Removing Box-Free Watermarks for Image-to-Image Models via Query-Based Reverse Engineering

Haonan An[1], Guang Hua[2], Hangcheng Cao[1*], Zhengru Fang[1], Guowen Xu[3], Susanto Rahardja[2], Yuguang Fang[1]

[1]Department of Computer Science, City University of Hong Kong, Hong Kong
[2]Infocomm Technology and Engineering Cluster, Singapore Institute of Technology, Singapore 828608
[3]School of Computer Science and Engineering, University of Electronic Science and Technology of China
{haonanan2-c, zhefang4-c}@my.cityu.edu.hk, {hangccao, my.fang}@cityu.edu.hk,
{ghua, susantorahardja}@ieee.org, guowen.xu@uestc.edu.cn

Abstract

The intellectual property of deep generative networks (GNets) can be protected using a cascaded hiding network (HNet) which embeds watermarks (or marks) into GNet outputs, known as box-free watermarking. Although both GNet and HNet are encapsulated in a black box (called operation network, or ONet), with only the generated and marked outputs from HNet being released to end users and deemed secure, in this paper, we reveal an overlooked vulnerability in such systems. Specifically, we show that the hidden GNet outputs can still be reliably estimated via query-based reverse engineering, leaking the generated and unmarked images, despite the attacker's limited knowledge of the system. Our first attempt is to reverse-engineer an inverse model for HNet under the stringent black-box condition, for which we propose to exploit the query process with specially curated input images. While effective, this method yields unsatisfactory image quality. To improve this, we subsequently propose an alternative method leveraging the equivalent additive property of box-free model watermarking and reverse-engineering a forward surrogate model of HNet, with better image quality preservation. Extensive experimental results on image processing and image generation tasks demonstrate that both attacks achieve impressive watermark removal success rates (100%) while also maintaining excellent image quality (reaching the highest PSNR of 34.69 dB), substantially outperforming existing attacks, highlighting the urgent need for robust defensive strategies to mitigate the identified vulnerability in box-free model watermarking.

Introduction

Recently, deep neural networks (DNNs), especially generative networks (GNets), have demonstrated their powerful ability to handle various tasks, surpassing previous state-of-the-art techniques. However, the resources required to train such models, whether in terms of time, money, or labor, are immense. For example, the widely used generative DNN application GPT-4 requires more than 24,000 GPUs for training (Liu et al. 2023b), resulting in significant expenses. Therefore, it is imperative to safeguard these assets from intellectual property infringement.

*Hangcheng Cao is the corresponding author.

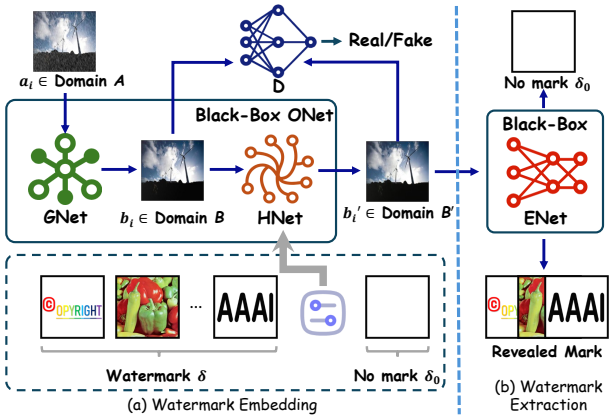

Figure 1: Flowchart of the victim model considered in this paper, where image denoising is used as an example.

Previous efforts in safeguarding the intellectual property of DNN models aim at two primary goals: 1) Ownership verification and 2) Model stealing tracing, also known as surrogate attack tracing. Both goals can be accomplished through watermarking techniques. The watermarking process embeds marks, e.g., encoded identity information, into the to-be-protected model or its outputs. Subsequently, a verification process can extract the embedded information to verify the ownership. In the case of model stealing tracing, model watermarking has shown its effectiveness in retaining the embedded marks in surrogate models, which can then be extracted through the verification process.

Among all categories of model watermarking, box-free watermarking stands out due to its ability to handle high-entropy outputs (e.g., images) and its flexibility in extracting watermarks solely from protected models' outputs. In this subdomain, Zhang et al. (Zhang et al. 2020) proposed the first box-free framework for image processing models. Although existing box-free methods have shown robustness against common image processing, surrogate attacks, and applicability to other generative tasks (Wu et al. 2020; Zhang et al. 2021; Huang et al. 2023; Zhang et al. 2024; An et al. 2025), in this paper, we reveal an overlooked vulnerability that query-based reverse engineering can leak the generated

and unmarked images, enabling watermark removal.

We build on the prior works (Wu et al. 2020; Zhang et al. 2020, 2021, 2024) as depicted in Figure 1. Considering image denoising as an example, the noisy image, denoted by $a_i \in A$, is processed by GNet, whose output, denoised image $b_i \in B$, is not released. Instead, b_i is further processed by the hiding network (HNet) for watermark embedding, resulting in a processed and marked image, $b'_i \in B'$, which is the actual output of the model. Notably, GNet and HNet are encapsulated as a black-box called operation network (ONet). The extraction network (ENet) takes b'_i as input to retrieve the embedded watermark, while an all-white image indicates the absence of watermarks.

Our attacks are grounded in a key observation that if an attacker can craft queries to circumvent the functionality of GNet, they may extract HNet's information to infer the watermark embedding mechanism, thereby enabling watermark removal against b'_i. Based on this, we first present a simple and intuitive method to reverse-engineer an inversion network of the HNet, denoted by HNet^{-1}, which is a watermark removal operator to b'_i. However, such an inversion-based removal attack resulted in limited image quality, which motivates us to propose the second attack. We notice that the watermark embedding process in the victim model in Figure 1 is additive in nature, which can be modeled by $b'_i = b_i + \delta'_{bi}$, where δ'_{bi} is the equivalent representation of the to-be-embedded watermark δ. Built on this, we demonstrate how the additive property can be exploited to estimate the private b_i. Both attacks are evaluated under the realistic and challenging black-box condition, where the attacker only has access to the victim model's inputs (a_i) and outputs (b'_i). Our contributions are as follows:

- We discover, via query-based reverse engineering, that encapsulating GNet and HNet into a black box is still insecure in protecting GNet and its outputs.
- We show that HNet can be inverted via specially crafted queries, which leads to watermark removal and leaking of GNet outputs.
- We further propose an improved remover leveraging on the additive equivalence of the watermark, which achieves better image quality.
- We discuss a query screener to be deployed at the API to defend against the proposed attacks.
- We conduct extensive experiments on image processing and generation tasks to demonstrate the effectiveness of our proposed methods.

A comprehensive review of related work is provided in the supplementary material.

Problem Formulation

Box-free Model Watermarking Basics

The existing shared box-free model watermarking workflow is depicted in Figure 1, and the related notations are summarized in Table 1. The workflow is modeled as

- Image operation: $b_i = \text{GNet}(a_i)$,
- Watermark embedding: $b'_i = \text{HNet}(\text{Concat}(b_i, \delta))$,

Notation	Definition
i, j	Sample index, $i \neq j$
δ_0	All-white image (no watermark)
δ	Watermark image
δ'_{ai}	Latent representation of δ for a_i
δ'_{bi}	Latent representation of δ for b_i
$a_i \in A$	To-be-processed image (input of GNet)
$b_i \in B$	Processed unmarked (by GNet) image
$b'_i \in B'$	Processed marked (by ONet) image
$\hat{b}_i \in \hat{B}$	Watermark removed b'_i
$e_i \in E$	Generic unmarked image ($B \subset E$)
$b_j \in B$	Processed unmarked (not by GNet) image
$b'_j \in B'$	b_j marked by HNet

Table 1: List of main notations.

- Watermark extraction: $\hat{\delta} = \text{ENet}(b'_i)$,

where GNet can be deraining, image generation, or an arbitrary image-to-image model, $\text{Concat}()$ is the channel-wise concatenation operation, and the extraction yields an estimated watermark $\hat{\delta}$ which is supposed to be δ. Note that ENet can also take an unmarked b_i or an arbitrary image, e_i, as input, and the output is expected to be δ_0. To protect GNet, it is pretrained and frozen, while the defender jointly trains HNet, D, and ENet, to minimize the combined loss

$$\mathcal{L}_{\text{Joint}} = \beta_1 \mathcal{L}_{\text{Fidelity}} + \beta_2 \mathcal{L}_{\text{Mark}} + \beta_3 \mathcal{L}_{\text{Adv}}, \tag{1}$$

where

$$\mathcal{L}_{\text{Fidelity}} = \sum_i \text{MSE}\left(b'_i, b_i\right), \tag{2}$$

$$\mathcal{L}_{\text{Mark}} = \sum_i \left[\text{MSE}(\hat{\delta}, \delta) + \text{MSE}\left(\text{ENet}(e_i), \delta_0\right)\right], \tag{3}$$

$$\mathcal{L}_{\text{Adv}} = \sum_i \left[\log(\text{D}(b_i)) + \log(1 - \text{D}(b'_i))\right]. \tag{4}$$

In the above framework, $\mathcal{L}_{\text{Fidelity}}$ ensures the marked image is visually indistinguishable from the processed unmarked image, $\mathcal{L}_{\text{Mark}}$ ensures successful watermark extraction from marked images b'_i as well as successful null extraction (all-white output) from unmarked images e_i which can be the processed unmarked b_i or any images not related to GNet. Additionally, the adversarial loss $\mathcal{L}_{\text{Adv}}$ improves the embedding quality so the discriminator cannot distinguish marked and unmarked images. Other loss functions, e.g., consistency loss (Zhang et al. 2020) and perceptual loss (Johnson, Alahi, and Fei-Fei 2016), can be further added to improve the performance.

Threat Model

We consider a stringent yet practical threat model in this paper, in which GNet and HNet are encapsulated together as a black-box API, referred to as operation network (ONet), and deployed as a cloud service. End users can thus only have their query image a_i and observe the processed and marked image b'_i. Meanwhile, ENet is assumed to be operated by an authorized party for watermark extraction and verification,

and it is inaccessible to end users including attackers. The goal of the attack is to remove the watermark embedded in ONet output without significantly degrading image quality, allowing for further attacks like watermark-free surrogate model training. That is to say, ideally, the attacker aims to restore from the observed b'_i to b_i, removing the watermark while preserving the effect of GNet. The generic attacking model is thus given by

$$\text{Remove}(a_i, b'_i) = b_i. \tag{5}$$

Note that $\text{ENet}(b_i) = \delta_0$ corresponds to the ideal null extraction, while the removal attack may also be considered successful if $\text{ENet}(\text{Remove}(a_i, b'_i))$ significantly differs from δ. The successful estimation of b_i from the only observable a_i and b'_i reveals the vulnerability from the black-box ONet, which is analyzed in the next section.

Proposed Methods

In this section, we first reveal the overlooked vulnerability in box-free model watermarking. Building on this, we propose query-based reverse engineering attacks that efficiently remove the embedded watermark. The first attack trains an inversion network of HNet to remove the embedded watermark with simple and intuitive insight. However, we observe the unsatisfactory image quality preservation with this method. To address this, we propose the second attack based on our observation of the additive equivalence property in box-free model watermarking, which demonstrates significantly improved performance in maintaining output image quality. Last, we summarize the proposed attacks and discuss a potential defensive mechanism.

Vulnerability Analysis

From an attacker's perspective, GNet and HNet appear tightly coupled in ONet because the intermediate output b_i is unobtainable. Therefore, breaching this coupling to extract information from either model becomes a critical step in compromising the watermarking system.

Since ONet is deployed as a commercial cloud service, its functionality is typically known to both users and potential attackers. Our attacks are based on a key observation: if the attacker can construct inputs $b_j \in B$, $j \neq i$ that satisfy approximate identity transformation under GNet, i.e.,

$$\text{GNet}(b_j) \approx b_j, \tag{6}$$

then GNet can be effectively bypassed, thereby exposing the watermarking mechanism implemented by HNet. For example, when GNet is instantiated as a deraining model, b_j can be images captured in rain-free environments. When GNet functions as an image generation model, such as Stable Diffusion (Rombach et al. 2022), b_j can be masked images where only a small patch is provided for operation. Building on this insight, we develop two attack strategies described in the following sections.

First Attack: Inversion of HNet

The query process with b_j as inputs can be expressed as

$$\text{ONet}(b_j, \delta) = \text{HNet}(\text{GNet}(b_j), \delta) \approx \text{HNet}(b_j, \delta) = b'_j, \tag{7}$$

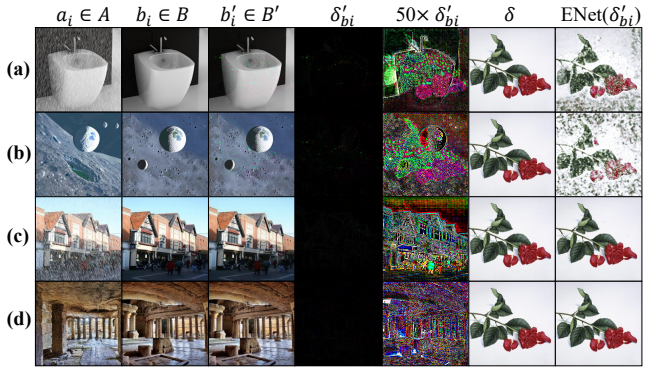

Figure 2: Demonstration of watermark extraction using the difference between b'_i and b_i as the input to ENet to verify the additive property. (a) Deraining with box-free model watermarking in (Wu et al. 2020). (b) Text-image-to-image-based image editing with box-free model watermarking in (Wu et al. 2020) and prompt "Emphasized planet appearance". (c) Deraining with box-free model watermarking in (Zhang et al. 2024). (d) Text-image-to-image-based image editing with box-free model watermarking in (Zhang et al. 2024) and prompt "Roman bath ruins".

where the approximation is based on (6). With a curated set of b_j and b'_j, we are able to train a surrogate model of the inversion of HNet, denoted as HNet^{-1}, to efficiently conduct watermark removal by minimizing following removal loss

$$\mathcal{L}_{\text{Removal}} = \sum_j \text{MSE}\left(\text{HNet}^{-1}(b'_j), b_j\right). \tag{8}$$

The number of queries needed to achieve a well-trained HNet^{-1} is within acceptable range, as shown in our experiments. However, an inferior performance in image quality for $\hat{b}_i = \text{HNet}^{-1}(b'_i)$ is observed when compared to b_i, which could negatively impact the performance of further attacks. One hypothesis is that the inherent high non-linearity of the DNN model renders the training of an optimal inverse network exceedingly challenging, particularly under conditions where the distribution of training data varies and the architecture of the HNet remains unknown to the attacker. In the next section, we introduce our second attack, which trains a surrogate model of HNet rather than its inversion. This method exploits the additive equivalent property observed in box-free model watermarking to directly estimate b_i, resulting in substantial image quality improvement.

Second Attack: Forward HNet

The Additive Equivalent We note that the watermark embedding process is a nonlinear high-dimensional mapping from the unmarked b_i to the marked b'_i governed by HNet, but it can also be equivalently expressed as an additive form in which the additive expression of watermark is the residual between b'_i and b_i, i.e.,

$$b'_i = b_i + \delta'_{bi}, \tag{9}$$

where δ'_{bi} is the additive equivalent of the watermark δ embedded into b_i, and in our considered models, δ'_{bi} is the la-

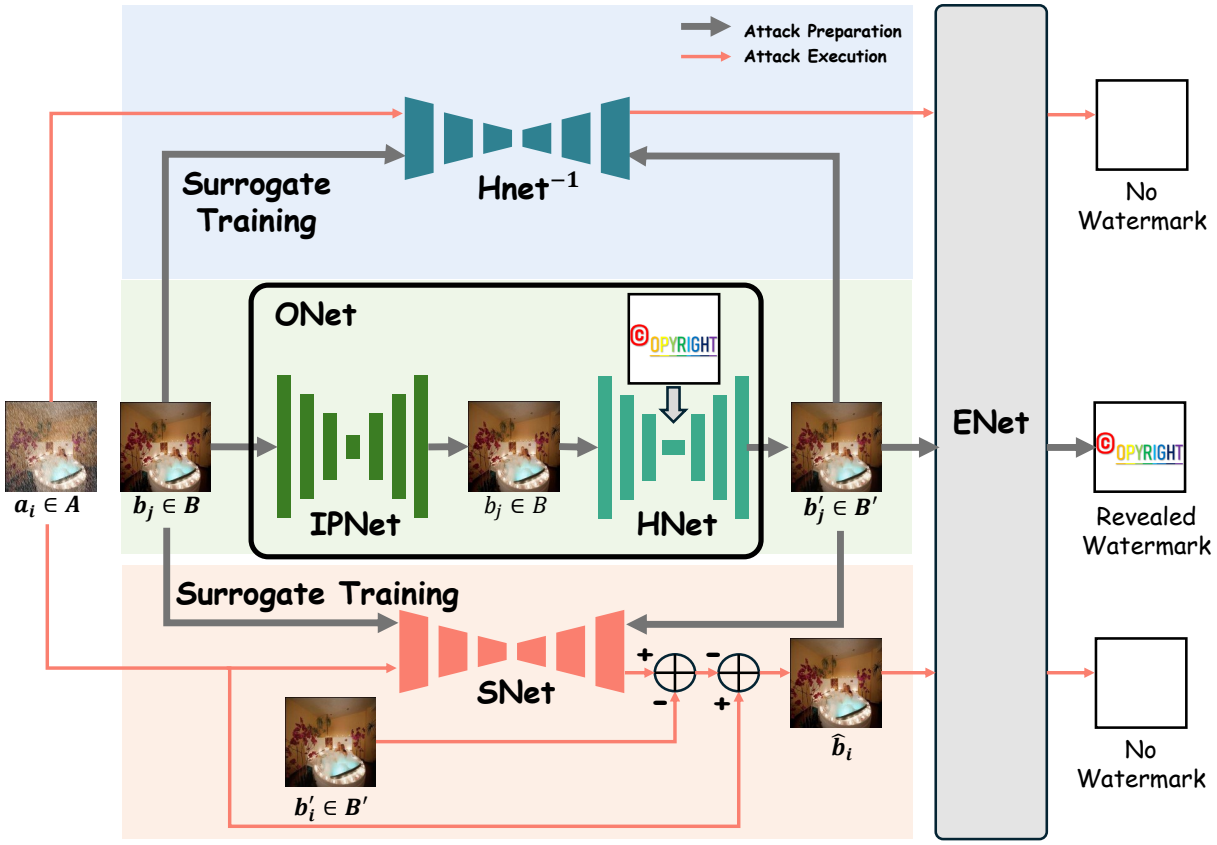

Figure 3: Workflow of victim model and the proposed attacks. Images in Domain B, which are watermark-free, can be estimated from all known quantities, showing watermark-free (all-white) outputs from ENet.

tent representation of δ. The justification for (9) stems from the first-order Taylor expansion of $\mathrm{HNet}(\mathrm{b_i}, \delta)$ around origin point of δ

$$\begin{aligned} &\mathrm{HNet}(b_i, \delta) \approx \mathrm{HNet}(b_i, 0) + \nabla_\delta \mathrm{HNet}(b_i, 0) \cdot \delta \\ &\Rightarrow b'_i - b_i \approx \nabla_\delta \mathrm{HNet}(b_i, 0) \cdot \delta = \delta'_{bi}, \end{aligned} \tag{10}$$

where $\mathrm{Concat}()$ is omitted for brevity in (10). It reveals that δ'_{bi} can be approximated as a certain transformation employed to watermark δ by HNet, i.e., the latent representation of δ. It then holds that

$$\begin{aligned} \delta'_{bi} &= b'_i - b_i \\ &= \mathrm{HNet}(\mathrm{Concat}(b_i, \delta)) - b_i. \end{aligned} \tag{11}$$

We can verify the above analysis by extracting the watermark using the residual as input to ENet and check if it holds that

$$\mathrm{ENet}(\delta'_{bi}) = \delta. \tag{12}$$

The verification results are shown in Figure 2. It can be seen from the δ'_{bi} column, or more clearly observed in its $50\times$ amplified visualization, that the residual keeps the texture information about the image content but has most semantic information lost. However, it is successfully verified that the watermark can still be reliably extracted from δ'_{bi}. This indicates that ENet can not only extract the watermark from marked images, i.e., $\mathrm{ENet}(b'_i) = \delta$, but also restore from the watermark's latent representation to its original image form as shown in (12), in the absence of attacks.

Estimate of b_i We now demonstrate that the additive nature of box-free watermarking can be exploited to expose the private information about b_i which should not be released in the black-box threat model. Replace b_i in (11) by a_i and according to (9), we have

$$\begin{aligned} \mathrm{HNet}(\mathrm{Concat}(a_i, \delta)) &= a_i + \delta'_{ai} \\ &\approx a_i + \delta'_{bi} \\ &= a_i + b'_i - b_i, \end{aligned} \tag{13}$$

which yields

$$b_i = a_i - (\mathrm{HNet}(\mathrm{Concat}(a_i, \delta)) - b'_i), \tag{14}$$

where the rationale of approximation $\delta'_{ai} \approx \delta'_{bi}$ is based on $\mathrm{ENet}(\delta'_{\mathrm{ai}}) \approx \mathrm{ENet}(\delta'_{\mathrm{bi}})$ and is verified in quantitative experiment. The portion $(\mathrm{HNet}(\mathrm{Concat}(a_i, \delta)) - b'_i)$ is the to-be-processed component for GNet, e.g., the noise component in denoising, the bone component in deboning, and transformation residual in image generation. Note that in (14), both a_i and b'_i are available to the attacker, while $\mathrm{HNet}(\mathrm{Concat}(a_i, \delta))$ is unknown since it is encapsulated in a black box. However, the attacker can curate special data

to query ONet, similar to the first attack, and the outputs of these queries can reveal the underlying functionality of HNet, enabling the creation of a surrogate hiding network that approximates forward HNet, as illustrated in the next subsection.

Attack Process We perform a query-based reverse engineering similar to the first attack by curating b_j to bypass GNet. With the curated b_j and b'_j pairs, instead of training an inverse of HNet, we train a surrogate model denoted by SNet that approximates HNet, establishing the guess mapping from processed but unmarked images to processed and watermarked images by minimizing the following simple loss function

$$\mathcal{L}_{\text{Surrogate}} = \sum_j \text{MSE}\left(\text{SNet}(b_j), b'_j\right). \tag{15}$$

It is important to note here that despite SNet approximates HNet, it does not concatenate the input with a mark but instead directly process b_j. The rationale lies in that b'_j inherently contains the mark δ embedded by HNet. Upon convergence, SNet grabs the functional essence of the black-box protected HNet. Therefore, we can replace the unknown component $\text{HNet}(\text{Concat}(a_i, \delta))$ in (14) by its approximate SNet, yielding

$$\hat{b}_i = a_i - (\text{SNet}(a_i) - b'_i), \tag{16}$$

obtaining the estimation of the processed but unmarked (equivalently, watermark-removed) image b_i, with all components known to the attacker. Note that (16) is our proposed realization of the generic removal attack in (5).

Summary and Further Discussion

To summarize, both attacks operate in two steps: attack preparation and attack execution, as illustrated in Figure 3. We note that these attacks essentially rely on the requirement for identity transformation to bypass GNet. As a countermeasure, we propose implementing API detection to evaluate the similarity between the input a_i and the output b_i of GNet. A practical implementation involves computing the Euclidean distance of a_i and b_i. If the distance falls below a predefined threshold, the system directly returns a_i with warning.

Experiments

We demonstrate the effectiveness of our proposed approaches by attacking two state-of-the-art box-free model watermarking methods (Wu et al. 2020; Zhang et al. 2024) for the tasks of image deraining and image generation, respectively, with the latter focusing on the text-image-to-image-based image editing using Stable Diffusion (Rombach et al. 2022). For the ease of notation, (Wu et al. 2020) is referred to as V_{Wu} and (Zhang et al. 2024) as V_{Zhang}. Notably, V_{Zhang} is the promoted version of (Zhang et al. 2021), which alleviates the vulnerability against normal image augmentation attacks. However, since our attack method does not involve any image augmentation operations, and the watermark embedding and extraction processes in (Zhang et al. 2021) and "V_{Zhang}" are identical, we treat both as the same victim and avoid redundant discussion.

Settings

Dataset Following victims, The PASCAL VOC dataset (Everingham et al. 2010) is used for image deraining task. It is composed of $12,000$ images from Domain A with raindrop noise, which is generated by algorithm (Zhang and Patel 2018), and $12,000$ derained images from Domain B. We equally divide the dataset into two parts, each containing $6,000$ noised and $6,000$ denoised images, to serve as training data for the victim models and both attacks, respectively. For image generation, we randomly generate $12,000$ images (served as a_i) by Stable Diffusion (Rombach et al. 2022) and also split them evenly for training the victim model and attacks. In addtion, all the images in both datasets are 256×256 RGB images.

Metric We evaluate the quality of watermark-removed images by two commonly used metrics, i.e., peak signal-to-noise ratio (PSNR) and multi-scale structural similarity index (MS-SSIM) (Wang, Simoncelli, and Bovik 2003), respectively. The watermark removal success rate of our proposed attack is defined as

$$\text{SR}_{\text{Remove}} = 1 - \text{SR}_{\text{Extract}}, \tag{17}$$

where $\text{SR}_{\text{Extract}}$ is the rate of successful watermark extractions, and a single extraction is successful if the normalized correlation coefficient between the ENet output and the ground-truth watermark δ is greater than 0.96.

Qualitative Results

Watermark Removal The qualitative results of our proposed attacks against V_{Wu} (Wu et al. 2020) and V_{Zhang} (Zhang et al. 2024) are presented in Figure 4. In each subfigure, the attack against deraining task is shown in the first row, while attack against image generation task is shown in the second row. Columns from left to right represent to-be processed image ($a_i \in A$), processed unmarked image ($b_i \in B$), processed marked image ($b'_i \in B'$), watermark removed image $\hat{b}_i \in \hat{B}$, embedded watermark (δ), ENet extracted watermark from $b'_i \in B'$, and ENet extracted output from watermark removed $\hat{b}_i$. It can be seen in the figures that both proposed attacks can successfully remove the watermark, although dispersed noise dots remain when attacking V_{Wu}.

Noise Estimation With a slight abuse of terms, we call the to-be-processed image component for GNet, i.e., $a_i - b_i$, collectively as "noise" (it is actually the rain component in deraining and transformation residual in image generation). Recall (14), we note that the key factor enabling the forward HNet attack is the tractable estimation of the noise component $(\text{SNet}(a_i) - b'_i)$. Figure 5 shows the qualitative experimental results of the proposed forward HNet attack for noise estimation in both deraining and image generation tasks against the two victim models. The columns from left to right represent the to-be-processed image ($a_i \in A$), processed unmarked image ($b_i \in B$), the ground-truth noise, and the estimated noise, respectively. Both noise images are amplified $10\times$ for better visibility. It can be seen that the estimated noise patterns show significantly high similarity

Figure 4: Qualitative demonstration of (a) HNet^{-1} attack on the deraining task and (b) image generation task, with the first row attacking V_{Wu} (Wu et al. 2020) and the second row attacking $\mathrm{V}_{\mathrm{Zhang}}$ (Zhang et al. 2024). Similarly, (c) Forward HNet attack on the deraining task and (d) image generation task, with the first row attacking V_{Wu} (Wu et al. 2020) and the second row attacking $\mathrm{V}_{\mathrm{Zhang}}$ (Zhang et al. 2024).

to their respective ground-truth patterns, and the normalized correlation coefficients between the estimate and ground-truth noise patterns are 0.883, 0.981, 0.892, and 0.979, respectively, from top to bottom.

Quantitative Results

The verification results for $\delta'_{ai} \approx \delta'_{bi}$ are shown in Table 2, while the quantitative experimental results applying our proposed attacks against the two victim models for deraining and image generation are summarized in Table 3, where the Correlation column refers to the average normalized correlation coefficients between the estimated and ground-truth noises for the proposed forward HNet attack.

Verification for $\delta'_{ai} \approx \delta'_{bi}$ Table 2 presents the average PSNR, MS-SSIM, and normalized correlation coefficient between δ'_{ai} and δ'_{bi}. The consistently high values of these metrics strongly support the hypothesis $\delta'_{ai} \approx \delta'_{bi}$ in the derivation of forward HNet attack.

Fidelity The fidelity of the proposed attacks is evaluated using PSNR and MS-SSIM metrics, comparing $\hat{b}_i$ with the unknown ground truth b_i. As shown in Table 3, for the HNet^{-1} attack, the average PSNR values exceed 33.38 dB for deraining task and 31.41 dB for image generation task, with all MS-SSIM values surpassing 0.986. For the forward

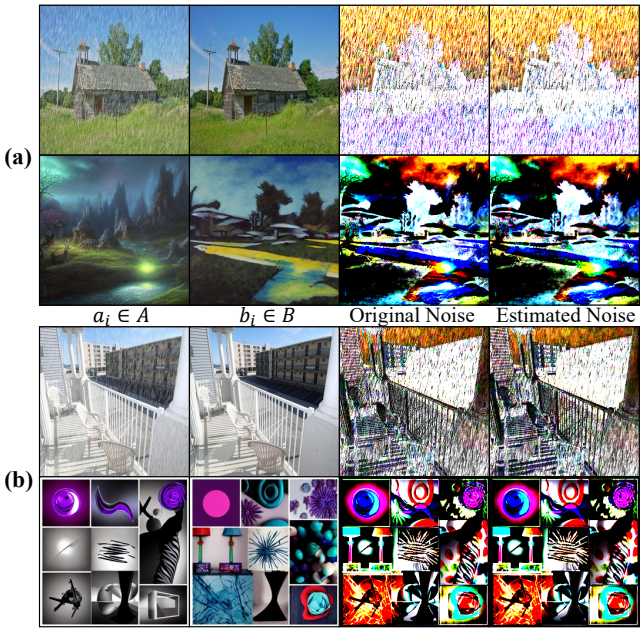

Figure 5: Qualitative demonstration of the proposed forward HNet attack on noise estimation performance for image deraining and generation tasks, where the noise is amplified by $10\times$ for better visibility. (a) Against V_{Wu} (Wu et al. 2020) (b) Against $\mathrm{V}_{\mathrm{Zhang}}$ (Zhang et al. 2024).

HNet attack, the average PSNR values are greater than 33.75 dB for deraining and 32.05 dB for image generation, while all MS-SSIM values are greater than 0.988, for both victims models, demonstrating very high-fidelity performance. The fidelity comparison of the two attacks demonstrate the superior image quality performance achieved by the forward HNet attack. In addition, we notice that the image quality of the generated image data is inferior to that of the deraining data. As illustrated in Figure 4, generated images are significantly more complex than those in the PASCAL VOC dataset (Everingham et al. 2010), resulting in the task-wise image quality differences.

Removal Success Rate All eight sets of experimental results shown in Table 3 have achieved 100% watermark removal ($\mathrm{SR}_{\mathrm{Remove}} = 1.000$). This represents a significant advancement over prior watermark removal methods not specifically designed for box-free model watermarking, as our attacks maintain the perfect removal rate while simultaneously preserving superior image quality.

Ablation Study

In (7), curated images that satisfy approximate identity transformation are used as input to bypass GNet encapsulated in ONet, which are subsequently combined with the corresponding watermarked images to conduct attacks. We demonstrate the necessity of bypassing GNet in the ablation study where image generation task is considered as an example. New HNet^{-1} and SNet are trained using pairs of images in A and their corresponding watermarked images in

Tasks	V_{Wu} (Wu et al. 2020)			V_{Zhang} (Zhang et al. 2024)		
	PSNR (dB)↑	MS-SSIM↑	Correlation↑	PSNR (dB)↑	MS-SSIM↑	Correlation↑
Deraining	34.19	0.948	0.800	34.41	0.9513	0.635
Image Generation	37.48	0.987	0.923	35.84	0.985	0.894

Table 2: Verification for approximation $\delta'_{ai} \approx \delta'_{bi}$ in forward HNet attack.

Victim Model	Removal Attack	Deraining				Image Generation			
		Correlation↑	PSNR (dB)↑	MS-SSIM↑	SR_{Remove} ↑	Correlation↑	PSNR (dB)↑	MS-SSIM↑	SR_{Remove} ↑
V_{Wu} (Wu et al. 2020)	JPEG-20%	–	26.53	0.940	1.000	–	24.71	0.936	1.000
	JPEG-50%		28.13	0.960	1.000		26.60	0.957	1.000
	AWGN-20dB		24.34	0.906	1.000		24.41	0.931	0.570
	AWGN-30dB		28.37	0.958	0.280		28.52	0.967	0.000
	Lattice-Interval2		13.82	0.681	1.000		13.73	0.734	1.000
	Lattice-Interval8		24.13	0.915	0.057		24.07	0.935	0.000
	WEvade-B-Q		35.98	0.856	0.380		43.11	0.954	0.140
	Regeneration-VAE		32.89	0.980	1.000		31.18	0.980	0.970
	Regeneration-Diff		22.74	0.845	1.000		21.38	0.841	1.000
	Inverse HNet **(Ours)**		33.38	0.987	1.000		31.75	0.986	1.000
	Forward HNet **(Ours)**	0.883	33.75	0.990	1.000	0.981	32.05	0.988	1.000
V_{Zhang} (Zhang et al. 2024)	JPEG-20%	–	27.57	0.955	1.000	–	24.93	0.943	1.000
	JPEG-50%		29.86	0.977	1.000		26.78	0.965	1.000
	AWGN-20dB		25.26	0.925	0.998		24.67	0.943	0.995
	AWGN-30dB		30.89	0.980	0.384		28.94	0.980	0.010
	Lattice-Interval2		13.89	0.695	1.000		13.75	0.745	1.000
	Lattice-Interval8		24.94	0.956	0.384		24.30	0.947	0.950
	WEvade-B-Q		30.95	0.796	0.530		35.00	0.884	0.450
	Regeneration-VAE		32.67	0.978	1.000		32.54	0.986	0.980
	Regeneration-Diff		22.70	0.858	1.000		21.61	0.853	1.000
	Inverse HNet **(Ours)**		33.98	0.988	1.000		31.41	0.991	1.000
	Forward HNet **(Ours)**	0.892	34.69	0.992	1.000	0.979	32.60	0.990	1.000

Table 3: Quantitative evaluations and comparisons of proposed attacks against existing methods, including JPEG, AWGN, Lattice (Liu et al. 2023a), WEvade-B-Q (Jiang, Zhang, and Gong 2023), and Regeneration (Zhao et al. 2024), where PSNR is in dB, and $0 \leq$ Correlation, MS-SSIM, $SR_{Remove} \leq 1$.

B'. The attack results are illustrated in Figure 6. We notice that for the two victims and the two attacks, $HNet^{-1}$ (the first row of each subfigure) removes the embedded watermark (the last column) but at the cost of significant image quality degradation (the fourth column), thereby failing to meet the fidelity requirement of our attack goal. In addition, forward HNet (the second row of each subfigure) fails to remove the embedded watermark (the last column), while incorporating the original components from a_i into $\hat{b}_i$. These results collectively demonstrate the need for circumventing GNet for conducting successful attacks.

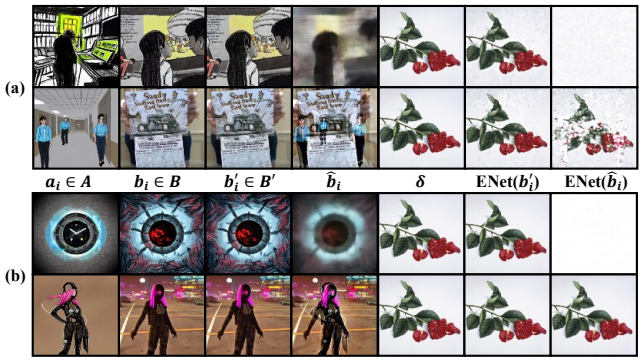

Figure 6: Ablation study on the need for bypassing GNet in the proposed attack for image generation task, against (a) V_{Wu} (Wu et al. 2020) and (b) V_{Zhang} (Zhang et al. 2024). In each subplot, the first row is $HNet^{-1}$ attack result, while the second is forward HNet attack result.

Conclusion

We have proposed two black-box watermark removal attacks driven by query-based reverse engineering against existing box-free model watermarking under the real-world black-box setting, specifically image-to-image models. We begin by showing the simple and efficient attack which removes watermarks by training an inversion model of HNet but with inferior output image quality. To fill the gap, we demonstrate the equivalent additive form of box-free model watermarking originally performed by the nonlinear HNet and exploit a simple surrogate training under the practical threat model. Then, a forward HNet attack is developed, which, based on the surrogate model of HNet, the query input, and the ONet output, can effectively estimate the unknown processed and unmarked image, thus achieving watermark removal with better image quality preservation. Finally, we point out the necessity of API detection in box-free model watermarking system against these attacks. Extensive experiments on deraining and image generation tasks demonstrated our attacks remove embedded watermarks at perfect success rates. Overall, our proposed removal attack reveals the vulnerabilities of box-free model watermarking in real-world scenarios, highlighting the need for more effective countermeasures.

Acknowledgments

The research work described in this paper was partially conducted in the JC STEM Lab of Smart City funded by The Hong Kong Jockey Club Charities Trust under Contract 2023-0108. The work was also supported in part by the Hong Kong SAR Government under the Global STEM Professorship and Research Talent Hub and Singapore Ministry of Education (MOE) under the Academic Research Fund (AcRF) Tier 1 Grant R-MA123-R205-0008. We sincerely appreciate the support provided.

References

An, H.; Hua, G.; Fang, Z.; Xu, G.; Rahardja, S.; and Fang, Y. 2025. Decoder Gradient Shield: Provable and High-Fidelity Prevention of Gradient-Based Box-Free Watermark Removal. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13424–13433.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Huang, Z.; Li, B.; Cai, Y.; Wang, R.; Guo, S.; Fang, L.; Chen, J.; and Wang, L. 2023. What can discriminator do? towards box-free ownership verification of generative adversarial networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5009–5019.

Jiang, Z.; Zhang, J.; and Gong, N. Z. 2023. Evading watermark based detection of AI-generated content. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 1168–1181.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 694–711. Springer.

Liu, H.; Xiang, T.; Guo, S.; Li, H.; Zhang, T.; and Liao, X. 2023a. Erase and Repair: An Efficient Box-Free Removal Attack on High-Capacity Deep Hiding. *IEEE Transactions on Information Forensics and Security*, 18: 5229–5242.

Liu, Y.; Han, T.; Ma, S.; and et al. 2023b. Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2): 100017.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Wang, Z.; Simoncelli, E.; and Bovik, A. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, volume 2, 1398–1402.

Wu, H.; Liu, G.; Yao, Y.; and Zhang, X. 2020. Watermarking neural networks with watermarked images. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(7): 2591–2601.

Zhang, H.; and Patel, V. M. 2018. Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 695–704.

Zhang, J.; Chen, D.; Liao, J.; Fang, H.; Zhang, W.; Zhou, W.; Cui, H.; and Yu, N. 2020. Model watermarking for image processing networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12805–12812.

Zhang, J.; Chen, D.; Liao, J.; Ma, Z.; Fang, H.; Zhang, W.; Feng, H.; Hua, G.; and Yu, N. 2024. Robust Model Watermarking for Image Processing Networks via Structure Consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–8.

Zhang, J.; Chen, D.; Liao, J.; Zhang, W.; Feng, H.; Hua, G.; and Yu, N. 2021. Deep model intellectual property protection via deep watermarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4005–4020.

Zhao, X.; Zhang, K.; Su, Z.; Vasan, S.; Grishchenko, I.; Kruegel, C.; Vigna, G.; Wang, Y.-X.; and Li, L. 2024. Invisible image watermarks are provably removable using generative ai. *Advances in Neural Information Processing Systems*, 37: 8643–8672.