Creating Blank Canvas Against AI-enabled Image Forgery

Qi Song, Ziyuan Luo, Renjie Wan*

Department of Computer Science, Hong Kong Baptist University
{qisong, ziyuanluo}@life.hkbu.edu.hk, {renjiewan}@hkbu.edu.hk

Abstract

AIGC-based image editing technology has greatly simplified the realistic-level image modification, causing serious potential risks of image forgery. This paper introduces a new approach to tampering detection using the Segment Anything Model (SAM). Instead of training SAM to identify tampered areas, we propose a novel strategy. The entire image is transformed into a blank canvas from the perspective of neural models. Any modifications to this blank canvas would be noticeable to the models. To achieve this idea, we introduce adversarial perturbations to prevent SAM from "seeing anything", allowing it to identify forged regions when the image is tampered with. Due to SAM's powerful perceiving capabilities, naive adversarial attacks cannot completely tame SAM. To thoroughly deceive SAM and make it blind to the image, we introduce a frequency-aware optimization strategy, which further enhances the capability of tamper localization. Extensive experimental results demonstrate the effectiveness of our method.

Code — https://github.com/qsong2001/blank_canvas

1 Introduction

Image tamper localization aims to mitigate the proliferation of forged imagery that threatens public trust and social stability. Contemporary AIGC frameworks (Rombach et al. 2022; Podell et al. 2024; Zhang, Rao, and Agrawala 2023; Lugmayr et al. 2022; Wu et al. 2020) generate images with unprecedented photorealism, rendering existing tamper localization approaches (Li and Huang 2019; Zhang et al. 2024; Xu et al. 2025b) increasingly inadequate.

Current tamper localization methods employ a post-hoc approach, analyzing content only after manipulation has occurred. Such a manner relies on specific forgery patterns learned during task-specific training (Dong et al. 2022; Sun et al. 2023; Ma et al. 2023; Ramesh et al. 2022; Asnani et al. 2023). However, such reliance undermines generalizability (Zhang et al. 2024) when encountering novel forgery contents absent from training data. As AI models advance toward generating increasingly diverse and novel content, these limitations will become more pronounced. To address this, we need **new thinking** beyond the current post-hoc strategy that relies on learned image tampering traces.

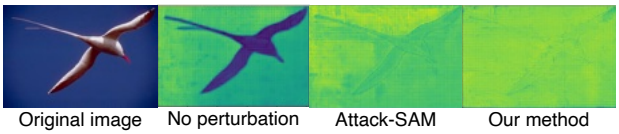

Figure 1: Predicted segmentation confidence maps from the segment anything model. Previous adversarial attack (Zhang et al. 2023) could not fully disrupt the SAM's perception, especially in edges and texture areas.

Recently, various powerful vision foundation models (Kirillov et al. 2023) have shown remarkable generalizability across various tasks. A good way is to leverage their generalizability in tamper localization. However, acquiring adequate datasets to develop a foundation model for tampering detection remains prohibitively challenging. Moreover, even with sufficient data, training such a foundation model requires substantial computational resources (Kwon et al. 2025), which may be inaccessible to many photo owners. To address this, we introduce a proactive approach: instead of relying on post-hoc localization, photo owners could embed invisible protective layers into images before distribution. These layers enable an off-the-shelf foundation model (*e.g.*, SAM) to automatically detect tampered regions without requiring task-specific tampering training.

This paper proposes a novel "blank-canvas" mechanism to achieve this envisioned scenario, leveraging the insight that tampering is more discernible in a simplified context. As illustrated in Fig. 2, tampering artefacts, often imperceptible in complex images due to complex textures, become prominently visible on a blank canvas. To exploit this, we suggest transforming the complex image into a "blank canvas" in the perspective of vision foundation model. Then, once such a "blank canvas" is tampered, the vision foundation model can readily identify the tampered areas. In our setting, we propose using the Segment Anything Model (SAM) (Kirillov et al. 2023), the popular off-the-shelf foundation model with the required capability in perceiving visual content. Then, we incorporate adversarial perturbations to images. The adversarial perturbations are designed to mislead SAM's mechanism, directing its focus toward manipulated regions rather than authentic content.

*Corresponding author.

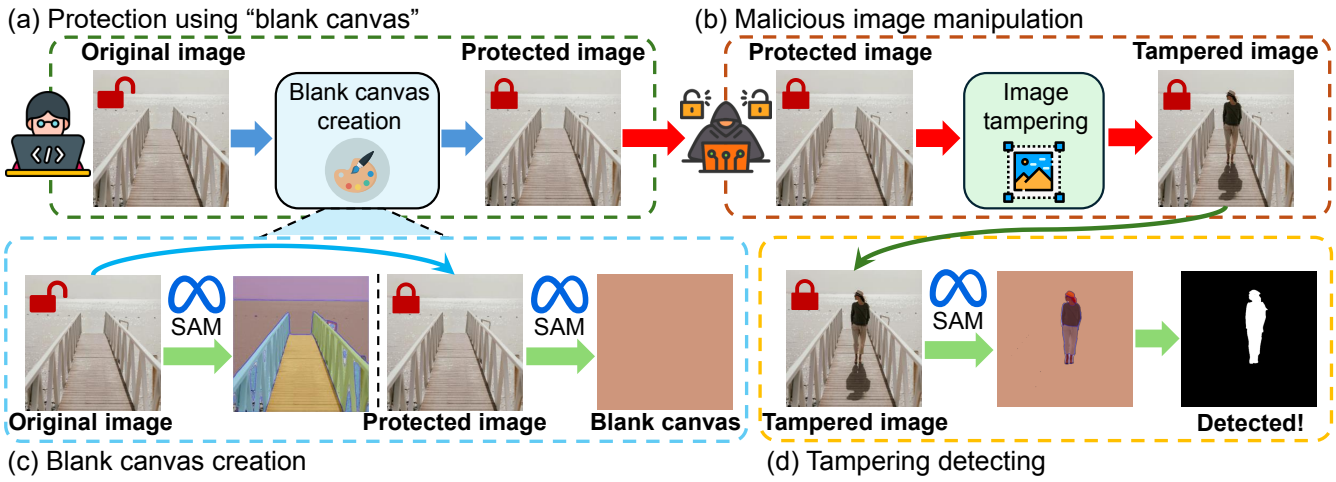

Figure 2: Overview of our tamper localization framework. (a) The protection process transforms an original image into a protected image through our blank canvas creation mechanism. (b) Illustration of potential malicious image manipulation on the protected image. (c) The blank canvas creation process demonstrates how the original image generates detailed SAM segmentation results while our protected image appears as a blank canvas to SAM (Kirillov et al. 2023). (d) The tampering detection phase shows SAM (Kirillov et al. 2023) successfully identifying the tampered region when the protected image is maliciously manipulated.

However, SAM is a powerful model, and conventional adversarial attacks (Madry et al. 2018; Zhang et al. 2023) cannot completely disrupt SAM perception (as further discussed in section 5.1). Segmentation models like SAM are naturally designed to distinguish textural regions and structural boundaries within an image. However, as shown in Fig. 1, the high-frequency structural patterns inherent in complex images (*e.g.*, edges, textures) retain perceptually salient contours. This challenge arises as SAM is inherently designed to identify different regions in an image based on the distinct features (Kirillov et al. 2023), which are often characterized by their frequency components in the spectral domain (Zhou et al. 2024; Xia et al. 2023). Consequently, this leads to fragmented segmentation outputs instead of achieving a coherent suppression of structural elements. To effectively obscure SAM's perception in these high-frequency areas, we propose a Frequency-aware optimization that manipulates the image's frequency components, creating a more robust disguise against SAM's perception. Our strategy ensures that SAM perceives the protected image as a blank canvas, successfully enabling the transition from “segment anything” to “segment nothing”. Malicious editing on the protected image could be identified as SAM perceives those areas of discrepancy. Our contributions can be summarized as:

- We introduce a novel method for tamper localization that utilizes adversarial perturbations to hinder the segmentation capabilities of the SAM, thereby enhancing its ability to detect tampered regions.
- We present a new concept that edits made on a “blank canvas” are more conspicuous and easier to identify, thereby contributing to the broader field of image security.
- We propose a frequency-aware optimization strategy designed to deceive the segment anything model effectively.

Extensive experimental results demonstrate the effectiveness of our solution for tamper localization.

2 Related Works

2.1 Tamper localization

The rapid advancements in AI-based image editing technology (Rombach et al. 2022; Podell et al. 2024; Zhang, Rao, and Agrawala 2023; Lugmayr et al. 2022; Wu et al. 2020) have significantly benefited photographers and image editors by enabling unprecedented creative possibilities. However, these powerful capabilities also pose significant challenges for information security. Different tamper localization methods are proposed to address this challenge. Existing approaches to image forensics predominantly concentrate on detecting predefined manipulation categories through passive analysis (Li and Zhou 2018; Zhu et al. 2018; Wu et al. 2022; Salloum, Ren, and Kuo 2018; Islam et al. 2020; Li and Huang 2019; Sun et al. 2024, 2023; Zhang et al. 2025). Beyond these specialized solutions, broader-spectrum detection frameworks (Ying et al. 2021; Kwon et al. 2021; Chen et al. 2021; Wu and et al. 2019; Li et al. 2018; Ying et al. 2023; Hu et al. 2023; Ying et al. 2022) attempt to identify tampering traces by analyzing inherent inconsistencies in forged images. Representative efforts include HiFi-Net (Guo et al. 2023), which implements hierarchical feature extraction for both synthetic and edited content, and Trufor (Guillaro et al. 2023) that combines a transformer-enhanced fusion of noise patterns with spectral characteristics. While OSN (Wu et al. 2022) enhances robustness against quality degradation through adaptive training strategies, SAFL-Net (Sun et al. 2023) enforces manipulation-sensitive feature learning via auxiliary constraints. Despite these advancements, current passive detectors (Dong et al. 2022; Sun et al. 2023; Ma et al. 2023; Ramesh et al. 2022; Asnani et al. 2023) frequently suffer from constrained generalizability and precision, typically excelling only on manipulation types encountered during training. Even proactive solutions like MaLP (Asnani et al. 2023), despite employing template-matching strategies, remain dependent on exten-

sive forged samples and retain architecture-level coupling to specific tamper characteristics. Recently, EditGuard (Zhang et al. 2024) proposed an active protection method that embeds information into images. Although effective in certain scenarios, this approach suffers from limited interpretability and practical constraints due to its reliance on non-intuitive steganographic content comparisons. To address these challenges, our method introduces visually detectable traces that enable transparent and human-verifiable tamper localization without requiring pre-registered steganographic references.

2.2 Adversarial attacks on SAM

As artificial intelligence systems become widely deployed, AI safety and assets are increasingly becoming a critical issue (Huang et al. 2024a; Luo et al. 2023, 2025b,c; Song et al. 2024a,b; Huang et al. 2025; Xu et al. 2025a; Li and Cheung 2025, 2024; Huang et al. 2024b; Luo et al. 2025a). Deep neural networks are widely known to be vulnerable to adversarial attacks CNN (Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2015; Kurakin, Goodfellow, and Bengio 2017) and vision transformer (ViT) (Dosovitskiy et al. 2021; Bhojanapalli et al. 2021; Mahmood, Mahmood, and Van Dijk 2021). This vulnerability has inspired numerous works investigating the model's robustness under various adversarial attacks. Typical adversarial attacks have primarily focused on manipulating image-level label predictions in image classification tasks. For semantic segmentation, prior work (Xie et al. 2017; Arnab, Miksik, and Torr 2018; Hendrik Metzen et al. 2017) studies adversarial attacks on traditional semantic segmentation models. Adversarial attacks on SAM differ from attacks on semantic segmentation models since the generated masks do not have semantic labels. Recent works (Shen, Li, and Wang 2024; Xia et al. 2023; Zhou et al. 2024; Zhang et al. 2023) have extended adversarial attacks to the emerging Segment Anything Model (SAM) (Kirillov et al. 2023) for prompt-based mask prediction. SAM-attack (Zhang et al. 2023) investigates the robustness of SAM against adversarial attacks. S-RA (Shen, Li, and Wang 2024) and Dark-SAM (Zhou et al. 2024) aim to build a unified, transfer-adversarial generation framework that combines a frequency-domain loss and an area-consistency loss. MUI-GRAT (Xia et al. 2023) investigates transferable attacks simultaneously compromising SAM and its downstream models, proposing a parameter-freezing strategy to enhance cross-model transferability. These studies reveal unique vulnerabilities in SAM's mask prediction paradigm, differing fundamentally from traditional semantic segmentation attacks due to SAM's class-agnostic segmentation nature.

3 Preliminaries

Segment anything model. Segment Anything Model (SAM) (Kirillov et al. 2023) is the first vision foundation model for image segmentation. Unlike conventional semantic segmentation methods (Minaee et al. 2021; Haralick and Shapiro 1985) that produce class-labeled masks, SAM outputs unlabeled segmentation results corresponding to prompt-specified regions. The fundamental data structure can be formally defined as a triplet $(x, \mathcal{P}, \mathcal{M})$, where $x \in \mathbb{R}^{H\times W\times 3}$ and $\mathcal{P}$ denotes input image and prompt vector (spatial coordinates or bounding boxes), $\mathcal{M} \in \{0,1\}^{H\times W}$ is target binary mask. For a given image x, multiple valid annotations can be generated through stochastic prompt sampling across the image plane, forming an augmented dataset: $\mathbb{D} = \{(x, \mathcal{P}_i, \mathcal{M}_i)\}_{i=1}^{N}$. The model's inference process can be expressed as:

$$\Phi = \text{SAM}(x, \mathcal{P}), \tag{1}$$

where $\Phi \in \mathbb{R}^{H\times W}$ denotes the confidence score matrix with equivalent spatial dimensions to the input image. The segmentation decision rule follows a threshold operation:

$$\mathcal{M}_{\text{pred}}[i,j] = \begin{cases} 1 & \Phi[i,j] > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The resultant binary mask $\mathcal{M}_{\text{pred}}$ preserves the original image resolution through its $H \times W$ dimensional structure. Notably, this architecture enables dynamic mask generation conditioned on various prompt types (points, boxes, or text) while maintaining spatial coherence in the output.

Our scenario. We aim to force SAM to perceive the protected image as a *blank canvas*. Then, any subsequent image manipulation could lead SAM to identify the tampered regions as anomalous segmentation responses. This analogy is to detecting chalk marks on a blank canvas: 1) Before tampering, SAM produces no masks; 2) Tampering acts like chalk strokes on a blank canvas, making altered pixels "noticeable" in SAM's output.

During usage, image owners can implement our method and transform their images into a "blank canvas" from the SAM's perspective before distribution. These images appear unchanged, as the added perturbations are invisible and do not significantly affect the original image quality. However, if malicious users attempt to alter these protected images, the modifications would be noticeable to the SAM models, such as those made on a blank canvas. Users and third-party investors can directly access the anomalous areas identified by a naive SAM model.

4 Proposed Method

This section presents our training-free approach for adapting SAM as a foundation model for tamper localization. Our approach contains two stages: 1) *Blank canvas creation*, which involves strategically suppressing SAM's segmentation capabilities across the images, thereby creating a perceived "blank canvas" from SAM's view. 2) *Tamper localization*, during which, if any tampering occurs, the manipulated areas become noticeable from SAM's perspective, allowing it to identify tampered regions.

4.1 Blank canvas creation

Instead of identifying tampered areas through learned artifacts passively, we proactively transform the image into a "blank canvas", an intentional special state where SAM's segmentation confidence is uniformly distributed. Any further manipulations inevitably introduce abnormal deviations, which SAM recognizes as anomalous segmentation targets, effectively identifying the tampered area.

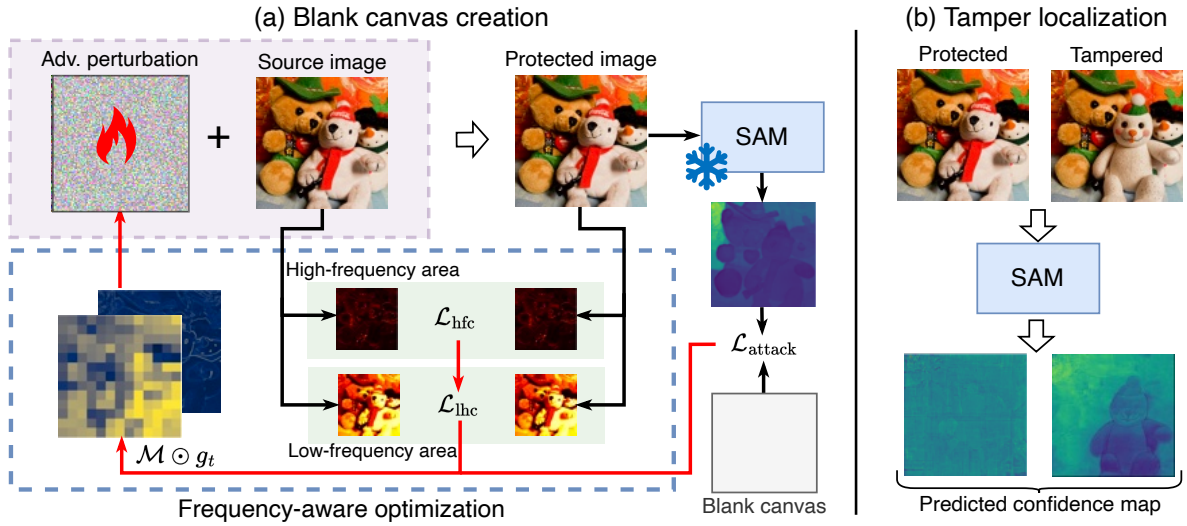

Figure 3: Overall of our method. We enable the source image to be a "blank canvas" from the perspective of SAM. Frequency-aware optimization is proposed to disrupt the high-frequency areas to deceive the SAM model fully. After the image is protected, the tampered locations become noticeable to SAM.

Specifically, we aim to force SAM to perceive the protected image as a blank canvas, *i.e.* , a state where all segmentation confidence scores converge to a certain value c (*i.e.* , $\Phi[i,j] \approx c,\ \forall i,j$). As described in Equation 1, a pixel x_{ij} is classified as masked if the predicted value y_{ij} is positive. The value $y_{ij} \in \Theta$ is derived from SAM and indicates the confidence that the position $\{i,j\}$ should be masked. Consequently, if the predicted values y across the images converge towards a typical constant, SAM interprets this image as a blank canvas. Specifically, our framework aims to identify such a perturbation δ:

$$\Phi' = \text{SAM}(x_{\text{clear}} + \delta, \mathcal{P}), \tag{3}$$

where x_{clear} represents the images requiring protection and $\mathcal{P}$ denotes the point prompt for SAM. The perturbation δ is designed to ensure that Φ' (the output of SAM) approaches a constant C. If every element in the matrix Φ' converges to this common constant, the entire image is perceived as a "blank canvas" from SAM's perspective.

To guarantee that SAM's output approaches the constant C, we employ the Mean Squared Error (MSE) loss, which is a natural fit for this optimization task. As articulated in Equation 4, the predicted value SAM(prompt, $x_{\text{clean}} + \delta$) is optimized to be close to a target threshold after the attack. Thus, we aim for the predicted values to equal the constant C, leading to the loss function defined as:

$$\mathcal{L}_{\text{attack}} = \texttt{MSE}(\Phi, C), \tag{4}$$

where Φ represents the output of SAM. The MSE loss in Equation 4 facilitates the prediction of negative values, aligning with our expectations for the perturbation δ. This loss term ensures that the predicted results remain close to the constant C, effectively rendering the images as a unified whole from SAM's perspective.

We can effectively obscure critical information that the model relies on for accurate segmentation by compelling SAM to perceive the protected images as blank canvases. However, we notice that the naive adversarial attack, *i.e.* , above $\mathcal{L}_{\text{attack}}$, is insufficient to deceive SAM, especially in high-frequency edges (further demonstrated in ablation study). To address this issue, we further introduce a ***Frequency-aware optimization***, which is designed to enhance adversarial perturbations through the frequency domain with adaptive spectral optimization. It comprises three synergistic components, designed to force adversarial perturbations that disrupt high-frequency areas while preserving the fidelity of low-frequency regions.

Wavelet-domain frequency decomposition. Firstly, we target the high-frequency perturbations that disrupt SAM's edge detection capabilities. We decompose the source image x and perturbed image $\tilde{x}$ into the frequency domain using the discrete wavelet transform (DWT) with the Daubechies-8 basis. This allows us to extract critical high-frequency components for disrupting SAM's edge perception. The loss can be computed as follows:

$$\mathcal{L}_{\text{hfc}} = \sum_{k=1}^{K} \| \underbrace{\mathcal{W}_k(\tilde{x}) \odot M_{\text{edge}}}_{\text{perturbed edges}} - \underbrace{\mathcal{W}_k(x) \odot M_{\text{edge}}}_{\text{original edges}} \|_F^2. \tag{5}$$

Here, $\mathcal{W}_k$ extracts wavelet coefficients at level k, and $\odot$ denotes the Hadamard product. The high-frequency perturbations are constrained by a Canny edge mask M_{edge}. This strategy aims to disrupt the high-frequency areas of the image to tame SAM fully for accurate tamper localization.

Structural preservation constraint. While we disrupt high-frequency information, the overall visual integrity of the image remains intact. This is crucial for avoiding detection by human observers or automated systems. We incorporate a structural preservation constraint that safeguards the low-frequency components using an adaptive Structural Similarity Index (SSIM) to maintain visual naturalness and prevent excessive distortion. The SSIM-based loss is defined as:

$$\mathcal{L}_{\text{lfc}} = \text{SSIM}(\phi_m, \tilde{\phi}_m), \tag{6}$$

Algorithm 1: Adaptive spectral optimization

Require: $\mathcal{L}_{\text{attack}}$, $\mathcal{L}_{\text{stealth}}$, Fourier transform $\mathcal{F}$, iterations T, step size α_0, perturbation bound ϵ, momentum decay $\mu \in [0, 1)$

Ensure: Adversarial perturbation δ

Initialize $\delta_0 \sim \mathcal{U}(-\epsilon, \epsilon)$
Initialize momentum: $m_0 = 0$
for t = 0 to T-1 **do**
  $g_t = \nabla_\delta(\mathcal{L}_{\text{attack}} + \mathcal{L}_{\text{stealth}})$ ▷ Compute gradient
  $\hat{g}_t = \mathcal{F}^{-1}(\mathcal{F}(g_t) \odot \mathcal{M})$ ▷ Spectral projection
  $m_{t+1} = \mu m_t + \frac{\hat{g}_t}{\|\hat{g}_t\|_1}$ ▷ Update momentum
  $\alpha_t = \alpha_0(1 - e^{-5t/T})$ ▷ Adaptive step
  $\delta_{t+1} = \text{Clip}_\epsilon[\delta_t + \alpha_t \cdot \text{sign}(m_{t+1})]$ ▷ Update δ
end for

where $\phi_m = \mathcal{L}_m^T x_{\text{lfc}}^{(m)} \mathcal{L}_m$, representing reconstructs the low frequency components at scale m. $\tilde{\phi}_m$. This component ensures that the low-frequency characteristics of the image are preserved, maintaining its natural appearance while still achieving the desired adversarial effect.

Adaptive spectral optimization. Finally, we introduce frequency-aware adversarial optimization with momentum, as detailed in algorithm 1. The motivation for this optimization approach is to derive the optimal perturbation δ^* that maximizes the effectiveness of the attack while adhering to constraints on perturbation magnitude. The optimization problem is formulated as:

$$\delta^* = \arg \max_{\|\delta\|_\infty \leq \epsilon} \underbrace{\mathbb{E}_{p\sim\mathcal{P}}[\|SAM(x+\delta, p)\|_1]}_{\mathcal{L}_{\text{attack}}} + \underbrace{\lambda\mathcal{L}_{\text{lfc}} - \beta\mathcal{L}_{\text{hfc}}}_{\mathcal{L}_{\text{stealth}}}. \tag{7}$$

In this context, $\mathcal{L}_{\text{attack}}$ combines mask prediction suppression and edge disruption, defined as follows:

$$\mathcal{L}_{\text{all}} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \|SAM(x+\delta, p_i)\|_2^2}_{\text{MSE suppression}} + \mathcal{L}_{\text{stealth}}. \tag{8}$$

The loss item $\mathcal{L}_{\text{attack}}$ is used to optimize the adversarial perturbation. We adaptively optimize the gradient based on the perturbation energy to maximize its impact on the high-frequency contents and minimize its impact on the low-frequency area. Specifically, a spectral projection mask $\mathcal{M}$ is designed to focus perturbation energy on high-frequency bands, defined as:

$$\mathcal{M}(u, v) = \begin{cases} 1, & \sqrt{u^2 + v^2} \geq f_{\text{cutoff}}, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

where f_{cutoff} acts as the cutoff frequency controlling the locality of the perturbation. u and v denote the frequency components in the frequency domain of the source image. We then use this map to control the adversarial optimization process via algorithm 1.

4.2 Tamper localization

Following the blank canvas initialization, the adversarial image $\tilde{x} = x + \delta$ must satisfy the condition:

$$\Phi_{\text{adv}}[i, j] = SAM(\tilde{x})[i, j] \approx C, \quad \forall i, j \in \{1, ..., H\} \times \{1, ..., W\}. \tag{10}$$

When tampering occurs on the blinded image $\tilde{x}$, the alteration Δx generates localized perturbation breaches, which can be captured as:

$$\mathcal{M}_{\text{tamper}} = \mathbb{I}\left(\|SAM(\tilde{x} + \Delta x)\|_2 > \tau_{\text{detect}}\right). \tag{11}$$

In this equation, τ_{detect} is adaptively determined using Otsu's method (Otsu et al. 1975). The tampered areas $\mathcal{M}_{\text{tamper}}$ are determined through evaluating the noticeable areas in the confidence map predicted by SAM.

4.3 Implementation details

To optimize the adversarial perturbations, we adopt the projected gradient descent (PGD) attack framework (Madry et al. 2018). In alignment with prior research focused on attacking vision models in a white-box setting, we establish the maximum allowable perturbation magnitude at $16/255$, with a step size of $2/255$ for PGD. The optimized target constant C is set to 15, as the predicted values in the background (non-mask) region typically hover around this value. For our experiments, we utilize the vanilla version of SAM (Kirillov et al. 2023) with ViT-H (Dosovitskiy et al. 2021) backbone and employ a single point prompt located at $(0, 0)$. This setup enables the prediction of a single mask in proximity to that point. Given that each mask is generated based on an input prompt, the attack is deemed successful if SAM fails to predict the original mask corresponding to that prompt accurately. Specifically, the attack succeeds in the mask removal task if Mask_{adv} is empty or if its area is at least as small as that of $\text{Mask}_{\text{clean}}$. To maintain generality, the $(\text{prompt}, \text{Mask})$ pair selected for the attack is randomly chosen from the image.

5 Experiments

Benchmarks. To comprehensively evaluate our method, we conduct experiments across four tamper localization dataset (Dong, Wang, and Tan 2013; Wen et al. 2016; Guan et al. 2019; Hsu and Chang 2006), comparing against eight state-of-the-art methods including SPAN (Hu et al. 2020), ManTraNet (Wu and et al. 2019), OSN (Wu et al. 2022), HiFi-Net (Guo et al. 2023), PSCC-Net (Liu et al. 2022), CAT-Net (Kwon et al. 2021), MVSS-Net (Dong et al. 2022), and FakeShield (Xu et al. 2025b). To simulate real-world forgery scenarios, we further evaluate on the AIGC-based editing dataset (Zhang et al. 2024). Editing methods, including SD Inpaint (Rombach et al. 2022), Controlnet (Zhang, Rao, and Agrawala 2023), SDXL (Podell et al. 2024) are used to manipulate images with the prompt to be "None". For fair comparison with EditGuard (Zhang et al. 2024) - the current state-of-the-art in proactive protection, the images are protected in advance before tampering. Following (Dong et al. 2022; Guillaro et al. 2023; Liu et al. 2022; Zhang et al. 2024),

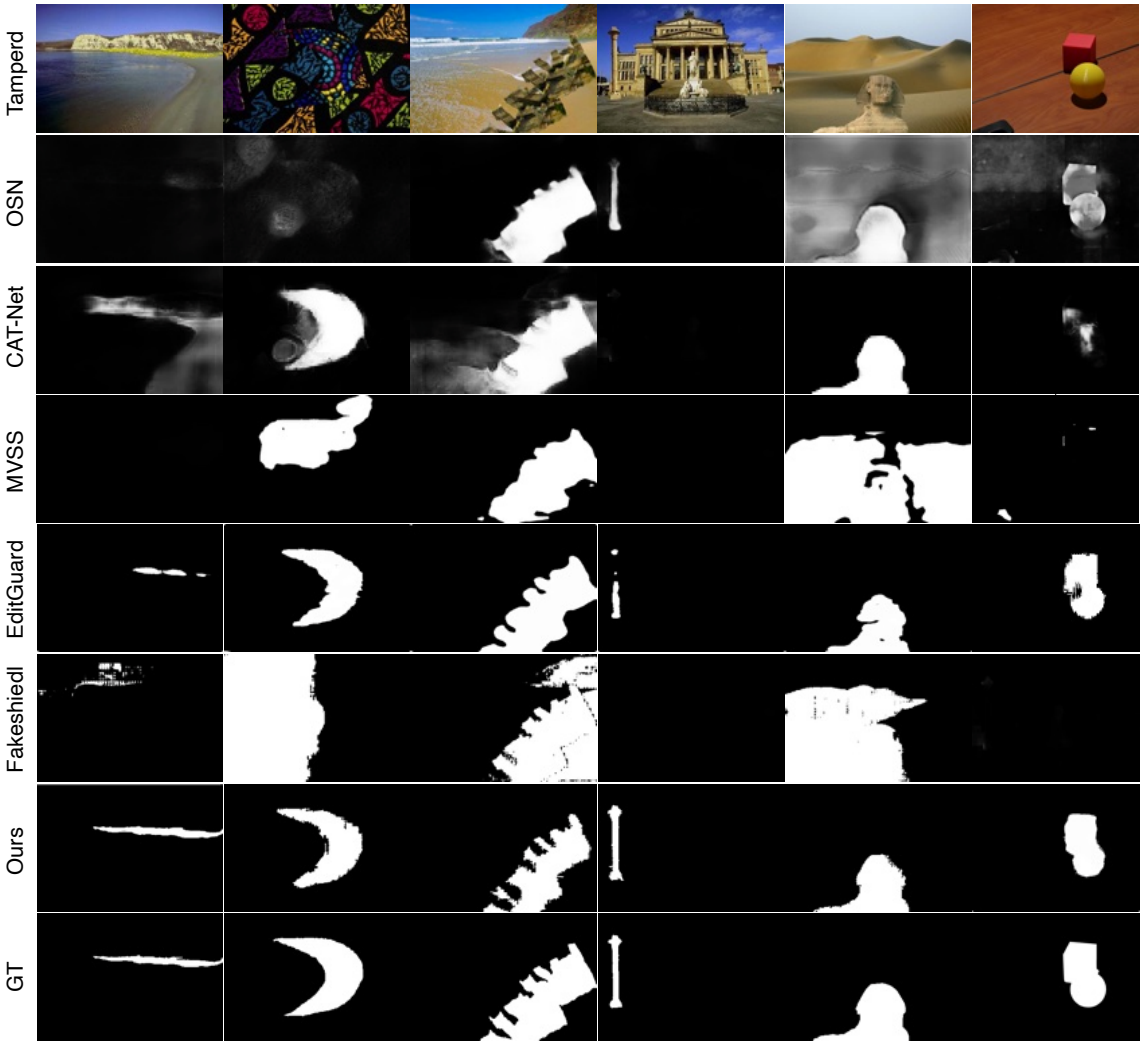

Figure 4: Visual results of localized tampering areas. Our method achieves superior performance compared to previous passive tamper localization methods, including OSN (Wu et al. 2022), CAT-Net (Kwon et al. 2021), MVSS (Dong et al. 2022), and FakeShield (Xu et al. 2025b). We also achieve comparable performance compared to protective methods EditGuard (Zhang et al. 2024) across each scene.

F1-score, IoU are used to evaluate the quality of tamper localization. Higher F1-score and IoU indicate better localization performance.

5.1 Experimental results

Results on classical benchmarks. For a comprehensive comparison with existing tamper localization methods, we conduct evaluations using classical benchmarks (Dong, Wang, and Tan 2013; Wen et al. 2016; Guan et al. 2019; Hsu and Chang 2006), as summarized in Tab. 1. Our method demonstrates competitive performance against the previous state-of-the-art method EditGuard (Zhang et al. 2024). In several cases, we achieve marginally better localization accuracy, surpassing it by small margins in F1-score across various datasets. respectively. This indicates the effectiveness of our proactive localization mechanism, which does not require any labeled data or tampered samples. As depicted in Fig. 4, our method effectively identifies pixel-level tampered areas, similar to EditGuard, while other methods tend to produce only rough outlines or show effectiveness in limited scenarios. This highlights the robustness and reliability of our approach in the realm of tamper localization.

Results on AIGC-based editing methods. We conduct experiments on various AIGC-based image editing methods using the AGE-Set dataset (Zhang et al. 2024), following the evaluation framework established by EditGuard (Zhang et al. 2024). The comparison of our method with several state-of-the-art tamper localization techniques is presented in Tab. 2. We observe that the F1-scores of existing passive forensic methods are generally below 0.7 when applied to AIGC-edited images. Notably, even after fine-tuning MVSS-Net on the constructed AIGC-edited dataset, the performance of MVSS-Net† remains inadequate, exhibiting significant limitations in handling various tampering techniques. In contrast, our method consistently achieves F1 Scores and AUCs exceeding 95%, while maintaining an IoU of around 90%

Method	CASIA1+		IMD2020		Columbia		NIST		DSO		Korus	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
SPAN (Hu et al. 2020)	0.11	0.14	0.09	0.14	0.14	0.20	0.16	0.21	0.14	0.24	0.06	0.10
ManTraNet (Wu and et al. 2019)	0.09	0.13	0.10	0.16	0.04	0.07	0.14	0.20	0.08	0.13	0.02	0.05
OSN (Wu et al. 2022)	0.47	0.51	0.38	0.47	0.58	0.69	0.25	0.33	0.32	0.45	0.14	0.19
HiFi-Net (Guo et al. 2023)	0.13	0.18	0.09	0.14	0.06	0.11	0.09	0.13	0.18	0.29	0.01	0.02
PSCC-Net (Liu et al. 2022)	0.36	0.46	0.22	0.32	0.64	0.74	0.18	0.26	0.22	0.33	0.15	0.22
CAT-Net (Kwon et al. 2021)	0.44	0.51	0.14	0.19	0.08	0.13	0.14	0.19	0.06	0.10	0.04	0.06
MVSS-Net (Dong et al. 2022)	0.40	0.48	0.23	0.31	0.48	0.61	0.24	0.29	0.23	0.34	0.12	0.17
FakeShield (Xu et al. 2025b)	0.56	0.62	0.52	0.58	0.68	0.76	0.34	0.39	0.50	0.54	0.22	0.26
EditGuard (Zhang et al. 2024)	0.60	**0.67**	0.55	0.62	0.70	0.78	**0.35**	0.40	0.52	0.56	0.22	0.28
Ours	**0.62**	**0.67**	**0.58**	**0.66**	**0.74**	**0.81**	0.31	**0.45**	**0.55**	**0.60**	**0.27**	**0.31**

Table 1: Comparison with other tamper localization methods on the classical tamper localization dataset.

Method	SD Inpaint		ControlNet		SDXL		RePaint		Lama	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
MVSS-Net (Dong et al. 2022)	0.178	0.103	0.178	0.103	0.037	0.028	0.104	0.082	0.024	0.022
PSCC-Net (Liu et al. 2022)	0.166	0.112	0.177	0.116	0.189	0.115	0.140	0.109	0.132	0.104
HiFi-Net (Guo et al. 2023)	0.547	0.128	0.542	0.123	0.828	0.261	0.681	0.339	0.483	0.029
MVSS-Net† (Dong et al. 2022)	0.694	0.575	0.678	0.558	0.482	0.359	0.185	0.111	0.393	0.275
EditGuard (Zhang et al. 2024)	0.966	0.936	0.968	**0.940**	0.965	0.936	**0.967**	0.938	**0.965**	0.934
Ours	**0.972**	**0.958**	**0.973**	0.938	**0.970**	**0.958**	0.961	**0.957**	0.954	**0.951**

Table 2: Comparison with other competitive tamper localisation methods under different AIGC-based editing methods. Note that † denotes the network fine-tuned with the AIGC-edited dataset (Zhang et al. 2024).

Figure 5: Ablation study on the proposed frequency-aware optimization. False positive results could occur in high-frequency areas as they are more likely to be noticed by SAM. The red/blue box indicates the false/true positives of localized regions.

Case	Degradation*	$\mathcal{L}_{\text{mse}}$	$\mathcal{L}_{\text{stealth}}$	`Ada.`	F1	IoU
(a)	Clean				0.352	0.378
(b)	Clean	✓	✓		0.934	0.928
(c)	Clean	✓		✓	0.931	0.921
Ours	Clean	✓	✓	✓	0.964	0.955
	Random	✓	✓	✓	0.945	0.931

Table 3: Our frequency-aware optimization, *i.e.* , combing $\mathcal{L}_{\text{stealth}}$ with `Ada.`, achieves the best results as it could fully disrupt the perception feature of SAM. Case (a) denotes that the images are not being protected with any adversarial perturbations.

across different tampering types. Our approach effectively captures subtle tampering traces introduced by AIGC-based editing methods, whereas other competing methods struggle to produce meaningful results.

Ablation study. To evaluate the contribution of each component in our proposed method, we conduct ablation studies on our frequency-aware perturbation (Tab. 3), visualized in Fig. 5, as well as the mean squared error loss ($\mathcal{L}_{\text{mse}}$) and stealth loss ($\mathcal{L}_{\text{stealth}}$). We observe that omitting frequency-aware optimization results in substantial performance degradation, reducing effectiveness to the level of unprotected images. In contrast, incorporating $\mathcal{L}_{\text{mse}}$ markedly improves results, with further gains from adding $\mathcal{L}_{\text{stealth}}$, underscoring their essential roles in disrupting perception features and boosting tampering-detection accuracy. Beyond these core components, supplemental experiments affirm robustness under both clean conditions and random degradations, while additional validation across diverse SAM variants demonstrates consistent effectiveness (please see supplement).

6 Conclusion

This paper introduces a novel proactive framework for tamper localization by transforming images into machine-interpretable "blank canvases" via frequency-optimized adversarial perturbations applied to vision foundation models. By suppressing the model's perception of original content while amplifying its sensitivity to synthetic alterations, our method addresses the limitations of conventional passive forensic approaches that rely on artifact detection or tampered training data. Our method shifts from post-hoc analysis to proactive protection and demonstrates strong generalizability across diverse manipulation scenarios, offering a scalable solution for real-world image authentication.

Acknowledgments

This work was carried out at the Renjie Group, Hong Kong Baptist University. Renjie Group is supported by the National Natural Science Foundation of China under Grant No. 62302415, Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515012822, and Research Grant Council (RGC) of Hong Kong SAR, under a GRF Grant 12203124 and an ECS Grant 22201125.

References

Arnab, A.; Miksik, O.; and Torr, P. H. 2018. On the robustness of semantic segmentation models to adversarial attacks. In *CVPR*.

Asnani, V.; Yin, X.; Hassner, T.; and Liu, X. 2023. Malp: Manipulation localization using a proactive scheme. In *CVPR*.

Bhojanapalli, S.; Chakrabarti, A.; Glasner, D.; Li, D.; Unterthiner, T.; and Veit, A. 2021. Understanding robustness of transformers for image classification. In *ICCV*.

Chen, X.; Dong, C.; Ji, J.; Cao, J.; and Li, X. 2021. Image manipulation detection by multi-view multi-scale supervision. In *ICCV*.

Dong, C.; Chen, X.; Hu, R.; Cao, J.; and Li, X. 2022. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *TPAMI*.

Dong, J.; Wang, W.; and Tan, T. 2013. Casia image tampering detection evaluation database. In *ChinaSIP*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *ICLR*.

Guan, H.; Kozak, M.; Robertson, E.; Lee, Y.; Yates, A. N.; Delgado, A.; Zhou, D.; Kheyrkhah, T.; Smith, J.; and Fiscus, J. 2019. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *WACVW*.

Guillaro, F.; Cozzolino, D.; Sud, A.; Dufour, N.; and Verdoliva, L. 2023. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *CVPR*.

Guo, X.; Liu, X.; Ren, Z.; Grosz, S.; Masi, I.; and Liu, X. 2023. Hierarchical fine-grained image forgery detection and localization. In *CVPR*.

Haralick, R. M.; and Shapiro, L. G. 1985. Image segmentation techniques. *CVGIP*.

Hendrik Metzen, J.; Chaithanya Kumar, M.; Brox, T.; and Fischer, V. 2017. Universal adversarial perturbations against semantic image segmentation. In *ICCV*.

Hsu, Y.-F.; and Chang, S.-F. 2006. Detecting image splicing using geometry invariants and camera characteristics consistency. In *ICME*.

Hu, X.; Ying, Q.; Qian, Z.; Li, S.; and Zhang, X. 2023. DRAW: Defending Camera-shooted RAW against Image Manipulation. In *ICCV*.

Hu, X.; Zhang, Z.; Jiang, Z.; Chaudhuri, S.; Yang, Z.; and Nevatia, R. 2020. SPAN: Spatial pyramid attention network for image manipulation localization. In *ECCV*.

Huang, X.; Cheung, K. C.; See, S.; and Wan, R. 2024a. Geometrysticker: Enabling ownership claim of recolorized neural radiance fields. In *ECCV*.

Huang, X.; Li, R.; Cheung, Y.-m.; Cheung, K. C.; See, S.; and Wan, R. 2024b. Gaussianmarker: Uncertainty-aware copyright protection of 3D gaussian splatting. In *NeurIPS*.

Huang, X.; Luo, Z.; Song, Q.; Wang, R.; and Wan, R. 2025. MarkSplatter: Generalizable watermarking for 3D gaussian splatting model via splatter image structure. In *ACM MM*.

Islam, A.; Long, C.; Basharat, A.; and Hoogs, A. 2020. Doa-gan: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization. In *CVPR*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2017. Adversarial machine learning at scale. *ICLR*.

Kwon, M.-J.; Lee, W.; Nam, S.-H.; Son, M.; and Kim, C. 2025. SAFIRE: Segment Any Forged Image Region. In *AAAI*.

Kwon, M.-J.; Yu, I.-J.; Nam, S.-H.; and Lee, H.-K. 2021. CAT-Net: Compression artifact tracing network for detection and localization of image splicing. In *WACV*.

Li, H.; and Huang, J. 2019. Localization of deep inpainting using high-pass fully convolutional network. In *ICCV*.

Li, R.; and Cheung, Y.-m. 2024. Variational multi-scale representation for estimating uncertainty in 3D gaussian splatting. In *NeurIPS*.

Li, R.; and Cheung, Y.-m. 2025. Modeling and Identifying Distractors with Curriculum for Robust 3D Gaussian Splatting. In *ACM MM*.

Li, Y.; Liu, D.; Li, H.; Li, L.; Li, Z.; and Wu, F. 2018. Learning a convolutional neural network for image compact-resolution. *TIP*.

Li, Y.; and Zhou, J. 2018. Fast and effective image copy-move forgery detection via hierarchical feature point matching. *TIFS*.

Liu, X.; Liu, Y.; Chen, J.; and Liu, X. 2022. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *TCSVT*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*.

Luo, Z.; Guo, Q.; Cheung, K. C.; See, S.; and Wan, R. 2023. Copyrnerf: Protecting the copyright of neural radiance fields. In *ICCV*.

Luo, Z.; Liu, J.; Li, H.; Rocha, A.; and Wan, R. 2025a. MantleMark: Migrating Watermarks from Multi-View Images to Radiance Fields via Frequency Modulation. *Authorea Preprints*.

Luo, Z.; Rocha, A.; Shi, B.; Guo, Q.; Li, H.; and Wan, R. 2025b. The nerf signature: Codebook-aided watermarking for neural radiance fields. *TPAMI*.

Luo, Z.; Zhao, Y.; Cheung, K. C.; See, S.; and Wan, R. 2025c. ImageSentinel: Protecting Visual Datasets from Unauthorized Retrieval-Augmented Image Generation. In *NeurIPS*.

Ma, X.; Du, B.; Liu, X.; Hammadi, A. Y. A.; and Zhou, J. 2023. IML-ViT: Image Manipulation Localization by Vision Transformer. *arXiv preprint arXiv:2307.14863*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. *ICLR*.

Mahmood, K.; Mahmood, R.; and Van Dijk, M. 2021. On the robustness of vision transformers to adversarial examples. In *ICCV*.

Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; and Terzopoulos, D. 2021. Image segmentation using deep learning: A survey. *TPAMI*.

Otsu, N.; et al. 1975. A threshold selection method from gray-level histograms. *Automatica*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. Sdxl: improving latent diffusion models for high-resolution image synthesis. In *ICLR*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Salloum, R.; Ren, Y.; and Kuo, C.-C. J. 2018. Image splicing localization using a multi-task fully convolutional network (MFCN). *J VIS COMMUN IMAGE R*.

Shen, Y.; Li, Z.; and Wang, G. 2024. Practical Region-level Attack against Segment Anything Models. In *CVPR*.

Song, Q.; Luo, Z.; Cheung, K. C.; See, S.; and Wan, R. 2024a. Geometry cloak: Preventing tgs-based 3D reconstruction from copyrighted images. In *NeurIPS*.

Song, Q.; Luo, Z.; Cheung, K. C.; See, S.; and Wan, R. 2024b. Protecting nerfs' copyright via plug-and-play watermarking base model. In *ECCV*.

Sun, Z.; Fang, H.; Cao, J.; Zhao, X.; and Wang, D. 2024. Rethinking image editing detection in the era of generative ai revolution. In *ACM MM*.

Sun, Z.; Jiang, H.; Wang, D.; Li, X.; and Cao, J. 2023. SAFL-Net: Semantic-Agnostic Feature Learning Network with Auxiliary Plugins for Image Manipulation Detection. In *ICCV*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Wen, B.; Zhu, Y.; Subramanian, R.; Ng, T.-T.; Shen, X.; and Winkler, S. 2016. COVERAGE—A novel database for copy-move forgery detection. In *ICIP*.

Wu, H.; Xiang, S.; Gabriben; and Niczem. 2020. FaceSwap. https://github.com/wuhuikai/FaceSwap.

Wu, H.; Zhou, J.; Tian, J.; and Liu, J. 2022. Robust image forgery detection over online social network shared images. In *CVPR*.

Wu, Y.; and et al. 2019. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *CVPR*.

Xia, S.; Yang, W.; Yu, Y.; Lin, X.; Ding, H.; DUAN, L.; and Jiang, X. 2023. Transferable Adversarial Attacks on SAM and Its Downstream Models. In *NeurIPS*.

Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; and Yuille, A. 2017. Adversarial examples for semantic segmentation and object detection. In *ICCV*.

Xu, Y.; Yin, Y.; Xing, Y.; and Chen, Y. 2025a. From Policy Comparison to Process Consistency and Beyond. In *ACM CIKM*.

Xu, Z.; Zhang, X.; Li, R.; Tang, Z.; Huang, Q.; and Zhang, J. 2025b. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *ICLR*.

Ying, Q.; Hu, X.; Zhang, X.; Qian, Z.; Li, S.; and Zhang, X. 2022. RWN: Robust Watermarking Network for Image Cropping Localization. In *ICIP*.

Ying, Q.; Qian, Z.; Zhou, H.; Xu, H.; Zhang, X.; and Li, S. 2021. From image to imuge: Immunized image generation. In *ACM MM*.

Ying, Q.; Zhou, H.; Qian, Z.; Li, S.; and Zhang, X. 2023. Learning to Immunize Images for Tamper Localization and Self-Recovery. *TPAMI*.

Zhang, C.; Zhang, C.; Kang, T.; Kim, D.; Bae, S.-H.; and Kweon, I. S. 2023. Attack-SAM: Towards attacking segment anything model with adversarial examples. *arXiv preprint arXiv:2305.00866*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Zhang, X.; Li, R.; Yu, J.; Xu, Y.; Li, W.; and Zhang, J. 2024. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *CVPR*.

Zhang, X.; Tang, Z.; Xu, Z.; Li, R.; Xu, Y.; Chen, B.; Gao, F.; and Zhang, J. 2025. Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking. In *CVPR*.

Zhou, Z.; Song, Y.; Li, M.; Hu, S.; Wang, X.; Zhang, L. Y.; Yao, D.; and Jin, H. 2024. Darksam: Fooling segment anything model to segment nothing. *NeurIPS*.

Zhu, X.; Qian, Y.; Zhao, X.; Sun, B.; and Sun, Y. 2018. A deep learning approach to patch-based image inpainting forensics. *SIGNAL PROCESS-IMAGE*.