures of the container image. Thus, deploying a dedicated network proves to be a more potent strategy in predicting the posterior mean of the vanished localization watermark information $\hat{\mathbf{Z}}=\mathbb{E}[\mathbf{Z}|\mathbf{I}_{rec}]$. Specifically, as shown in Fig. 5, we stack M residual blocks $\mathrm{Res}(\cdot)$ [19] and M channel-wise transformer blocks $\mathrm{Trans}(\cdot)$ [70] to extract the local and non-local features $\mathbf{F}_c$.

$$\mathbf{F}_c = \mathrm{Trans}(\mathrm{Res}(\mathrm{DWT}(\mathbf{I}_{rec}))) + \mathrm{Res}(\mathrm{DWT}(\mathbf{I}_{rec})). \quad (2)$$

Considering that the container image is prone to various degradations during network transmission, we pre-define N learnable embedding tensors as degradation prompts $\mathbf{P} = [\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N]$, where N denotes the number of degradation types and is set to 3. These learned prompts $\mathbf{P}$ can adaptively learn a diverse range of degradation representations and are integrated with the intrinsic features extracted from $\mathbf{I}_{rec}$, enabling the proposed IBSN to handle multiple types of degradations using a single set of parameters. To better foster the interaction between the input features $\mathbf{F}_c$ and the degradation prompt $\mathbf{P}$, the features $\mathbf{F}_c$ are passed to a global average pooling (GAP) layer, a 1×1 convolution, and a softmax operator to produce a set of dynamic weight coefficients. Each degradation prompt $\mathbf{P}_i$ is combined using these dynamic coefficients $\mathbf{w}_{p\circledast i}$ and subsequently integrated via an upsampling operator $\uparrow$ and 3×3 convolution to obtain the enhanced representation $\mathbf{P}_c$.

$$\begin{aligned} \mathbf{P}_c &= \mathrm{Conv}_{3\times3}((\sum\nolimits_{i=1}^{N} \mathbf{w}_{p\circledast i}\mathbf{P}_i)_{\uparrow}), \\ where \quad \mathbf{w}_p &= \mathrm{Softmax}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{GAP}\left(\mathbf{F}_c\right)\right)\right). \end{aligned} \quad (3)$$

Finally, we utilize a 3×3 convolution to fuse the learned degradation representation conditioned on the extracted features $\mathbf{F}_c$ to enrich the degradation-specific context, obtaining $\hat{\mathbf{Z}}$. This process can be formulated as:

$$\hat{\mathbf{Z}} = \mathrm{Conv}_{3\times3}\left([\mathbf{P}_c; \mathbf{F}_c]\right) \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times12}. \quad (4)$$

4.4. Bit Encryption and Recovery Modules

As shown in Fig. 4, to encode the copyright watermark $\mathbf{w}_{cop}$ into $\mathbf{I}_{med}$, we firstly expand $\mathbf{w}_{cop}\in\{0,1\}^L$ via stacked MLPs and reshape it into several $L\times L$ message feature

Table 1. Localization precision (F1-Score) and bit accuracy (BA) comparison with other competitive methods on [9, 16, 20, 58].

Method	CAISAv1 [9]		Coverage [58]		NIST16 [16]		Columbia [20]	
	F1	BA(%)	F1	BA(%)	F1	BA(%)	F1	BA(%)
ManTraNet [63]	0.320	-	0.486	-	0.225	-	0.650	-
SPAN [22]	0.169	-	0.428	-	0.363	-	0.873	-
CAT-Net v2 [32]	0.852	-	0.582	-	0.417	-	0.923	-
OSN [61]	0.676	-	0.472	-	0.449	-	0.836	-
MVSS-Net [8]	0.650	-	0.659	-	0.372	-	0.781	-
PSCC-Net [39]	0.670	-	0.615	-	0.210	-	0.760	-
TruFor [17]	0.822	-	0.735	-	0.470	-	0.914	-
EditGuard (Ours)	**0.954**	**99.91**	**0.955**	**100**	**0.911**	**99.88**	**0.988**	**99.93**

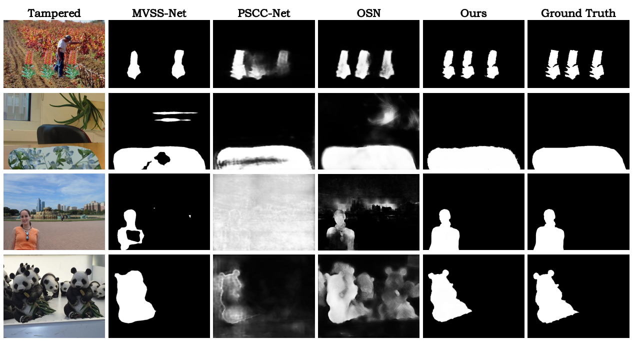

Figure 6. Localization precision comparisons of our EditGuard and competitive methods [8, 39, 61] on four classical benchmarks.

maps. Meanwhile, $\mathbf{I}_{med}$ is fed to a U-style feature enhancement network to extract features of each downsampling and upsampling layer. Finally, the message features will be upscaled and integrated with multi-level image features via the fusion mechanism [21, 62], achieving the modulation of bit-image information. In the decoding end, $\mathbf{I}_{rec}$ is fed into a U-shaped sub-network and downsampled to a size of $L\times L$. The recovered copyright watermark $\hat{\mathbf{w}}_{cop}$ is then extracted via an MLP. More details are presented in **S.M.**.

4.5. Construct EditGuard via the IBSN

To stabilize the optimization of the proposed IBSN, we propose a bi-level optimization strategy. Given an arbitrary original image $\mathbf{I}_{med}$ and watermark $\mathbf{w}_{cop}$, we first train the bit encryption and recovery module via the ℓ_2 loss.

$$\ell_{cop} = \|\mathbf{I}_{con} - \mathbf{I}_{med}\|_2^2 + \lambda \, \|\hat{\mathbf{w}}_{cop} - \mathbf{w}_{cop}\|_2^2 \,, \quad (5)$$

where λ is set to 10. Furthermore, we freeze the weights of BEM and BRM and jointly train the IHM and IRM. Given a random original image $\mathbf{I}_{ori}$, localization watermark $\mathbf{W}_{loc}$ and copyright watermark $\mathbf{w}_{cop}$, the loss function is:

$$\ell_{loc} = \|\hat{\mathbf{I}}_{ori} - \mathbf{I}_{ori}\|_1 + \alpha \, \|\mathbf{I}_{con} - \mathbf{I}_{ori}\|_2^2 + \beta \|\hat{\mathbf{W}}_{loc} - \mathbf{W}_{loc}\|_1, \quad (6)$$

where α and β are respectively set to 100 and 1. During training, we only introduced degradations to $\mathbf{I}_{con}$ without being exposed to any tampering. After acquiring a pre-trained IBSN, we can construct the proposed EditGuard via the components of IBSN. As plotted in Fig. 4, the dual-watermark encoder of EditGuard is composed of IHM and BEM, which correspond to the localization and copyright watermarking in Fig. 3 respectively. The copyright extractor strictly corresponds to BRM. The tamper locater

Table 2. Visual quality of the container image $\mathbf{I}_{con}$ and bit accuracy comparison with other pure watermarking methods.

Method	Image Size	M. L.	PSNR (dB)	SSIM	NIQE(↓)	BA(%)
MBRS [26]	128×128	30	26.57	0.886	7.219	**100**
CIN [44]	128×128	30	**41.35**	**0.981**	7.171	99.99
PIMoG [10]	128×128	30	36.22	0.941	7.113	99.99
SepMark [62]	128×128	30	35.42	0.931	7.095	99.86
EditGuard	128×128	30	36.93	0.944	**5.567**	99.89
EditGuard	512×512	64	37.77	0.949	4.257	99.95

includes IRM and a mask extractor (ME). Note that we need to pre-define a localization watermark $\mathbf{W}_{loc}$, which is shared between the encoding and decoding ends. The choice of $\mathbf{W}_{loc}$ is very general to our method. It can be any natural image or even a solid color image. Finally, by comparing the pre-defined watermark $\mathbf{W}_{loc}$ with the decoded one $\hat{\mathbf{W}}_{loc}$, we can obtain a binary mask $\hat{\mathbf{M}}\in\mathbb{R}^{H\times W}$:

$$\hat{\mathbf{M}}[i,j] = \theta_\tau(\max(|\hat{\mathbf{W}}_{loc}[i,j,:] - \mathbf{W}_{loc}[i,j,:]|)). \quad (7)$$

where $i \in [0, H)$ and $j \in [0, W)$. $\theta_\tau(z) = 1 \ (z \geq \tau)$. τ is set to 0.2. $|\cdot|$ is an absolute value operation.

5. Experiments

5.1. Implementation Details

We trained our EditGuard via the training set of COCO [38] **without any tampered data**. Thus, for tamper localization, our method is actually zero-shot. The Adam [30] is used for training 250K iterations with β_1=0.9 and β_2=0.5. The learning rate is initialized to 1×10^{-4} and decreases by half for every 30K iterations, with the batch size set to 4. We embed a 64-bit copyright watermark and a simple localization watermark such as a pure blue image ([R, G, B] = [0, 0, 255]) to original images. Following [8, 17, 39], F1-score, AUC, IoU, and bit accuracy are used to evaluate localization and copyright protection performance. Since no prior methods can simultaneously achieve this dual forensics, we conducted separate comparisons with tamper localization and image watermarking methods.

5.2. Comparison with Localization Methods

For a fair comparison with tamper localization methods, we conducted extensive evaluations on four classical benchmarks [9, 16, 20, 58], as reported on Tab. 1. Since EditGuard is a proactive approach, we initially embed watermarks into authentic images and then paste the tampered areas into the container images. Remarkably, even for tamper types that existing methods specialize in, the localization accuracy of EditGuard consistently outperforms the SOTA method [17] across four datasets by margins of **0.102, 0.116, 0.441, and 0.065 in F1-score without any labeled data or tampered samples required**, which verifies the superiority of our proactive localization mechanism. As shown in Fig. 6, our EditGuard can precisely pinpoint pixel-level tampered areas but other methods can only produce a rough outline or are only effective in some cases. Mean-

Table 3. Comparison with other competitive tamper localization methods under different AIGC-based editing methods. Note that † denotes the network finetuned in our constructed AGE-Set-C.

Method	Stable Diffusion Inpaint [51]				Controlnet [72]				SDXL [49]			
	F1	AUC	IoU	BA(%)	F1	AUC	IoU	BA(%)	F1	AUC	IoU	BA(%)
MVSS-Net [8]	0.178	0.488	0.103	-	0.178	0.492	0.103	-	0.037	0.503	0.028	-
OSN [61]	0.174	0.486	0.101	-	0.191	0.644	0.110	-	0.200	0.755	0.118	-
PSCC-Net [39]	0.166	0.501	0.112	-	0.177	0.565	0.116	-	0.189	0.704	0.115	-
IML-VIT [45]	0.213	0.596	0.135	-	0.200	0.576	0.128	-	0.221	0.603	0.145	-
HiFi-Net [18]	0.547	0.734	0.128	-	0.542	0.735	0.123	-	0.633	0.828	0.261	-
MVSS-Net† [8]	0.694	0.939	0.575	-	0.678	0.925	0.558	-	0.482	0.884	0.359	-
EditGuard (Ours)	**0.966**	**0.971**	**0.936**	**99.95**	**0.968**	**0.987**	**0.940**	**99.96**	**0.965**	**0.989**	**0.936**	**99.96**

Method	RePaint [43]				Lama [57]				FaceSwap [60]			
	F1	AUC	IoU	BA(%)	F1	AUC	IoU	BA(%)	F1	AUC	IoU	BA(%)
MVSS-Net [8]	0.104	0.546	0.082	-	0.024	0.505	0.022	-	0.285	0.612	0.192	-
OSN [61]	0.183	0.644	0.105	-	0.170	0.430	0.099	-	0.308	0.791	0.171	-
PSCC-Net [39]	0.140	0.469	0.109	-	0.132	0.329	0.104	-	0.157	0.346	0.180	-
IML-VIT [45]	0.103	0.497	0.059	-	0.105	0.465	0.064	-	0.105	0.465	0.064	-
HiFi-Net [18]	0.681	0.896	0.339	-	0.483	0.721	0.029	-	0.781	0.890	0.478	-
MVSS-Net† [8]	0.185	0.529	0.111	-	0.393	0.829	0.275	-	0.459	0.739	0.333	-
EditGuard (Ours)	**0.967**	**0.977**	**0.938**	**99.95**	**0.965**	**0.969**	**0.934**	**99.95**	**0.896**	**0.943**	**0.876**	**99.86**

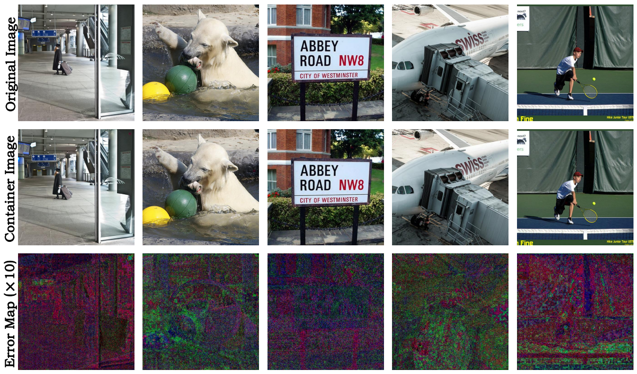

Figure 7. Visual results of the container image $\mathbf{I}_{con}$ and the error map of the proposed EditGuard. Here, localization and copyright watermarks are randomly selected from the dataset.

while, our bit accuracy remains over **99.8%** while all other methods can not realize effective copyright protection.

5.3. Comparison with Watermarking Methods

To evaluate the visual quality of $\mathbf{I}_{con}$, we compared EditGuard with other watermarking methods on $1K$ testing images from COCO [38] under the tampering of stable diffusion inpaint [51]. For a fair comparison, we also retrained our EditGuard on 128×128 original images and 30 bits. Tab. 2 reports that the fidelity of our container image far surpasses that of SepMark [62], PIMoG [10], and MBRS [26] but is slightly inferior to CIN [44]. Meanwhile, our method exhibits the best performance in perceptual quality measures like NIQE. As shown in Fig. 7, dual-watermarked images do not have noticeable artifacts and noise, making them imperceptible to the human eyes. When suffer malicious tampering, our method outperforms SepMark and is very close to PIMoG and CIN in bit accuracy. Note that other competitive methods only hide 30 bits, with a capacity of 30/(128×128). In contrast, our EditGuard hides both an RGB localization watermark and a 1D copyright watermark, with a capacity far greater than 30/(128×128). Here, we do not claim to achieve the best visual quality and bit accuracy, but just to demonstrate that our method is comparable to the current image watermarking methods.

5.4. Extension to AIGC-based Editing Methods

Dataset Preparation: We constructed a dataset tailored for AIGC Editing methods, dubbed AGE-Set, comprising two sub-datasets. The first AGE-Set-C is a batch-processed coarse tamper dataset. Its original images are sourced from COCO 2017 [38] and CelebA [41], containing $30K$ training images and $1.2K$ testing images. We used some SOTA editing methods such as Stable Diffusion Inpaint [51], Controlnet [72], SDXL [49] to manipulate images with the prompt to be "None", and employed some unconditional methods like Repaint [43], Lama [57], and Faceswap [60]. **Note that we only use the tampered data to train other methods, not our EditGuard.** The second sub-dataset AGE-Set-F includes 100 finely edited images. It is edited manually via some sophisticated software such as SD-Web-UI, Photoshop, and Adobe Firefly. These AIGC-based editing methods can achieve a good fusion of the tampered and unchanged areas, making it hard for the naked eye to catch artifacts. More details are presented in **S.M.**.

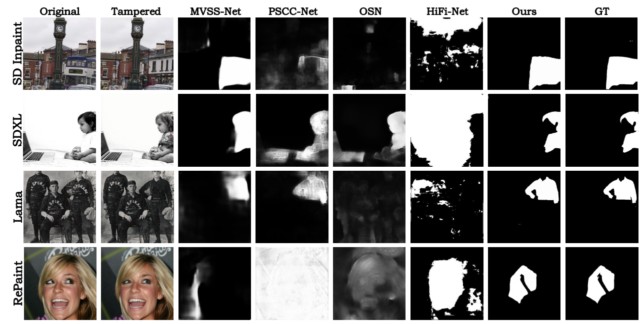

Figure 8. Localization performance comparisons of our EditGuard and other methods [8, 18, 39, 61] on our constructed AGE-Set-C.

AGE-Set-C: Tab. 3 presents the comparison of our EditGuard and some SOTA tamper localization methods [8, 18, 39, 45, 61]. We observe that the F1-scores of other passive forensic methods are generally lower than 0.7 when applied to AGE-Set-C. Meanwhile, even when we try our best to finetune MVSS-Net using AGE-Set-C, the accuracy of MVSS-Net† remains unsatisfactory, and they exhibit catastrophic forgetting across various tamper methods. In contrast, our method can guarantee an F1-score and AUC of over **95%**, maintaining around **90%** IOU, regardless of tampering types. As shown in Fig. 8, our EditGuard can accurately capture these imperceptible tampering traces produced by AIGC-based editing methods, but other methods are almost ineffective. Moreover, our EditGuard can effectively recover copyright information with a bit accuracy exceeding **99.8%**. Noting that none of the comparison methods offer copyright protection capabilities.

AGE-Set-F: To further highlight the practicality of our EditGuard, we conducted subjective comparisons with other methods on the meticulously tampered AGE-Set-F.

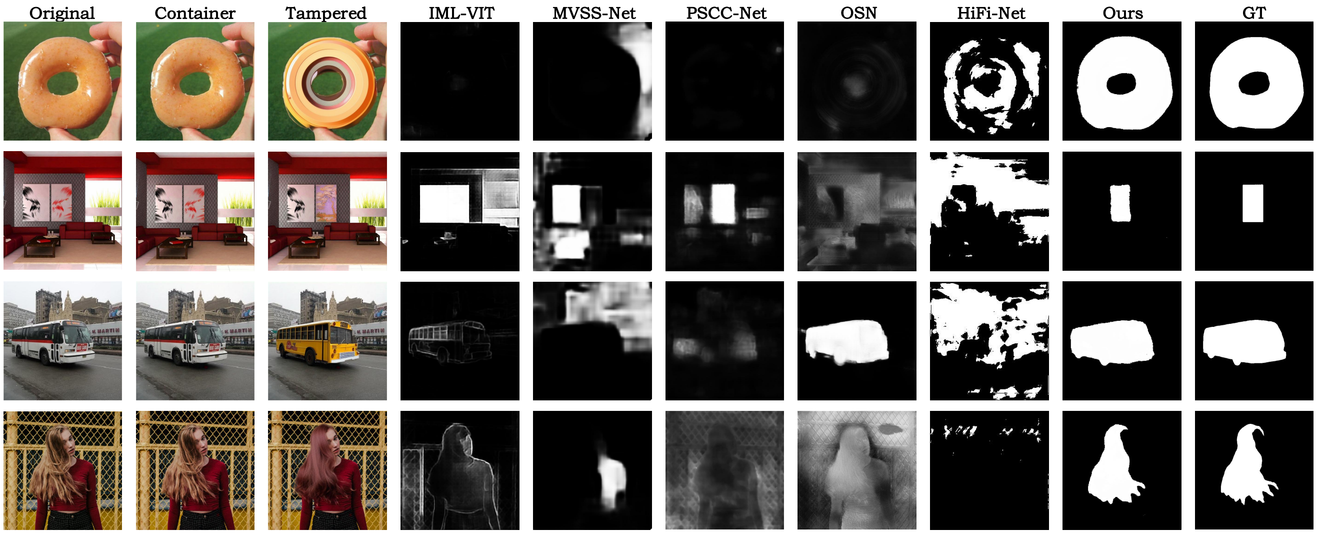

Figure 9. Localization precision comparisons of our EditGuard and other competitive methods on the meticulously tampered AGE-Set-F.

Table 4. Localization and bit recovery performance of our EditGuard and MVSS-Net† [8] under different levels of degradations.

Methods	Metrics	Clean	Gaussian Noise		JPEG			Poisson
			σ=1	σ=5	Q=70	Q=80	Q=90	
MVSS-Net† [8]	F1	0.694	0.644	0.619	0.458	0.507	0.558	0.652
	BA(%)	-	-	-	-	-	-	-
EditGuard (Ours)	F1	0.966	0.937	0.932	0.920	0.920	0.925	0.943
	BA(%)	99.95	99.94	99.37	97.69	98.16	98.23	99.91

The tamper types in this subset did not appear in the training set. As shown in Fig. 9, when faced with real-world tampering, even the most powerful tamper localization methods almost entirely fail. This is due to their mechanisms to look for image artifacts and explore instance-wise semantic information. However, our EditGuard, which locates tampered masks via the natural fragility and locality of I2I steganography, can still clearly annotate the tampered area.

5.5. Robustness Analysis

As shown in Tab. 4, we conducted robustness analysis on the tampering of "Stable Diffusion Inpaint" [51] under Gaussian noise with σ=1 and 5, JPEG compression with Q=70, 80 and 90, and Poisson noise with α=4 [71]. We observed that our method still maintains a high localization accuracy (F1-score>0.9) and bit accuracy with a very slight performance decrease under various levels of degradations, while MVSS-Net† [8] exhibits a noticeable performance degradation compared to its results in clean conditions. It is attributed to our prompt-based estimation that can effectively learn the degradation representation.

5.6. Ablation Study

To verify the effectiveness of each component of the EditGuard, we conducted ablation studies on bi-level optimization (BO), lightweight feature interaction module (LFIM), transformer block (TB), and prompt-based fusion (PF) under the tampering of "Stable Diffusion Inpaint". As listed in Tab. 5, without BO, the joint training of all components

Table 5. Abalation studies on the core components of EditGuard.

Case	Degradation Type $\mathcal{D}(\cdot)$	PF	TB	LFIM	BO	F1	AUC	IoU	BA(%)
(a)	Clean	✓	✓	✓		-	-	-	49.17
(b)	Clean	✓	✓		✓	0.950	0.960	0.904	99.73
(c)	Clean	✓		✓	✓	0.957	0.966	0.927	99.51
(d)	Random Degradations		✓	✓	✓	0.903	0.933	0.841	99.12
Ours	Clean	✓	✓	✓	✓	0.966	0.971	0.936	99.95
	Random Degradations	✓	✓	✓	✓	0.938	0.964	0.887	99.36

cannot converge effectively, resulting in bit accuracy that is close to random guessing. Without LFIM and TB, the IoU of EditGuard will suffer 0.032 and 0.009 declines since these two modules can better perform feature fusion. Without PF, the robustness of the EditGuard will significantly decline. We observed that the F1/AUC/IoU of our method far surpasses that of case (d) by 0.035/0.031/0.046 under "Random Degradations", which indicates that the PF effectively enables a single network to support watermark recovery under various degradations. "Random Degradations" denotes that we randomly set the $\mathcal{D}(\cdot)$ to various levels of Gaussian noise, Poisson noise, and JPEG compression.

6. Conclusion

We present the first attempt to design a deep versatile watermarking mechanism **EditGuard**. It enhances the credibility of images by embedding imperceptible localization and copyright watermarks, and decoding accurate copyright information and tampered areas, making it a reliable tool for artistic creation and legal forensic analysis. In the future, we will focus on improving the robustness of EditGuard and strive not only to offer pixel-wise localization results but also to provide semantic-wise outcomes. Additionally, we plan to further expand EditGuard to a broader range of modalities and applications, including video, audio, and 3D scenes. Our efforts at information authenticity serve not only the AIGC industry, but the trust in our digital world, ensuring that every pixel tells the truth and the rights of each individual are safeguarded.

References

[1] Mahdi Ahmadi, Alireza Norouzi, Nader Karimi, Shadrokh Samavi, and Ali Emami. Redmark: Framework for residual diffusion watermarking based on deep networks. *Expert Systems with Applications*, 146:113157, 2020. 3

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

[3] Shumeet Baluja. Hiding images in plain sight: Deep steganography. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 3

[4] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 4

[5] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

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

[7] Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, and Jiliang Tang. Diffusionshield: A watermark for copyright protection against generative diffusion models. *arXiv preprint arXiv:2306.04642*, 2023. 3

[8] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2022. 2, 6, 7, 8

[9] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *Proceedings of the IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, 2013. 6

[10] Han Fang, Zhaoyang Jia, Zehua Ma, Ee-Chien Chang, and Weiming Zhang. 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)*, 2022. 3, 6, 7

[11] Han Fang, Yupeng Qiu, Kejiang Chen, Jiyi Zhang, Weiming Zhang, and Ee-Chien Chang. Flow-based robust watermarking with invertible noise layer for black-box distortions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 3

[12] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[13] Anirban Ghoshal. Artists lose first copyright battle in the fight against ai-generated images. https://www.computerworld.com/article/3709691/artists-lose-first-copyright-battle-in-the-fight-against-ai-generated-images.html, 2023. 2

[14] Gabriel Goh, James Betker, Li Jing, Aditya Ramesh, et al. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf, 2023. 1

[15] Google DeepMind. Synthid. https://deepmind.google/technologies/synthid/, 2023. 2

[16] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *Proceedings of the IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019. 6

[17] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. 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)*, 2023. 2, 6

[18] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 7

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[20] Yu-Feng Hsu and Shih-Fu Chang. Detecting image splicing using geometry invariants and camera characteristics consistency. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2006. 6

[21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[22] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 6

[23] Xiaoxiao Hu, Qichao Ying, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. Draw: Defending camera-shooted raw against image manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[24] Nasir N Hurrah, Shabir A Parah, Nazir A Loan, Javaid A Sheikh, Mohammad Elhoseny, and Khan Muhammad. Dual watermarking framework for privacy protection and content authentication of multimedia. *Future generation computer Systems*, 94:654–673, 2019. 2

[25] Ashraful Islam, Chengjiang Long, Arslan Basharat, and Anthony Hoogs. 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)*, 2020. 2

[26] Zhaoyang Jia, Han Fang, and Weiming Zhang. 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)*, 2021. 6, 7

[27] Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong. Evading watermark based detection of ai-generated content. *arXiv preprint arXiv:2305.03807*, 2023. 3

[28] Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. Hinet: Deep image hiding by invertible network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[29] Asra Kamili, Nasir N Hurrah, Shabir A Parah, Ghulam Mohiuddin Bhat, and Khan Muhammad. Dwfcat: Dual watermarking framework for industrial image authentication and tamper localization. *IEEE Transactions on Industrial Informatics*, 17(7):5108–5117, 2020. 2

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[31] Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. 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)*, 2021. 2

[32] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022. 6

[33] Haodong Li and Jiwu Huang. Localization of deep inpainting using high-pass fully convolutional network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[34] Yuanman Li and Jiantao Zhou. Fast and effective image copy-move forgery detection via hierarchical feature point matching. *IEEE Transactions on Information Forensics and Security*, 14(5):1307–1322, 2018. 2

[35] Yue Li, Dong Liu, Houqiang Li, Li Li, Zhu Li, and Feng Wu. Learning a convolutional neural network for image compact-resolution. *IEEE Transactions on Image Processing*, 28(3): 1092–1107, 2018. 2

[36] Siau-Chuin Liew, Siau-Way Liew, and Jasni Mohd Zain. Tamper localization and lossless recovery watermarking scheme with roi segmentation and multilevel authentication. *Journal of digital imaging*, 26:316–325, 2013. 2

[37] Chia-Chen Lin, Ting-Lin Lee, Ya-Fen Chang, Pei-Feng Shiu, and Bohan Zhang. Fragile watermarking for tamper localization and self-recovery based on ambtc and vq. *Electronics*, 12(2):415, 2023. 2

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 6, 7

[39] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (11):7505–7517, 2022. 6, 7

[40] Yang Liu, Mengxi Guo, Jian Zhang, Yuesheng Zhu, and Xiaodong Xie. A novel two-stage separable deep learning framework for practical blind watermarking. In *Proceedings of the ACM International Conference on Multimedia (MM)*, 2019. 3

[41] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 7

[42] Shao-Ping Lu, Rong Wang, Tao Zhong, and Paul L Rosin. Large-capacity image steganography based on invertible neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 4

[43] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 7

[44] Rui Ma, Mengxi Guo, Yi Hou, Fan Yang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Towards blind watermarking: Combining invertible and non-invertible mechanisms. In *Proceedings of the ACM International Conference on Multimedia (MM)*, 2022. 2, 3, 6, 7

[45] Xiaochen Ma, Bo Du, Xianggen Liu, Ahmed Y Al Hammadi, and Jizhe Zhou. Iml-vit: Image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*, 2023. 2, 7

[46] Chong Mou, Youmin Xu, Jiechong Song, Chen Zhao, Bernard Ghanem, and Jian Zhang. Large-capacity and flexible video steganography via invertible neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 4

[47] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024. 1

[48] Neena Raj NR and R Shreelekshmi. Fragile watermarking scheme for tamper localization in images using logistic map and singular value decomposition. *Journal of Visual Communication and Image Representation*, 85:103500, 2022. 2

[49] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 7

[50] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3, 7, 8

[52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the International conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1

[53] Ronald Salloum, Yuzhuo Ren, and C-C Jay Kuo. Image splicing localization using a multi-task fully convolutional network (mfcn). *Journal of Visual Communication and Image Representation*, 51:201–209, 2018. 2

[54] Deepa Shivaram. The white house and big tech companies release commitments on managing ai. https://www.nprillinois.org/2023-07-21/the-white-house-and-big-tech-companies-release-commitments-on-managing-ai, 2023. 2

[55] Xiaopeng Sun, Weiqi Li, Zhenyu Zhang, Qiufang Ma, Xuhan Sheng, Ming Cheng, Haoyu Ma, Shijie Zhao, Jian Zhang, Junlin Li, et al. Opdn: Omnidirectional position-aware deformable network for omnidirectional image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, pages 1293–1301, 2023. 1

[56] 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. 2

[57] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. 7

[58] 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. 6

[59] 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. 3

[60] Huikai Wu, Shaocheng Xiang, Gabriben, and Niczem. Faceswap. https://github.com/wuhuikai/FaceSwap, 2020. 7

[61] 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. 2, 6, 7

[62] Xiaoshuai Wu, Xin Liao, and Bo Ou. Sepmark: Deep separable watermarking for unified source tracing and deepfake detection. In *Proceedings of the ACM international conference on Multimedia (MM)*, 2023. 2, 3, 6, 7

[63] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. 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)*, 2019. 2, 6

[64] Youmin Xu, Chong Mou, Yujie Hu, Jingfen Xie, and Jian Zhang. Robust invertible image steganography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 4

[65] Qichao Ying, Zhenxing Qian, Hang Zhou, Haisheng Xu, Xinpeng Zhang, and Siyi Li. From image to imuge: Immunized image generation. In *Proceedings of the ACM international conference on Multimedia (MM)*, 2021. 2

[66] Qichao Ying, Xiaoxiao Hu, Xiangyu Zhang, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. Rwn: Robust watermarking network for image cropping localization. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2022.

[67] Qichao Ying, Hang Zhou, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. Learning to immunize images for tamper localization and self-recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[68] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1

[69] Jiwen Yu, Xuanyu Zhang, Youmin Xu, and Jian Zhang. Cross: Diffusion model makes controllable, robust and secure image steganography. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[70] 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. 5

[71] Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Radu Timofte, and Luc Van Gool. Practical blind denoising via swin-conv-unet and data synthesis. *arXiv preprint arXiv:2203.13278*, 2022. 8

[72] 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. 1, 7

[73] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023. 3

[74] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[75] Xinshan Zhu, Yongjun Qian, Xianfeng Zhao, Biao Sun, and Ya Sun. A deep learning approach to patch-based image inpainting forensics. *Signal Processing: Image Communication*, 67:90–99, 2018. 2