Query-efficient Attack for Black-box Image Inpainting Forensics via Reinforcement Learning

Xianbo Mo[1], Shunquan Tan[1*], Bin Li[2], Jiwu Huang[1]

[1]Faculty of Engineering, Shenzhen MSU-BIT University, China
[2]Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University, China
6420240106@smbu.edu.cn,tanshunquan@gmail.com,libin@szu.edu.cn,jwhuang@smbu.edu.cn

Abstract

Recently, image inpainting has become a common tool for manipulating nature images in a malicious manner, which has led to the rapid advancement of inpainting forensics. Although current forensics methods have shown precise location of inpainting regions and reliable robustness against image post-processing operations, it remains unclear whether they can effectively resist the possible attacks in real-world scenarios. To identify potential flaws, we propose a novel black-box anti-forensics framework to attack inpainting forensics methods, which employs reinforcement learning to generate a query-efficient countermeasure, named RLGC. To this end, we define reinforcement learning paradigm to model the Markov Decision Process of query-based black-box anti-forensics scenario. Specifically, pixel-wise agents are used to modulate anti-forensics images based on action selection and query forensics methods to obtain corresponding outputs. Later, reward function evaluates attack effect and image distortion with these outputs. To maximize the cumulative reward, policy and value networks are integrated and trained by Asynchronous Advantage Actor-Critic algorithm. Experimental results demonstrate that, without visually detectable distortion on anti-forensics images, RLGC achieves remarkable attack effects in a highly query-effcient way against various black-box inpainting forensics methods, even outperforming the most representative white-box attack method.

Introduction

The rapid development of image forgery technologies such as image inpainting has seriously challenged the authenticity and integrity of digital images (Verdoliva 2020). As digital images are frequently employed as evidence and reliable records in contexts such as criminal investigations, journalistic reporting, and intellectual property protection, the detection and localization of inpainted regions within images have emerged as significant research challenges. Numerous forensic methods(Mayer and Stamm 2018; Li and Huang 2019; Wu and Zhou 2022; Yang, Cai, and Kot 2022; Zhang et al. 2023) based on deep learning have been proposed to address these issues in the past decade.

Although deep learning technology has significantly advanced inpainting forensics, current research (Akhtar and Mian 2018) has shown that it is vulnerable to adversarial attacks. These attacks introduce perturbations to manipulate input data, thereby misleading the target network into producing incorrect predictions. They are typically classified as either white-box or black-box attacks, depending on the level of prior knowledge about the target network. White-box attacks have full access to the details of the target network, while black-box attacks have no access to these internal details. In inpainting forensics, adversarial attacks pose a severe challenge. The reason is that inpainting forensics methods remain the only reliable way to identify inpainting regions, as human eyes struggle to detect manipulated areas within meticulously crafted inpainting images. Consequently, the incorrect forensics results caused by adversarial attacks are of utmost concern. Moreover, since anti-forensics attacks expose vulnerabilities in existing forensics methods, research on these attacks can further promote the design of more robust forensics methods. To address these challenges, many anti-forensics methods have been proposed in recent years (Barni et al. 2019; Carlini and Farid 2020; Xie, Ni, and Shi 2021; Ding et al. 2022; Fan, Hu, and Ding 2024).

However, existing attack methods are inappropriate for image inpainting forensics. First, current anti-forensics methods have focused on attacking image-wise classifiers, whereas pixel-wise segmentation networks are used in inpainting forensics. Secondly, in real-world scenarios, inpainting forensics operate as black-box systems, with only query results available to attackers. This makes query-based attacks the most suitable approach for inpainting forensics. Unfortunately, existing anti-forensics methods are predominantly designed for white-box attacks, with several relying on transfer-based black-box attacks. Besides, for query-based black-box adversarial examples methods(Li and Chen 2021; Andriushchenko et al. 2020; Maho, Furon, and Le Merrer 2021) designed for computer vision tasks, current research(Li et al. 2022) has shown that they can be easily detected based on the similarity of successive queries. Nevertheless, this detection method suffers from extremely high false alarm rates when the number of queries is less than 10. Therefore, there is an urgent need for query-efficient anti-forensics methods tailored for black-box image inpainting forensics, capable of constructing successful attacks with fewer than 10 queries.

To address this, we propose RLGC(**R**einforcement

*Corresponding author.

Learning to **G**enerate **C**ountermeasure) to conduct highly query-efficient attack for black-box image inpainting forensics methods. Our goal is to utilize query results to build adversarial attacks and achieve minimum visual distortion on the original inpainting images. Specifically, we limit the query times of RLGC to less than 10 times to evade the detection of query-based attack defense method(Li et al. 2022). To this end, we first model attack scenario based on RL(Reinforcement Learning) paradigm. Given the original inpainting images, which correspond to the initial state, the agent selects actions based on its policy and conducts state transitions to modulate adversarial perturbations for anti-forensics images. These images are then used to query target forensics networks, obtaining corresponding outputs. The reward function evaluates the attack effect and visual distortion according to these outputs. To maximize the cumulative reward, our policy and value network are integrated by Asynchronous Advantage Actor-Critic (A3C) framework, where the advantage-based loss functions optimize network parameters. Through iterative interactions between agents and image inpainting forensics methods, RLGC's attack efficiency can be constantly optimized until it achieves the given goal. Our contributions are as follows: **(1)** We propose the first query-based anti-forensics framework targeting black-box inpainting forensic methods, thereby eliminating the dependence on the transferability of white-box attacks. **(2)** We first apply a reinforcement learning (RL) paradigm within the anti-forensics framework, enabling pixel-wise agents to learn highly query-efficient policies based on inpainting forensics results. **(3)** We propose a novel method for generating perturbations that incrementally introduce small magnitudes(+1/-1/0) of noise, thus mitigating the risk of generating excessively strong noise that could leave conspicuous attack traces.

Background and Preliminaries

Image Inpainting Forensics

Given a mask $\{\mathbf{M} = (m_{i,j})^{(w\times h)}, m_{i,j} \in \{0,1\}\}$, the damaged image can be calculated as: $\mathbf{D} = (d_{i,j,k})^{(w\times h\times c)} = (x_{i,j,k} * m_{i,j})^{(w\times h\times c)}$, image inpainting is to obtain an inpainting image $\mathbf{Y}$ that satisfies:

$$\min_{Y\in\mathcal{I}} \|\mathbf{X}-\mathbf{Y}\|,\ \mathbf{Y} = (y_{i,j,k})^{(w\times h\times c)} = \theta_i((d_{i,j,k})^{(w\times h\times c)}) \tag{1}$$

where $\|\cdot\|$ is L2 norm, and θ_i is the inpainting algorithm.

Over the years, advancements in inpainting frameworks have resulted in synthetic images that are increasingly difficult to distinguish from authentic ones. Consequently, various forensics methods have been proposed for detecting image inpainting (Mayer and Stamm 2018; Li and Huang 2019; Wu and Zhou 2022; Yang, Cai, and Kot 2022; Zhang et al. 2023). These methods not only determine the authenticity of an image but also locate its synthetic regions. Given a ground truth mask $\boldsymbol{M}$, the objective of inpainting forensics methods θ_f is to predict the mask $\boldsymbol{M^p}$ while adhering to the following constraints:

$$\min \|\boldsymbol{M}-\boldsymbol{M^p}\|,\ \boldsymbol{M^p} = (m^p_{i,j})^{(w\times h)} = \theta_f((y_{i,j,k})^{(w\times h\times c)}) \tag{2}$$

Black-box Adversarial Examples

Black-box attacks can be categorized into two types: **(1)** Transfer-based attacks: A local substitute model is first trained. Adversarial examples are then generated using white-box attacks such as FGSM(Goodfellow, Shlens, and Szegedy 2015) and i-FGSM(Kurakin, Goodfellow, and Bengio 2017) on the substitute model, which are later directly used to attack target models. The success of these attacks depends on the transferability of white-box adversarial examples, which is influenced by the discrepancy between target and substitute models. **(2)** Query-based attacks: They can be classified into two categories: score-based attacks(Li and Chen 2021; Andriushchenko et al. 2020) and decision-based attacks(Maho, Furon, and Le Merrer 2021). Score-based attacks involve adding slight perturbations to the input and observing the response of target models. In contrast, decision-based attacks start from a point already in the adversarial region and use binary search to find a point on the decision boundary between the starting point and the clean example.

Reinforcement Learning

RL is one of the three fundamental machine learning paradigms, which focuses on creating intelligent agents that can take actions to maximize cumulative reward. To model a real-world scenario, a tuple $(\mathcal{S}, \mathcal{A}, \pi, r, \delta)$ is defined based on the existing background knowledge, where $\mathcal{S}$ denotes the set of states or the environment, $\mathcal{A}$ represents the action set, π reflects the probability of state transition, r signifies the reward generated by the state transition, and δ is the discount factor for rewards. Typically, practical RL algorithms are based on infinite-horizon MDP (Markov Decision Process) with successive state transitions. In a single state transition, the process starts from the current state $s_c \in \mathcal{S}$. Then, an action $a_c \in \mathcal{A}$ is selected according to $\pi(a_c|s_c)$. As a result, s_c transitions to next state $s_n \in \mathcal{S}$, which leads to the reward $r(s_c, s_n)$. In recent years, with the rapid development of deep learning technology, RL has been integrated with deep learning to form the deep RL paradigm.

A3C(Mnih et al. 2016) is a deep RL-based algorithm. The foundation of A3C is an actor-critic framework, where the actor selects its actions for the current state s_c based on $\pi(a_c|s_c)$, while the critic evaluates the value of the next state s_n. Typically, deep learning-based policy and value networks are used as the actor and critic in A3C. To train these networks, A3C leverages the advantage of the actor over the critic, which is the difference between the expected reward and value. We denote the policy network and value network as P and V respectively, and represent their parameters as θ_p and θ_v. At time step t, the expected reward of N following states $\{s_{(t+i)}|i = 0, 1, ..., N-1\}$ is calculated as:

$$\overline{R}^N_{(t)} = \sum_{i=0}^{N-1} \lambda^i r_{(t+i)} + \lambda^N V(s_{(t+N)}), \tag{3}$$

where r_t is the reward of state s_t, $V(s_{(t+N)})$ is the value of state $s_{(t+N)}$, and λ is discount factor. Then, advantage function of actor over critic can be represented as:

$$A(a_t, s_t) = \overline{R}^N_{(t)} - V(s_t). \tag{4}$$

From the respect of critic, its target is to minimize $A(a_t, s_t)$ through the gradient descent algorithm as:

$$d_{\theta_v} = \nabla_{\theta_v}((A(a_t, s_t))^2), \tag{5}$$

where d_{θ_v} is the gradient of V. On the other hand, actor's target is to maximize $A(a(t), s(t))$, thus P's gradient d_{θ_p} is:

$$d_{\theta_p} = \nabla_{\theta_p}(-\log \pi(a_t|s_t)(A(a_t, s_t))), \tag{6}$$

where $\log \pi(a_t|s_t)$ is the probability map outputted by P.

In addition, A3C uses asynchronous gradient descent with multiple agents running independently on separate threads, sharing policy and value networks. They gather training data through state transitions, calculate gradients, and update networks asynchronously, allowing for more efficient training and improved policy learning.

Proposed Method

To propose a practical anti-forensics framework, two major challenges need to be addressed. The first one arises from the fact that most forensics methods are black-box systems. As copyright protection and security concerns prevent the disclosure of such methods' details, only the forensics results are available to users querying these black-box systems. Thus, a practical anti-forensics framework should be developed based on only query results. Building upon this assumption, the simplest query-based attack on black-box inpainting forensics systems can be conducted by adding random noise (perturbation) to inpainting images until the query results indicate that it successfully disrupts the outputs. Fig. 1 illustrates this procedure. However, this type of attack is likely to be query-intensive due to the lack of prior knowledge about the target forensics system and the inefficiency in utilizing query results. Additionally, it may result in visually detectable distortions in inpainting images due to the cumulative effect of excessive random noise.

Thus, the second challenge entails minimizing the total number of queries n required while retaining optimal attack performance, which corresponds to build a query-efficient anti-forensics framework. However, the smaller n usually means a compromised attack performance, thereby defeating the primary objective of the anti-forensics attack. Therefore, it is imperative to balance between visual quality and attack efficacy.

As depicted in Fig. 1, it is evident that at any given time step $(0 < t \leq n)$, the current inpainting image $\boldsymbol{X_t}$ can be simplified to depend solely on $\boldsymbol{x_{t-1}}$ and its corresponding perturbation $\boldsymbol{\xi_{t-1}}$. This dependency can be mathematically expressed based on transition probability π as follows:

$$\pi(\boldsymbol{X_t}|\boldsymbol{X_{0:t-1}},\ \boldsymbol{\xi_{0:t-1}}) = \pi(\boldsymbol{X_t}|\boldsymbol{X_{t-1}}, \boldsymbol{\xi_{t-1}}) \tag{7}$$

This equation confirms that the state transition from $\boldsymbol{x_{t-1}}$ to $\boldsymbol{x_t}$ satisfies the Markov property, where t ranges from 1 to n. Thus, we can model query procedure as a MDP guided by a given policy. And an effective policy is desired to make RLGC query-efficient. We propose a CNN-based policy to accomplish this. Additionally, we employ the A3C algorithm to better optimize our policy network. Prior to this, we need to define the fundamental elements of the RL paradigm used in RLGC.

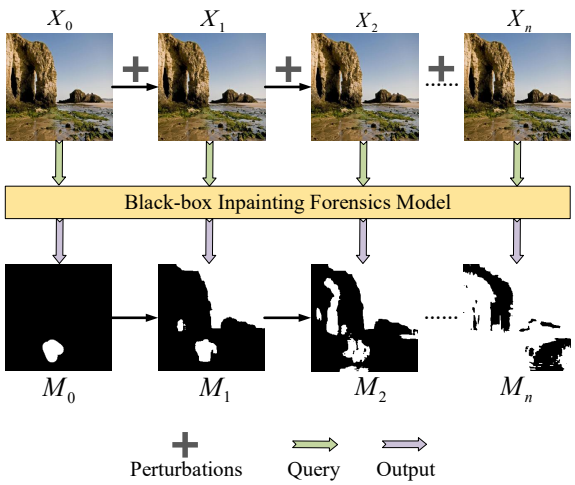

Figure 1: The illustration of querying black-box inpainting .

Elements Definition

Environmental model In RLGC, inpainting forensics methods serve as environmental model, with IID-Net(Wu and Zhou 2022) being utilized. IID-Net is selected due to its excellent detection performance and robustness against various image post-processing operations.

Agent Building upon the multi-threaded asynchronous parallel concept of the A3C framework, we assign an individual agent to each pixel. The objective is to empower each agent to adaptively determine its direction and magnitude of the perturbation by taking into consideration the distribution of neighboring pixels.

State Our state set $\mathcal{S}$ consists of images set $\mathcal{I}$, forming a high-dimensional space with the size of $256^{(w\times l\times c)}$. However, it is unnecessary to explore the entire state space as even small perturbations can lead to excellent attack performance. Specifically, given an original inpainting image $\boldsymbol{X_0} \in \mathcal{I}$, it serves as the initial state $\boldsymbol{S_0}$.

Action RLGC leverages actions as a mean to modulate perturbations for attacking forensics models. To help agents achieve more precise control, we set the magnitude of each perturbation to 1. For color images with three channels, the image-wise action map $\boldsymbol{A}$ can be denoted as $\boldsymbol{A} = \{(a^R_{i,j,k}, a^G_{i,j,k}, a^B_{i,j,k})^{(w*h*c)} | a^R_{i,j,k}, a^G_{i,j,k}, a^B_{i,j,k} \in \{0, -1, +1\}\}$, where $(a^R_{i,j,k}, a^G_{i,j,k}, a^B_{i,j,k})$ are corresponding to R, G, B channel of color images.

State transition The transition of $\boldsymbol{S_t}$ to $\boldsymbol{S_{t+1}}$, denoted as $T(\boldsymbol{S_{t+1}}|\boldsymbol{S_t}, \boldsymbol{A_t})$, can be formulated as:

$$\begin{aligned} T(\boldsymbol{S_{t+1}}|\boldsymbol{S_t}, \boldsymbol{A_t}) : \boldsymbol{X_{t+1}} &= \boldsymbol{X_t} + \boldsymbol{A_t} \\ (x^{t+1}_{i,j,k})^{w*h*c} &= (x^t_{i,j,k} + a^t_{i,j,k})^{w\times h\times c} \end{aligned} \tag{8}$$

where $\boldsymbol{X}_t$ and $\boldsymbol{X_{t+1}}$ are corresponding to $\boldsymbol{S_t}$ and $\boldsymbol{S_{t+1}}$. $\boldsymbol{A_t}$ is the action map that agents take at $\boldsymbol{S_t}$.

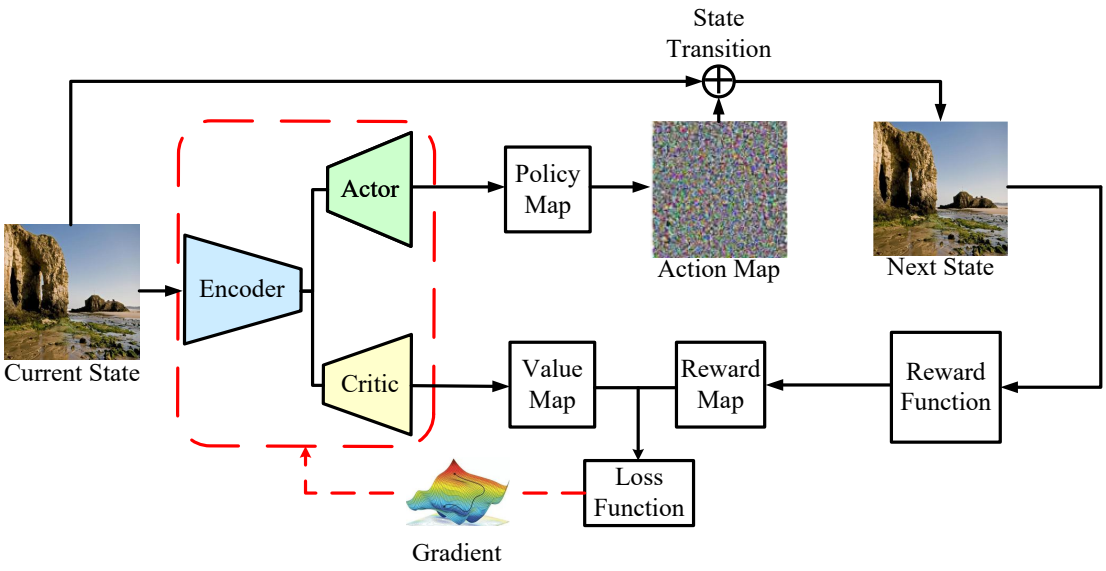

Figure 2: The illustration of a state transition of our proposed anti-forensics framework.

Reward function RLGC considers attack effect and visual distortion in reward function. Given an arbitrary state transition $T(\boldsymbol{S_{t+1}}|\boldsymbol{S_t}, \boldsymbol{A_t})$, its reward map $\boldsymbol{R} = (r_{i,j,k})^{w\times h\times c}$ can be calculated as:

$$\boldsymbol{R}(T(\boldsymbol{S_{t+1}}|\boldsymbol{S_t}, \boldsymbol{A_t})) = \omega_d \times \boldsymbol{R_D}(T(\boldsymbol{S_{t+1}}|\boldsymbol{S_t}, \boldsymbol{A_t}), \boldsymbol{S_0}) + \omega_a \times \boldsymbol{R_A}(T(\boldsymbol{S_{t+1}}|\boldsymbol{S_t}, \boldsymbol{A_t}), \boldsymbol{M}) \tag{9}$$

where $\boldsymbol{R_D}(T(\boldsymbol{S_{t+1}}|\boldsymbol{S_t}, \boldsymbol{A_t}), \boldsymbol{S_0})$ corresponds to visual distortion difference; $\boldsymbol{R_A}(T(\boldsymbol{S_{t+1}}|\boldsymbol{S_t}, \boldsymbol{A_t}), \boldsymbol{M})$ corresponds to attack performance difference; $\boldsymbol{M}$ is ground truth mask. Specifically $\boldsymbol{R_D}$ and $\boldsymbol{R_A}$ are calculated as follows:

$$\begin{aligned}&\boldsymbol{R_D}(T(\boldsymbol{S_{t+1}}|\boldsymbol{S_t}, \boldsymbol{A_t}), \boldsymbol{S_0}) = (r^d_{i,j,k})^{w\times h\times c}\\ &= [(\boldsymbol{X_t} - \boldsymbol{X_0}) * (\boldsymbol{X_t} - \boldsymbol{X_0})] -\\ &[(\boldsymbol{X_{t+1}} - \boldsymbol{X_0}) * (\boldsymbol{X_{t+1}} - \boldsymbol{X_0})]\\ &= \{(x^t_{i,j,k} - x^0_{i,j,k})^2 - (x^{t+1}_{i,j,k} - x^0_{i,j,k})^2|\\ &i \in \{1, 2, ..., w\}, j \in \{1, 2, ..., h\}, k \in \{1, 2, .., c\}\}\end{aligned} \tag{10}$$

$$\begin{aligned}&\boldsymbol{R_A}(T(\boldsymbol{S_{t+1}}|\boldsymbol{S_t}, \boldsymbol{A_t}), \boldsymbol{M}) = (r^a_{i,j,k})^{w\times h\times c}\\ &= [(\boldsymbol{M_t} - \boldsymbol{M}) * (\boldsymbol{M_t} - \boldsymbol{M})] -\\ &[(\boldsymbol{M_{t+1}} - \boldsymbol{M}) * (\boldsymbol{M_{t+1}} - \boldsymbol{M})]\\ &= \{(m^t_{i,j} - m_{i,j})^2 - (m^{t+1}_{i,j} - m_{i,j})^2|\\ &i \in \{1, 2, ..., w\}, j \in \{1, 2, ..., h\}, k \in \{1, 2, .., c\}\}\end{aligned} \tag{11}$$

where $*$ is Hadamard Product symbol, $\boldsymbol{X_{t+1}}$, $\boldsymbol{X_t}$ are the inpainting images corresponding to $\boldsymbol{S_{t+1}}$, $\boldsymbol{S_t}$, respectively. $\boldsymbol{M_{t+1}}$ and $\boldsymbol{M_t}$ are the predicted masks outputted by the target forensics model.

Policy Optimization Network

The A3C-based policy optimization network in RLGC can be divided into three different modules as follows:

Encoder The encoder module enables RLGC to process high-dimensional states in an efficient manner. By compressing these states into lower-dimensional representations, redundant information can be removed, facilitating the learning process of our agent with the most relevant data for the anti-forensics task. To this end, we have utilized the ImageNet(Deng et al. 2009) pre-trained EfficientNet(Tan and Le 2019) to initialize our encoder module. This pre-trained network provides a plethora of useful feature information derived from natural images, with its intermediate layer features utilized for the input and concatenated layers of the actor and critic. Hence, our encoder module employs the down-sampled blocks extracted from EfficientNet B1.

Actor It generates a policy that directs the attack based on the features derived from the encoder. To achieve this, the actor module provides probability distributions for the sampling process of the action set, which consists of 27 items. Therefore, the output of the actor module can be expressed as a probability distribution over the action set as $\boldsymbol{P} = \{(p_{i,j,k})^{(w\times h\times 27)}, |\sum_{k=1}^{27} p_{i,j,k} = 1, \forall(i,j) \in \{1, 2, ..., w\} \times \{1, 2, ..., h\}\}$.

Critic It is used for value function approximation. The goal of the critic module is to estimate the value function of the current attacked image based on the features provided by encoder, which is defined as the expected sum of future rewards that an agent can receive from the current attacked image. Thus, we denote the value map as $\boldsymbol{V} = (v_{i,j,k})^{(w\times h\times c)}$, whose size is the same as reward map.

The actor and critic are both responsible for processing the encoder's features to accomplish their specific tasks. Consequently, we have designed a similar network structure for both components. The middle layers of actor and critics are same as the upsampling module of UNet(Ronneberger,

Fischer, and Brox 2015). For the activation function of output layer, it is Softmax for actor, while it is Tanh for critic.

Training and Testing Procedures

In this section, we describe how RL elements collaborate with policy optimization algorithm to construct query-based attack via state transitions, as depicted in Fig. 2. Started from arbitrary current state s_t, encoder model takes it as input and outputs the corresponding features for actor and critic. Later, the action map $\boldsymbol{A_t}$ is sampled from policy map $\boldsymbol{P_t}$ outputted by actor, while the value map $\boldsymbol{V_t}$ is directly generated by critic. After conducting state transition $T(s_{t+1}|s_t, \boldsymbol{A_t})$, the reward map $\boldsymbol{R_t}$ is obtained through calculating reward function (Equation (9)). In the training procedure, $\boldsymbol{V_t}$ and $\boldsymbol{R_t}$ generated by 6 state successive state transitions are used to calculate A3C's loss function(Equation (3) and (4)).

In the testing procedure, the number of state transitions may vary. To ensure that RLGC maximizes the distance between the predicted mask and ground truth, we terminate state transition procedure when the attack performance of next state is worse than that of current state. We utilize the F1-score as the evaluation metric for our attack performance. On the other hand, considering that longer state transition procedures tend to result in more severe image distortions, we have set a threshold for the decline in F1-score between current and next states, which is fixed at 0.02. In other words, if the difference in F1-score between the current and next state exceeds 0.02, state transition procedure will continue. Otherwise, state transition procedure will be terminated, and we will consider current state as the terminal state. The corresponding inpainting image generated at the terminal state will be deemed as the optimal attacked sample for target forensics methods. Furthermore, it is important to note that, during the first two state transitions, termination will not occur, as small perturbations during this early stage may not result in stable attack performance.

Experiments

Experimental Setup

Dataset In black-box attack, the dataset settings are typically undisclosed to attackers. To showcase the effectiveness of RLGC in this scenario, we utilized two distinct datasets, which were introduced by IID-Net (Wu and Zhou 2022), for inpainting forensics methods and RLGC as follows:

Dataset for Inpainting Forensics Methods: It is denoted as $\mathcal{D}_F$, which contains 48,000 pairs of inpainting images and ground-truth masks. The original images come from the *Places* dataset (JPEG lossy compression)(Zhou et al. 2017) or the Dresden dataset (NEF lossless compression)(Gloe and Böhme 2010) with a proportion of 1:1. The masks are randomly sampled from (Liu et al. 2018), and all the inpainting images are generated by CA(Yu et al. 2018).

Dataset for RLGC: It is denoted as $\mathcal{D}_A$, which contains 11,000 pairs of inpainting images and ground-truth masks. Its original images come from two additional datasets, CelabA(Karras et al. 2018) and ImageNet(Deng et al. 2009). And eleven different representative inpainting methods are used to generate inpainting images, including seven deep learning-based ones proposed in recent years (CA(Yu et al. 2018), GC(Yu et al. 2019), SH(Yan et al. 2018), EC(Nazeri et al. 2019), LB(Wu, Zhou, and Li 2021), RN(Yu et al. 2020), and LR(Guo et al. 2017)), and four traditional ones (TE(Telea 2004), NS(Bertalmio, Bertozzi, and Sapiro 2001), PM(Herling and Broll 2014), and SG(Huang et al. 2014)). Note that TE and NS were published before 2005, but they have been included in the OpenCV extension package as the built-in default inpainting methods, indicating their wide usage and the meaningfulness of the results based on them. The proportion of training, validating and testing is 4:1:5.

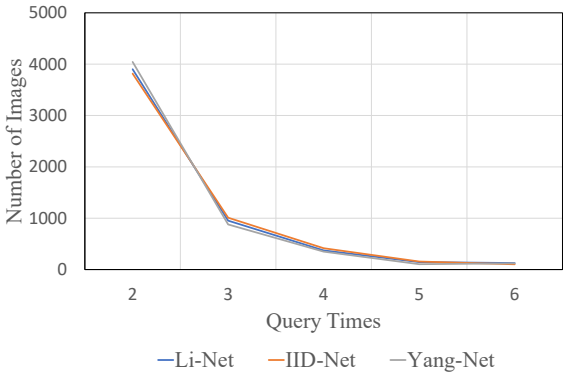

Figure 3: The query times required for RLGC.

Attack Comparison

To conduct a comprehensive comparison, we evaluated the attack performance of RLGC(query-based black-box attack method with up to 6 query times) against two other attack methods: FGSM(Goodfellow, Shlens, and Szegedy 2015)(gradient-based white-box attack method) and Square Attack(Andriushchenko et al. 2020)(query-based black-box attack method with up to 250 query times). And the maximal magnitude of attack perturbation was set to 4 for FGSM and RLGC, while it was 50 for Square Attack. From Table 1, it is evident that both RLGC and FGSM outperform Square Attack by significant margin. When comparing FGSM and RLGC, despite RLGC operating in a black-box manner and FGSM in a white-box manner, RLGC consistently achieves better attack performance in almost all scenarios. These results demonstrate that RLGC not only provides a robust attack mechanism but also does so with fewer queries, making it an efficient and effective method for compromising image inpainting forensics methods and query-based attack defense method(Li et al. 2022).

Evaluation of Query Efficiency

To assess the query efficiency, we recorded the number of query times required for RLGC to generate the final attacked inpainting images against different forensics methods. The results are presented in Fig. 3. In Fig. 3, the x-axis represents query times to attack, while the y-axis represents the number of images in testing set of $\mathcal{D}_A$. It can be observed

Forensics Methods	Metrics	Original Results	Anti-Forensics Methods	Attacked Results
Li-Net	**F1-score**	71.71%	FGSM	25.92% (↓ 45.79%)
			Square Attack	31.34%(↓ 40.37%)
			RLGC	**24.45%(↓ 47.26%)**
	IOU	62.13%	FGSM	16.76%(↓ 45.37%)
			Square Attack	24.22%(↓ 37.91%)
			RLGC	**16.15%(↓ 45.98%)**
	AUC	83.13%	FGSM	62.08%(↓ 21.05%)
			Square Attack	70.02%(↓ 13.11%)
			RLGC	**61.52%(↓ 21.61%)**
IID-Net	**F1-score**	84.85%	FGSM	22.40%(↓ 62.45%)
			Square Attack	52.82%(↓ 32.03%)
			RLGC	**21.33%(↓ 63.52%)**
	IOU	78.32%	FGSM	15.78%(↓ 63.54%)
			Square Attack	40.00%(↓ 38.32%)
			RLGC	**14.93%(↓ 63.39%)**
	AUC	98.21%	FGSM	**59.93%(↓ 38.28%)**
			Square Attack	92.65%(↓ 5.56%)
			RLGC	61.44%(↓ 36.77%)
Yang-Net	**F1-score**	85.07%	FGSM	29.89%(↓ 55.18%)
			Square Attack	43.78%(↓ 41.29%)
			RLGC	**21.79%(↓ 63.28%)**
	IOU	79.56%	FGSM	20.01%(↓ 59.55%)
			Square Attack	35.64%(↓ 43.92%)
			RLGC	**14.42%(↓ 65.14%)**
	AUC	95.58%	FGSM	74.08%(↓ 21.50%)
			Square Attack	82.91%(↓ 12.67%)
			RLGC	**59.68%(↓ 35.90%)**

Table 1: Attack performance of different anti-forensics methods against state-of-the-art forensics models.

that most of the images only require 2 query times to generate their final attack inpainting images, whose percentage is 3,953/5,500 = 71.87%. Furthermore, we have also calculated the average number of query times, which is 2.44. This indicates that RLGC exhibits a extremely low query cost while achieving excellent attack performance. Based on these results, we can conclude that RLGC is a highly query-efficient anti-forensics framework for attack different forensics methods in black-box scenario.

Visual Analysis

To better compare the attack performance of RLGC with other comparative attack methods, we conducted a visual analysis on the predicted mask outputted by IID-Net. The results are presented in Fig. 4. Both FGSM and RLGC significantly disturbed the predicted masks compared to corresponding ground truth masks. However, Square Attack's attack performance was found to be unstable. For RLGC and FGSM, although the evaluation of attack performance in Table 1 with metrics of F1-score, IOU, and AUC suggests that RLGC's superiority over FGSM may not be substantial, the visual analysis reveals significant differences between them. For instance, as depicted in Fig. 4, the attack effect caused by FGSM classifies most pixels as inpainting pixels, resulting in obviously larger inpainting regions in the corresponding predicted masks than the original regions. This indicates that FGSM attack forensics methods by causing the higher false alarm rate. Conversely, RLGC prefers to classify most pixels as original pixels, resulting in larger original regions in the predicted masks. Importantly, in the context of anti-forensics attack, the primary goal is to conceal inpainting regions while preserving the original regions in predicted masks. Therefore, we argue that RLGC aligns more closely with actual anti-forensics goals compared to FGSM.

Forensics Methods	PSNR	SSIM
Li-Net	42.69	0.9798
IID-Net	42.58	0.9789
Yang-Net	42.50	0.9756

Table 2: Image quality of attack images generated by RLGC.

Moreover, image distortion caused by the attack of RLGC is not visually noticeable in Fig. 4. It achieves by the fact that image quality is directly associated with query efficiency in RLGC, as each query time introduces modifications of $\{-1, +1, 0\}$ to the attack images while its average number of query times is 2.44. To further validate the quality of RLGC's attack images, we conducted image quality assessment with metrics of Peak Signal to Noise Ratio (PSNR) and Structure Similarity Index Measure (SSIM). The average PSNR and SSIM scores between attack images generated by RLGC and original images are shown in Table 2. The results demonstrate RLGC achieves excellent image quality after assigning perturbations on attack images. Based on

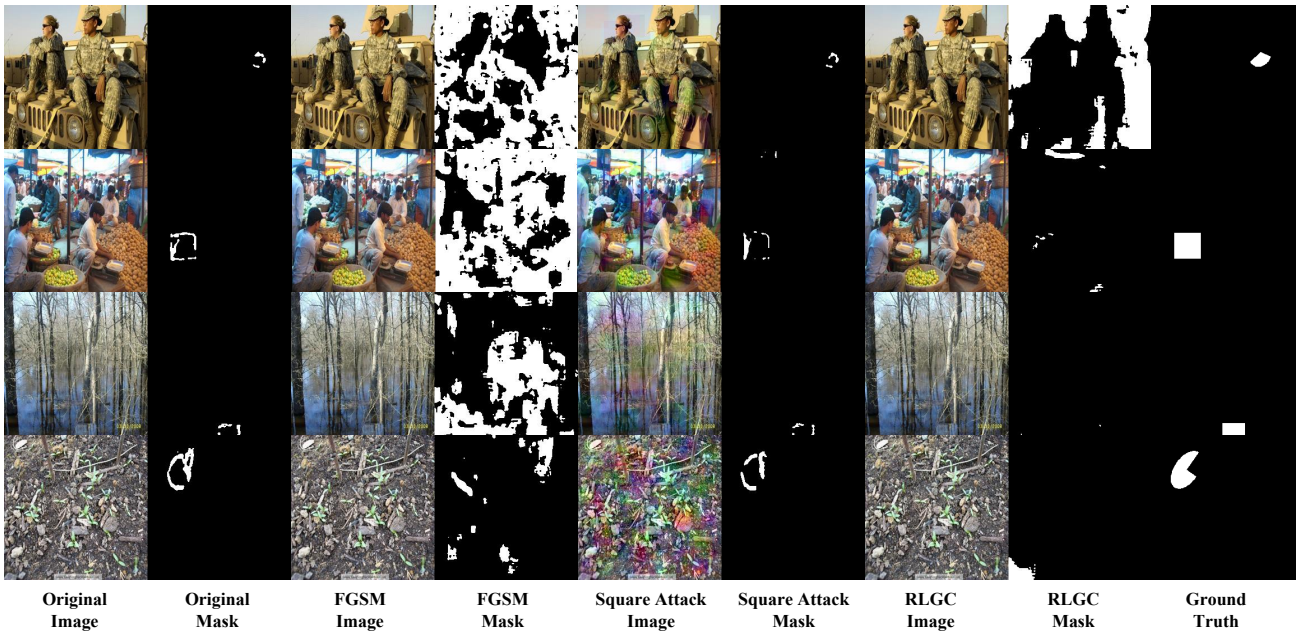

Figure 4: The visualization of inpainting images and its corresponding masks outputted by IID-Net.

Training On	F1-score	IOU	AUC
$\mathcal{D}_{NS}$	22.13%	15.63%	64.56%
$\mathcal{D}_{TE}$	24.22%	17.31%	71.47%
$\mathcal{D}_{PM}$	24.36%	19.13%	72.93%
$\mathcal{D}_{SG}$	24.19%	17.33%	69.33%
$\mathcal{D}_{LR}$	22.42%	15.93%	62.40%
$\mathcal{D}_{CA}$	23.42%	16.67%	68.78%
$\mathcal{D}_{SH}$	24.06%	17.40%	71.84%
$\mathcal{D}_{EC}$	23.97%	17.30%	71.70%
$\mathcal{D}_{GC}$	24.69%	17.78%	69.87%
$\mathcal{D}_{RN}$	23.96%	17.33%	72.07%
$\mathcal{D}_{LB}$	25.52%	18.68%	72.59%
$\mathcal{D}_{A}$	21.33%	14.93%	61.44%

Table 3: The location performance of RLGC with single training inpainting method against IID-Net.

these observations, we can conclude that the generated attack images of RLGC are not visually distinguishable compared with original images by human eyes, making RLGC's attacks more covert.

Limited Training Scenario

In this section, the effectiveness of RLGC is evaluated with limitation of inpainting method in training dataset. For example, we selected the images generated by inpainting method of GC, denoted as $\mathcal{D}_{GC}$, which contains a total of 1,000 inpainting images. We use only 400 images for training, 100 images for validation, and the remaining 500 images and the other 10,000 inpainting images not generated by GC, for testing. Additionally, we reduce the number of training iterations, which is only 300 overall and one well-trained model is save for each 50 iterations. In this context, RLGC's attack results are shown in Table 3.

From Table 3, we find that RLGC still achieves remarkable attack performance against IID-Net. For example, the IOU scores of all training subsets are distributed in the interval of 15% to 20%. These results highlight the reliable generalization capability of RLGC for mismatched inpainting methods between training and testing datasets. It is crucial for real-world applications since there are always other inpainting methods that are not included in training datasets.

Time Cost

RLGC's training takes around 25 hours when conducted on a single NVIDIA L40 GPU. And the average time to attack one image is 3.37 seconds. These results demonstrate RLGC's efficient and practical characteristic.

Conclusions and Future Work

In this paper, we present a query-based anti-forensics framework for attacking black-box inpainting forensics methods, using RL-based techniques. It achieves both high attack performance and negligible image distortion based on query-efficient attack. Experiments demonstrate that RLGC is effective in transferring across different inpainting methods and detectors, even when experimental settings for training and testing differ. In the future, we aim to address several issues in the future. First, we will expand our anti-forensics scenarios to include more image forgery operations. Second, we aim to leverage the power of RL to automatically generate forgery images, thus addressing the pressing need for more well-crafted forgery datasets.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant 62272314, U23B2022) and Guangdong Provincial Key Laboratory (Grant 2023B1212060076).

References

Akhtar, N.; and Mian, A. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6: 14410–14430.

Andriushchenko, M.; Croce, F.; Flammarion, N.; and Hein, M. 2020. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, 484–501. Springer.

Barni, M.; Kallas, K.; Nowroozi, E.; and Tondi, B. 2019. On the transferability of adversarial examples against CNN-based image forensics. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8286–8290.

Bertalmio, M.; Bertozzi, A. L.; and Sapiro, G. 2001. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 355–362.

Carlini, N.; and Farid, H. 2020. Evading deepfake-image detectors with white-and black-box attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops(CVPRW)*, 658–659.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 248–255.

Ding, F.; Shen, Z.; Zhu, G.; Kwong, S.; Zhou, Y.; and Lyu, S. 2022. ExS-GAN: Synthesizing anti-forensics images via extra supervised GAN. *IEEE Transactions on Cybernetics*, 53(11): 7162–7173.

Fan, B.; Hu, S.; and Ding, F. 2024. Synthesizing black-box anti-forensics deepfakes with high visual quality. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4545–4549. IEEE.

Gloe, T.; and Böhme, R. 2010. The'Dresden Image Database'for benchmarking digital image forensics. In *Proceedings of the ACM Symposium on Applied Computing*, 1584–1590.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *3rd International Conference on Learning Representations(ICLR)*.

Guo, Q.; Gao, S.; Zhang, X.; Yin, Y.; and Zhang, C. 2017. Patch-based image inpainting via two-stage low rank approximation. *IEEE transactions on Visualization and Computer Graphics*, 24(6): 2023–2036.

Herling, J.; and Broll, W. 2014. High-quality real-time video inpainting with PixMix. *IEEE Transactions on Visualization and Computer Graphics*, 20(6): 866–879.

Huang, J.-B.; Kang, S. B.; Ahuja, N.; and Kopf, J. 2014. Image completion using planar structure guidance. *ACM Transactions on Graphics (TOG)*, 33(4): 1–10.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2017. Adversarial examples in the physical world. In *Proceedings of the International Conference on Learning Representations(ICLR)*.

Li, H.; and Huang, J. 2019. Localization of Deep Inpainting Using High-pass Fully Convolutional Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8301–8310.

Li, H.; Shan, S.; Wenger, E.; Zhang, J.; Zheng, H.; and Zhao, B. Y. 2022. Blacklight: Scalable defense for neural networks against {Query-Based}{Black-Box} attacks. In *31st USENIX Security Symposium (USENIX Security 22)*, 2117–2134.

Li, N.; and Chen, Z. 2021. Toward Visual Distortion in Black-Box Attacks. *IEEE Transactions on Image Processing*, 30: 6156–6167.

Liu, G.; Reda, F. A.; Shih, K. J.; Wang, T.-C.; Tao, A.; and Catanzaro, B. 2018. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 85–100.

Maho, T.; Furon, T.; and Le Merrer, E. 2021. Surfree: a fast surrogate-free black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10430–10439.

Mayer, O.; and Stamm, M. C. 2018. Accurate and Efficient Image Forgery Detection Using Lateral Chromatic Aberration. *IEEE Transactions on Information Forensics and Security*, 13: 1762–1777.

Mnih, V.; Badia, A.; Adria, P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning(ICML)*, 1928–1937.

Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; and Ebrahimi, M. 2019. EdgeConnect: Structure Guided Image Inpainting using Edge Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Tan, M.; and Le, Q. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97, 6105–6114.

Telea, A. 2004. An Image Inpainting Technique Based on the Fast Marching Method. *Journal of Graphics Tools*, 9(1): 23–34.

Verdoliva, L. 2020. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5): 910–932.

Wu, H.; and Zhou, J. 2022. IID-Net: Image Inpainting Detection Network via Neural Architecture Search and Attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1172–1185.

Wu, H.; Zhou, J.; and Li, Y. 2021. Deep generative model for image inpainting with local binary pattern learning and spatial attention. *IEEE Transactions on Multimedia*, 24: 4016–4027.

Xie, H.; Ni, J.; and Shi, Y.-Q. 2021. Dual-domain generative adversarial network for digital image operation antiforensics. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1701–1706.

Yan, Z.; Li, X.; Li, M.; Zuo, W.; and Shan, S. 2018. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 1–17.

Yang, W.; Cai, R.; and Kot, A. 2022. Image Inpainting Detection via Enriched Attentive Pattern with Near Original Image Augmentation. In *Proceedings of the 30th ACM International Conference on Multimedia(ACM MM)*, 2816–2824.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2018. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5505–5514.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2019. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international Conference on Computer Vision (CVPR)*, 4471–4480.

Yu, T.; Guo, Z.; Jin, X.; Wu, S.; Chen, Z.; Li, W.; Zhang, Z.; and Liu, S. 2020. Region normalization for image inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, 12733–12740.

Zhang, Y.; Fu, Z.; Qi, S.; Xue, M.; Cao, X.; and Xiang, Y. 2023. PS-Net: A Learning Strategy for Accurately Exposing the Professional Photoshop Inpainting. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1452–1464.