

DP-TRAE: A Dual-Phase Merging Transferable Reversible Adversarial Example for Image Privacy Protection

Xia Du, Jiajie Zhu, Jizhe Zhou*, Chi-man Pun, *Senior Member, IEEE*, Zheng Lin, Cong Wu, *Member, IEEE*, Zhe Chen, *Member, IEEE*, and Jun Luo, *Fellow, IEEE*

Abstract—In the field of digital security, Reversible Adversarial Examples (RAE) combine adversarial attacks with reversible data hiding techniques to effectively protect sensitive data and prevent unauthorized analysis by malicious Deep Neural Networks (DNNs). However, existing RAE techniques primarily focus on white-box attacks, lacking a comprehensive evaluation of their effectiveness in black-box scenarios. This limitation impedes their broader deployment in complex, dynamic environments. Furthermore, traditional black-box attacks are often characterized by poor transferability and high query costs, significantly limiting their practical applicability. To address these challenges, we propose the Dual-Phase Merging Transferable Reversible Attack method, which generates highly transferable initial adversarial perturbations in a white-box model and employs a memory-augmented black-box strategy to effectively mislead target models. Experimental results demonstrate the superiority of our approach, achieving a 99.0% attack success rate and 100% recovery rate in black-box scenarios, highlighting its robustness in privacy protection. Moreover, we successfully implemented a black-box attack on a commercial model, further substantiating the potential of this approach for practical use.

Index Terms—Adversarial attack, privacy protection, black-box attack.

I. INTRODUCTION

DEEP Neural Networks (DNNs) have initiated a technological revolution in various fields [1]–[14], such as image recognition, natural language processing, and autonomous driving. Despite rapid progress in artificial intelligence across these domains, concerns about security and privacy have also increased [15]–[18]. Malicious actors often exploit unauthorized DNNs to analyze and steal users’ private data to their

The part of this work has been published in ACM MM 2024.

Xia Du and Jiajie Zhu are with the School of Computer and Information Engineering, Xiamen University of Technology, Xiamen, 361000, China (email: duxia@xmut.edu.cn; cosmos36@163.com).

Jizhe Zhou is with the School of Computer Science, Engineering Research Center of Machine Learning and Industry Intelligence, Sichuan University, Chengdu, China, 610020, China (email: yb87409@um.edu.mo).

Chi-man Pun is with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, 999078, China (email: cmpun@umac.mo).

Zheng Lin and Cong Wu are with the Department of Electrical and Electronic Engineering, University of Hong Kong, Pok Fu Lam, Hong Kong SAR, China (e-mail: linzheng@eee.hku.hk; congwu@hku.hk).

Zhe Chen is with the Institute of Space Internet, Fudan University, Shanghai, China, and the School of Computer Science, Fudan University, Shanghai, China (e-mail: zhechen@fudan.edu.cn).

Jun Luo is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: junluo@ntu.edu.sg).

* denotes Corresponding author.

Corresponding author: Jizhe Zhou (yb87409@um.edu.mo).

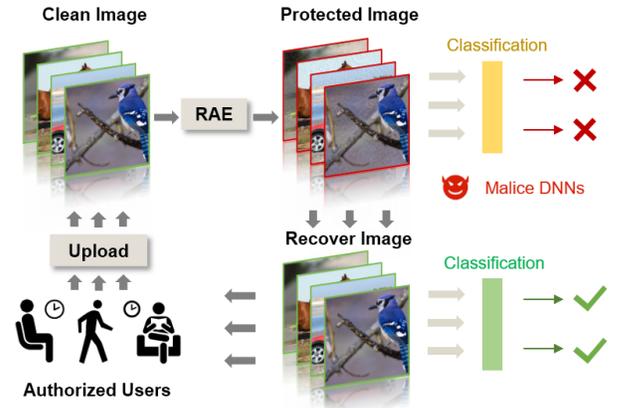


Fig. 1. RAEs prevent malicious DNNs from stealing privacy data and can recover the image quality when necessary.

advantage [19]–[23]. For example, Cambridge Analytica used unauthorized personal data from Facebook users for targeted political advertising [24].

Recent research has exposed critical vulnerabilities in DNNs [25]–[27], particularly their susceptibility to variations in input data quality and distribution [28]. In image recognition tasks, even minor pixel modifications can significantly mislead classification results [29]–[32]. Adversarial attacks exploit these weaknesses by introducing subtle perturbations that mislead DNNs incorrect predictions [33]. Recent studies have proposed using adversarial attacks to protect image privacy from malicious DNN analysis [34]–[36]. However, the adversarial noise introduced directly is often irreversible, leading to degraded image quality and reduced data usability, especially for digital images. Therefore, there is an urgent need for a technique that can protect image privacy while preserving visual quality. Our research indicates that Reversible Adversarial Examples (RAEs) [37] offer significant potential for data protection: this approach ensures data security while enabling the adversarial perturbations to be reversed, restoring the original image. As illustrated in Figure 1, RAEs can generate controlled and reversible adversarial perturbations to effectively mislead unauthorized DNNs, thus safeguarding user privacy without compromising data usability.

The field of RAEs is still in its early stages of development. Liu *et al.* first introduced the concept of reversible adversarial attacks by integrating reversible data hiding techniques with adversarial examples [37]. Xiong *et al.* further expanded re-

versible adversarial attacks to black-box scenarios, employing ensemble model techniques to demonstrate the potential of RAEs across multiple models [38]. Although their approach exhibited strong transferability and misleading capability, it faced limitations in effectively attacking previously unexploited models. Meanwhile, Zhang *et al.* [39] proposed a method utilizing RGAN technology to replace reversible data hiding techniques, efficiently generating adversarial examples via an attacking encoder network and reversing them through a recovery decoder network. While this approach successfully restored the original images, it was less effective against unknown models.

Due to the specific nature of RAE, most current RAEs rely on the transferability of white-box attacks to achieve attacks. However, the effectiveness of this transferability is predominantly contingent upon the intensity of the applied perturbations. The RDH method imposes strict limitations on the magnitude of perturbations, leading to a design conflict that significantly reduces the success rate of attacks against unknown black-box models.

To overcome the potential challenges of existing RAE techniques, particularly the balance between implementing effective adversarial attacks and adhering to the strict perturbation limits of reversible data hiding [40], [41] techniques, we propose the Dual-Phase Merging Transferable Reversible Adversarial Example (DP-TRAE), a novel reversible adversarial attack on black-box models. The preliminary version [42] of this work was presented at ACM MM 2024. DP-TRAE divides the entire attack into two phases: the Stepwise Adaptive White-box Attack (SA-WA) and a Memory-Assisted Expansion Black-box Attack (MAE-BA). The motivation behind DP-TRAE is to combine the high transferability of white-box attacks with the targeted nature of black-box attacks, thus reducing the overall cost of the attack. Our intuition suggests that, compared to random perturbations, conducting a black-box attack on top of adversarial perturbations initialized by a white-box attack can mislead the target model more efficiently, similar to modifying an existing masterpiece with a clear direction in mind. Although white-box adversarial noise may not fully align with the gradient ascent direction of an unknown model, it provides a superior initial direction for the attack. Among them, The SA-WA method introduces additional perturbation guidance for gradient-sensitive regions, thereby enhancing the efficiency of misdirection. It adaptively adjusts the magnitude of perturbations to mitigate overfitting, ensuring more robust attack performance. MAE-BA method records the impact of each perturbation on the results, accumulating and utilizing historical data to select the most promising update points and enhance the perturbation intensity in neighboring regions. Furthermore, to address the conflict between perturbation and RDH storage capacity [37], [38], [43], we regularize the perturbations and compress them using Huffman coding, effectively mitigating the trade-off between perturbation intensity and storage requirements. In particular, the main contributions of this work are as follows:

- We propose a novel Adaptive Transferable Reversible Adversarial Attack framework for black-box attacks, integrating multiple attack strategies to enhance both transferability

and efficiency effectively. To the best of our knowledge, our approach is the first successful application of RAE attacks on commercial black-box models.

- We propose the MAE-BA and SA-WA to accelerate the generation and compression of effective perturbations and employ Huffman coding to further compress the final perturbation information. These approaches tackle the key challenge of preserving the integrity of adversarial examples while guaranteeing their reversibility.

- Experimental results affirm the superiority of our attack framework, achieving a 99.0% Attack Success Rate (ASR) for reversible adversarial examples on specific models, with a 100% restoration rate for the recovered images. These outcomes validate the practical feasibility of our approach.

The rest of this paper is organized as follows. Section II describes the relevant background and technology. Section III provides the detailed design of the DP-TRAE. Section IV fully analyzes the experimental results and verifies the feasibility of the proposed DP-TRAE. Finally, the conclusions and future outlooks are presented in Section V.

II. RELATED WORK AND BACKGROUND

In this section, we review the existing work on adversarial attacks and the related techniques used in the proposed methods.

A. Adversarial Attack

Adversarial attacks generate perturbations to mislead model decisions. Goodfellow *et al.* [44] introduced the Fast Gradient Sign Method (FGSM), which exposed the vulnerability of deep learning models by adding small perturbations along the gradient direction. Despite FGSM's computational efficiency, its performance in complex scenarios is limited. To address these limitations, Kurakin *et al.* [45] proposed the iterative-FGSM (I-FGSM), enhancing attack effectiveness through multiple minor iterative updates. Dong *et al.* [46] extended I-FGSM by incorporating a momentum term, resulting in the Momentum Iterative Method (MI), which stabilizes the update direction and significantly improves the transferability of adversarial examples across different models.

Attacks based on a single input often exhibit poor transferability due to overfitting to a specific model. To alleviate this issue, data augmentation techniques have been integrated into adversarial attacks to increase input diversity. Xie *et al.* proposed the Diverse Input Method (DI) [47], which enhances adversarial attack effectiveness by applying random transformations (e.g., scaling and cropping). Dong *et al.* [48] further proposed the Translation-Invariant Method (TI), which uses convolution to smooth the gradient, expanding the perturbation's spatial extent and improving generalizability to unseen models.

In the black-box attack setting, Guo *et al.* developed the Simple Black-box Attack (SimBA) [49], which modifies individual input dimensions to evaluate their impact on model output, generating compelling adversarial examples with minimal queries. This straightforward approach demonstrates the efficiency of black-box attacks without relying on gradient

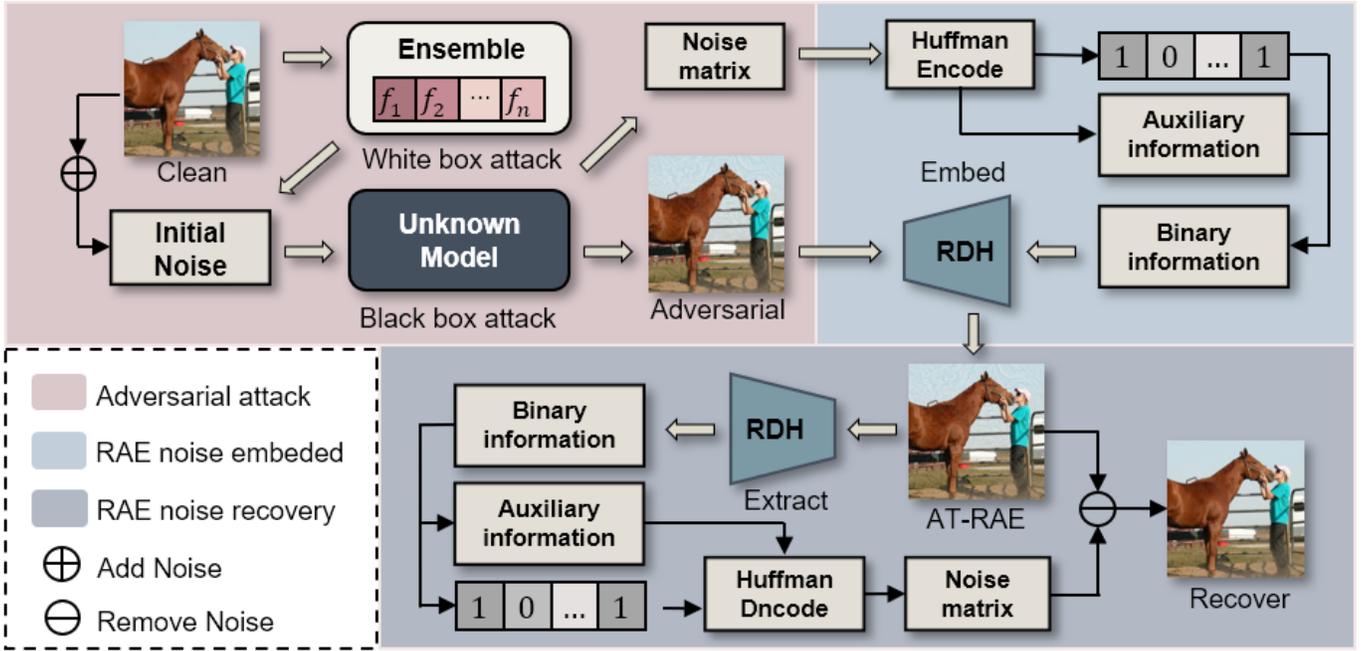


Fig. 2. An overview of the proposed DP-TRAE.

information. On the other hand, Maho *et al.* proposed the SurFree attack optimizes the query path using geometric considerations [50], eliminating the need for a substitute model. This approach substantially reduces query complexity while maintaining high attack success rates.

B. Reversible Data Hiding

Image steganography involves embedding secret information within images in an imperceptible way to the human eye. The Least Significant Bit (LSB) [51] substitution is a popular and highly effective technique known for its simplicity and ease of implementation while maintaining high imperceptibility. Discrete Cosine Transform embeds data in the frequency domain, providing robustness against compression. Discrete Wavelet Transform offers better localization in both spatial and frequency domains [52]. Grayscale invariant reversible steganography, which allows perfect image recovery after data extraction while maintaining robustness to grayscale variations, has also emerged as an important method [53]. Recently, deep learning-based approaches have been introduced to further enhance the security and capacity of steganographic systems [54], [55]. In a nutshell, it can be described as:

$$X_{RAE} = R(X, Mes), \quad (1)$$

where X is the carrier, and Mes is the embed message. In this paper, considering the balance between capacity and efficiency, we employ the LSB method for data embedding.

III. METHODOLOGY

A. Overview

This section elaborates on the DP-TRAE framework, which comprises three core modules illustrated in Figure 2: the

SA-WA module for white-box attacks, MAE-BA for black-box scenarios, and a reversible mechanism ensuring image preservation. DP-TRAE begins by performing rapid adversarial preprocessing on the input clean image to generate robust adversarial examples, thereby reducing the query overhead for the second-stage MAE-BA. MAE-BA estimates the gradient direction by querying superpixel blocks and utilizes historical query results to effectively improve attack efficiency. SA-WA and MAE-BA can be used independently to adapt to different attack scenarios. Finally, Huffman coding is applied to compress the information of the matrix, which is then embedded using RDH to generate DP-TRAE. During restoration, RDH extracts and reverses the perturbation matrix to losslessly recover the image.

B. Preliminary

Consider the white-box model f and the unknown target model b , where $x \in \mathcal{X}$ is a benign input with dimensions $H \times W \times C$ and the corresponding ground-truth label $y \in \mathcal{Y}$. Let $f(x)$ and $b(x)$ denote the prediction results of the white-box and black-box models, respectively. Consistent with prior work, we assume that the complete gradient information of f is available while b is entirely unknown, providing only the output labels and the associated probabilities.

For white-box attacks, given a benign input x and a loss function \mathcal{J} (e.g., the Cross-Entropy loss), the aim of the attack can be formulated as:

$$\delta = \arg \max \mathcal{J}(f(x + \delta), y), \quad (2)$$

$$f(x + \delta) \neq y, \quad s.t. \|\delta\|_{\infty} < \epsilon, \quad (3)$$

where δ represents the adversarial perturbation and ϵ is a constant that controls the norm constraint.

MI-FGSM is a classic attack algorithm that incorporates momentum to stabilize the optimization process to maximize Eq. 2 and improve attack success rates. In MI-FGSM, the adversarial perturbation is iteratively updated, and the accumulated gradient is used to determine the direction of the perturbation. Given a step size α , a number of iterations T , and a decay factor μ , the iterative update is defined as follows:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_{\delta} \mathcal{J}(f(x_t + \delta_t), y)}{\|\nabla_{\delta} \mathcal{J}(f(x_t + \delta_t), y)\|_1}, \quad (4)$$

$$\delta_{t+1} = \delta_t + \alpha \cdot \text{sign}(g_{t+1}), \quad (5)$$

where g_t is the accumulated gradient at iteration t , and δ_t represents the adversarial perturbation at iteration t . The use of the momentum term μ helps in accumulating the gradient information from previous steps, thereby generating more effective perturbations.

However, a single image input limits the robustness of the attack. Therefore, prior works have utilized a combination of DI and TI to increase input diversity, effectively preventing overfitting to a single model. And the iterative equation becomes:

$$g_{t+1} = \mu \cdot g_t + \frac{T \cdot \nabla_{\delta} \mathcal{J}(f(D(x + \delta_t), y))}{\|T \cdot \nabla_{\delta} \mathcal{J}(f(D(x + \delta_t), y))\|_1}, \quad (6)$$

where T is the convolution kernel in TI and D is the diverse input transformation in DI. We employ the combination of DI, TI, and MI as the basis of the attack to gain more transferability.

C. Stepwise Adaptive Attack

The SA-WA module aims to produce highly transferable adversarial perturbations. In white-box attacks, internal details of the target model are accessible, which allows for the effective use of gradient information to craft adversarial perturbations. During this attack process, we observed that the gradients of the model exhibit non-uniform variations across different input regions. The SA-WA leverages this by applying more substantial perturbations to gradient-sensitive areas, thereby prioritizing regions where the model's decision boundaries are most susceptible to deviation. This approach enhances the efficiency of generating adversarial perturbations, allowing significant attack effects to be observed in the initial iterations.

However, focusing solely on enhancing perturbations in gradient-sensitive regions may lead to overfitting to a single model, resulting in suboptimal performance and reduced attack transferability. Therefore, we introduced the adaptive strategy for scaling the perturbations in sensitive regions. Specifically, as the number of iterations increases, we progressively reduce the amplification applied to these regions. By doing so, we mitigate the risk of overfitting during later iterations, ensuring a balance between attack effectiveness and model transferability, and the SA-WA algorithm is presented in Algorithm 1.

In addition to focusing on attack performance, it is essential to consider the embedding and recovery of perturbations in

subsequent steps, as RDH technology imposes strict storage limitations, and the magnitude of adversarial perturbations at each position varies, which complicates direct encoding and storage. For instance, with a ϵ size of 8/255, storing the perturbation for a single pixel across three channels would require 48 bits, not even accounting for cases where the perturbation is zero. Therefore, it is crucial to compress the perturbation while preserving its effectiveness.

Algorithm 1 DI-TI-MI-SA

Require: Clean image x , model f , loss function \mathcal{J} , Iterations N , step size α , perturbation limit ϵ , original label y

Ensure: Adversarial perturbation δ

- 1: Initialize adversarial perturbation: $\delta = 0, X_{adv} = X$
 - 2: **for** $i \leftarrow 0$ to $N - 1$ **do**
 - 3: $X_{adv} = X + \delta_i$
 - 4: Apply DI: $X_{adv} = D(X_{adv})$
 - 5: Calculate gradients: $g_{i+1} = \nabla_{\delta} f(x_{adv}, y)$
 - 6: Apply TI, MI: utilize Eq. 6 and make $g_{temp} = g_{i+1}$
 - 7: Apply SA: The top $(N - i)/2N$ of g_{temp} magnitude changes are labeled as 2, while the rest are labeled as 1.
 - 8: Update perturbation: $\delta_{i+1} = \delta_i + \alpha \cdot \text{sign}(g_{t+1}) \cdot g_{temp}$
 - 9: Clip perturbation: $\delta_{i+1} = \text{clip}(\delta_{i+1}, -\epsilon, \epsilon)$
 - 10: **end for**
 - 11: **return** δ
-

To address this, SA-WA incorporates a compression mechanism directly into the iterative attack process. By embedding the compression step, SA-WA minimizes the impact of compression on attack performance, allowing for efficient storage without significantly compromising the attack's effectiveness. SA-WA first aggregates the gradients of channels, reducing the three-dimensional perturbation δ to one dimension:

$$g_t = \mathbb{E}_{c \in \{R, G, B\}} [g_{c_t}] = \frac{1}{3} \sum_{c \in \{R, G, B\}} g_{c_t}. \quad (7)$$

At this stage, the data storage volume is reduced to one-third of its original size. The perturbations are then compressed using a threshold-based approach, where precise values are replaced with uniform noise:

$$E(|\delta|) = \begin{cases} 2, & \frac{\epsilon}{2} \leq |\delta| \leq \epsilon \\ 1, & 0 < |\delta| \leq \frac{\epsilon}{2} \\ 0, & |\delta| = 0 \end{cases} \quad (8)$$

$$\delta_{t+1} = \text{sign}(\delta_{t+1}) \cdot E(|\delta_{t+1}|) \cdot \xi, \quad (9)$$

where ξ represents the stage threshold used to compress the perturbation values within a piecewise distribution. Compressed perturbation δ will be applied as the initial perturbation matrix for the MAE-BA method.

D. Memory-Assisted Expansion Attack

For black-box models, the SimBA method for generating adversarial perturbations is a simple yet effective approach. SimBA iteratively applies perturbations in randomly selected directions and observes the resulting change in the target class

probability. Perturbations that effectively reduce the target class probability are retained, enabling SimBA to reliably lower the model’s prediction confidence through repeated adjustments. This approach resembles finite difference methods for gradient estimation [56], but with a key distinction: SimBA’s stochastic coordinate perturbations do not directly estimate gradients, instead leveraging observed outcomes to generate effective perturbations. In this work, we adopt stochastic coordinate perturbations as the default strategy.

However, pixel-level stochastic coordinate perturbations, which involve perturbing individual pixels, result in a high number of queries, leading to increased costs, especially when attacks are repeated over multiple iterations. To address this, the MAE-BA method introduces the concept of superpixel blocks to reduce query complexity and associated costs. Superpixel blocks group neighboring pixels into cohesive regions, allowing for coordinated exploration of the input image via larger pixel groups rather than individual units. Operating on these aggregated blocks significantly reduces the required queries while maintaining perturbation effectiveness.

Algorithm 2 MAE-BA

Require: Clean image x , White adversarial noise δ , Black model b , Iterations N , stage threshold ξ , perturbation limit ϵ , original label y , enhance step size s , Expand size Ep

Ensure: Adversarial perturbation δ

- 1: Initialize Empty Memory list H
- 2: **for** $i \leftarrow 0$ to $N - 1$ **do**
- 3: **if** $b(x + \delta) \neq y$ **then**
- 4: Select random direction q
- 5: Select random superpixel blocks E_j
- 6: $P_{pre} = P_b(y|x + \delta)$
- 7: **if** $(i + 1)\%s = 0$ **then**
- 8: Replace index j with the largest value in H
- 9: Expand superpixel blocks E_j size with Ep
- 10: $H[j] = 0$
- 11: **end if**
- 12: $\delta = \delta + q \cdot \xi \cdot E_j$
- 13: $P_{next} = P_b(y|x + \delta)$
- 14: **if** $P_{pre} < P_{next}$ **then**
- 15: $\delta = \delta - q \cdot \xi \cdot E_j$
- 16: $P_{next} = P_b(y|x + \delta)$
- 17: **end if**
- 18: $H \leftarrow (P_{pre}/P_{next}, j)$
- 19: Clip perturbation: $\delta = \text{clip}(\delta, -\epsilon, \epsilon)$
- 20: **end if**
- 21: **end for**
- 22: **return** δ

In addition to employing superpixel blocks, we utilize information obtained from historical queries, a factor often neglected in typical black-box attack scenarios. Historical information provides insights into the sensitivity of specific image regions to perturbations, guiding the informed selection of subsequent query directions. Leveraging this historical knowledge enhances search efficiency by avoiding redundant queries and accelerating convergence toward effective adversarial perturbations. Specifically, we prioritize superpixel

blocks previously identified as promising based on historical query data, expanding the exploration scope around these regions. By focusing on the neighborhoods of promising superpixel blocks, we exploit the intrinsic local coherence of the gradient, where small perturbations in adjacent regions tend to yield similar effects on model output. This approach reduces redundant queries in less sensitive areas, concentrating instead on regions where minor adjustments are more likely to induce significant changes in model behavior. Furthermore, integrating historical query information as adaptive feedback guides the attack towards areas of maximum vulnerability, improving both the effectiveness and query efficiency of the adversarial perturbation process and the whole algorithm is presented in Algorithm 2.

E. Embed and Recover

After compressing the perturbations using Eq. 7 and Eq. 8, we observed a significant imbalance in the value distributions of different perturbations, as well as considerable variation among individual perturbations. Therefore, we decided to employ Huffman coding to effectively encode and store the perturbation information. By leveraging the advantages of Huffman coding, we can efficiently encode these perturbations with a higher compression ratio, thereby reducing storage requirements and improving processing efficiency. The detailed steps of the entire perturbation embedding algorithm are presented in Algorithm 3.

Algorithm 3 Encode and Embed

Require: Clean image x , Adversarial noise δ , stage threshold ξ , Flexible encryptor F , RDH technology R

Ensure: Reversible Adversarial Examples DP-TRAE

- 1: Initialize Huffman Tree T , Message matrix $Mse(H, W)$
- 2: **for** $h \leftarrow 0$ to $H - 1$ **do**
- 3: **for** $w \leftarrow 0$ to $W - 1$ **do**
- 4: Calculation *stage*: $stage = \delta_{hw}/\xi$
- 5: $Mse_{hw} = \xi$
- 6: Contract Huffman Tree: $T.append(stage)$
- 7: **end for**
- 8: **end for**
- 9: Compress 2D Mes into 1D
- 10: Encrypted message: $Mse = F(Mes + T)$
- 11: Generate DP-TRAE: DP-TRAE= $R(x + \delta, Mse)$
- 12: **return** DP-TRAE

When performing the recovery operation, we only need to reverse the embedding process to losslessly restore the perturbation information. Furthermore, during the encoding, compression, and storage processes, we can flexibly introduce specific encryption measures to ensure that the information remains secure during transmission and storage.

IV. EXPERIMENT

A. Experiment setup

Dataset and Environment. In this study, we employed the ILSVRC2012 dataset [57] to assess the effectiveness of various adversarial attack methods across different deep

TABLE I
ASR (%) ON SEVERAL MODELS UNDER ATTACK SCENARIOS USING RES-50, VGG-16, AND INC-v3 AS THE WHITE-BOX MODELS, RESPECTIVELY.

Source : RN-50	Target model										
	Attack	RN-34	RN-50	RN-152	DN-121	VGG-16	VGG-19	Inc-v3	Alexnet	Mob-v2	Mob-v3
Liu [37]	27.8	99.9	29.0	26.1	26.4	23.8	11.4	11.7	21.2	6.6	28.4
RIT [43]	34.7	99.8	31.3	31.8	26.8	27.0	15.1	19.1	24.0	9.4	31.9
DP-RAE [42]	47.1	100	43.9	44.7	45.2	46.6	24.1	26.7	43.5	16.5	43.8
DP-TRAE (Ours)	71.3	99.3	68.4	74.3	81.0	81.4	50.3	37.6	73.7	35.1	67.2
Source : Inc-v3	Target model										
	Attack	RN-34	RN-50	RN-152	DN-121	VGG-16	VGG-19	Inc-v3	Alexnet	Mob-v2	Mob-v3
Liu [37]	5.0	4.0	4.0	4.9	6.2	6.9	97.3	9.2	7.3	3.9	14.9
RIT [43]	9.7	8.4	6.8	9.6	9.8	10.7	97.6	15.2	12.3	8.1	18.8
DP-RAE [42]	20.1	18.5	14.1	18.2	19.2	19.2	98.1	25.4	25.3	15.4	27.4
DP-TRAE (Ours)	41.5	40.1	32.5	44.6	45.1	47.3	96.9	30.5	45.4	24.6	44.9
Source : VGG16	Target model										
	Attack	RN-34	RN-50	RN-152	DN-121	VGG-16	VGG-19	Inc-v3	Alexnet	Mob-v2	Mob-v3
Liu [37]	10.2	9.1	5.7	11.6	99.5	59.6	6.9	10.4	16.4	5.5	23.5
RIT [43]	11.8	11.1	5.6	14.5	99.6	53.5	9.5	15.2	16.2	7.5	24.5
DP-RAE [42]	16.2	17.2	10.9	17.6	99.7	79.4	12.3	21.9	26.2	12.5	31.4
DP-TRAE (Ours)	30.7	31.7	21.3	35.5	100	87.4	21.4	33.7	47.0	23.0	43.2

learning models. We randomly selected 1,000 images that the target models could classify accurately. All experiments were conducted on an NVIDIA A40 GPU, which ensured efficient processing capabilities for the extensive computations involved.

Regarding model selection, we focused on a range of models with diverse architectures and their corresponding sub-models, including ResNet34 (RN-34) [58], ResNet50 (RN-50) [58], ResNet152 (RN-152) [58], DenseNet121 (DN-121) [59], VGGNet16-BN (VGG16) [60], VGGNet19-BN (VGG19) [60], Inception-v3 (Inc-v3) [61], Alexnet [62], MobileNet-v2 (Mob-v2) [63], MobileNet-v3 (Mob-v3) [64]. This approach allowed us to evaluate the transferability of adversarial attacks across different model types, providing insights into their robustness and vulnerabilities.

Attack Setting. We set the maximum perturbation ϵ for the attack to $8/255$, the stage threshold was set to half of the δ , and the step size α equals the stage threshold. The number of iterations for the white-box attack was set to 10, while the maximum number of iterations for the black-box attack was set to 1000. For the black-box attack, the expand size Ep was set to 4, with an enhanced step size as $s = 5$.

Evaluation Metrics. Regarding the evaluation of attack performance, we employed Attack Success Rate (ASR) as the primary metric to assess the effectiveness of misleading different models. ASR is defined as the proportion of images that successfully deceive the target model out of the total number of input images. The higher ASR values reflect greater attack performance.

For recovery performance, we employed Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [65], which are widely used metrics for evaluating image quality. The higher PSNR value implies that the restored image

more closely resembles the original. On the other hand, SSIM assesses the similarity between two images and with values ranging from 0 to 1. A higher SSIM value indicates a more remarkable similarity to the original image regarding structural details. Together, these metrics provide a comprehensive assessment of the visual quality of the recovered images.

In addition, we evaluated the recognition accuracy of the recovered images, referred to as Success Rate (SR). SR measures the proportion of recovered images that the model correctly classifies. This metric provides insight into whether adversarial samples, after undergoing the recovery process, can successfully restore the classifier’s predictions to their original labels. A higher SR indicates that the recovery process effectively mitigates the impact of adversarial attacks, restoring the model’s ability to correctly classify the images while preserving high visual quality.

B. Attack performance

In this section, we evaluate the attack performance of DP-TRAE from two perspectives: the attack performance of DP-TRAE in white-box scenarios and the attack performance of DP-TRAE in black-box scenarios.

White-box scenarios. In this experimental setup, we compared DP-TRAE with several existing RAEs, including the RAE [37] proposed by Liu *et al.*, the RAE based on Reversible Image Transformation (RIT) [43] and our previous work DP-RAE [42]. We selected three structurally diverse models to conduct the attacks and used it to generate adversarial examples, which were then tested on the other models to assess single-model transferability. Table I shows the results. First, the proposed DP-TRAE consistently outperformed across most scenarios. This is because that the method proposed by Liu *et al.* fails to compress the perturbation magnitude, resulting in

TABLE II
THE ASR (%) ON SEVERAL MODELS UNDER ATTACK SCENARIOS USING RES-50, VGG-16, AND INC-V3 AS THE ENSEMBLE MODELS.

Attack	Test model							Ensemble model		
	RN-34	RN-152	DN-121	Mob-v2	Mob-v3	VGG-19	AlexNet	RN-50	Inc-v3	VGG-16
Liu [37]	29.5	28.8	34.8	31.0	6.5	59.5	9.5	99.1	99.0	99.4
RIT [43]	43.4	36.8	42.9	36.8	14.1	63.8	20.4	99.1	98.8	99.4
DP-RAE [42]	60.4	57.3	62.7	61.1	22.0	86.5	30.8	99.6	99.3	99.7
DP-TRAE (Ours)	83.2	81.4	87.6	85.6	44.4	96.2	42.9	99.2	99.0	99.9

TABLE III
THE ASR (%) ON DIFFERENT MODELS, WITH DN-121 AS THE BLACK-BOX MODEL, AND DP-TRAE UTILIZING THE ENSEMBLE PREPROCESSING OPERATION BASED ON SA-WA.

Attack	Target model										
	RN-34	RN-50	RN-152	DN-121	VGG-16	VGG-19	Inc-v3	Alexnet	Mob-v2	Mob-v3	Average
Simba [49]	1.0	0.4	0.2	94.6	1.2	1.0	0.9	1.7	0.3	0.2	10.2
Simba-DCT [49]	0.7	0.6	0.5	95.9	0.9	0.8	1.5	2.6	0.7	1.0	10.5
Surfree [50]	13.3	11.6	4.1	97.2	24.8	21.6	8.2	54	33.8	23.6	29.2
DP-RAE [42]	60.2	99.5	56.4	99.0	99.5	82	99.1	29.6	60.2	22.0	70.8
DP-TRAE (Ours)	80.9	98.0	79.7	99.0	99.6	92.2	97.1	41.9	83.8	43.2	81.5

over-detailed storage processes that significantly increase the steganographic storage overhead. This issue becomes particularly pronounced when the ϵ parameter is large, forcing Liu’s approach to sacrifice partial attack performance while meeting steganographic requirements. The RIT method achieves reversible adversarial attacks by directly disguising the original image as adversarial examples. However, its inability to losslessly preserve adversarial perturbation details during sample generation ultimately compromises the effectiveness of the attack. Our previous work DP-RAE employs grayscale-invariant steganography. While maintaining image grayscale properties, it significantly reduced storage efficiency and DP-RAE did not compress the perturbation information effectively. To preserve attack efficiency, the method relies on super-pixel blocks for perturbation compression, which inevitably degrades attack performance and leads to limited cross-model transferability. In contrast, DP-TRAE adopts a more concise and efficient RDH technique while leveraging Huffman coding to compress the perturbations. This approach eliminates the need for additional regional compression, effectively preserves attack performance, and demonstrates superior transferability compared to DP-RAE. Additionally, we observed that perturbations generated using the RN-50 exhibited better transferability across different models. It is attributed to the residual connections of RN-50, which are used to capture hierarchical and generalized feature representations. Residual connections help preserve crucial information across layers, resulting in perturbations that generalize better, making them more effective when used to deceive other models. Inc-V3 relies on convolutional blocks that focus on multi-scale feature extraction, which may result in more specialized perturbations to that specific architecture and be less effective on other models with different structures. Similarly, VGG16, with its simpler and more uniform convolutional stack, may lack the ability

to generate perturbations that capture complex, transferable features, leading to a reduced effectiveness when transferred to dissimilar models. Moreover, it can be observed that the generated perturbations demonstrate stronger transferability when applied to homologous models. Through the above analysis, it is suggested that both the architectural characteristics of the model used to generate adversarial examples and the nature of the learned feature representations play a crucial role in determining the transferability of adversarial attacks.

To further assess the transferability of adversarial perturbations, we employed an ensemble attack strategy, which enhances the cross-model transferability by optimizing perturbations across multiple models simultaneously. The perturbations were generated by integrating several white-box models, each assigned with simple weighting factors. As presented in Table II, the results demonstrate that the DP-TRAE consistently improves ASR in all cases. This improvement is attributed to introducing the stage threshold and Huffman coding compression mechanism, which effectively reduces the storage requirement for unit perturbations while ensuring the success rate and effectiveness of the attack, leading to a more refined application of perturbations. Additionally, the extra perturbation in gradient-sensitive regions accelerates the generation process.

Black-box attack scenario. Due to the current lack of research on black-box reversible attack methods, we compared DP-TRAE with several existing state-of-the-art black-box attack methods, including Simba [49], Simba-DCT [49], Surfree [50], and our previous work DP-RAE [42]. The experimental results, as shown in Table III, indicate that although these query-based black-box attacks exhibit strong performance when targeting specific models, they generally suffer from poor transferability. Specifically, these methods typically rely on optimizations tailored to specific models,

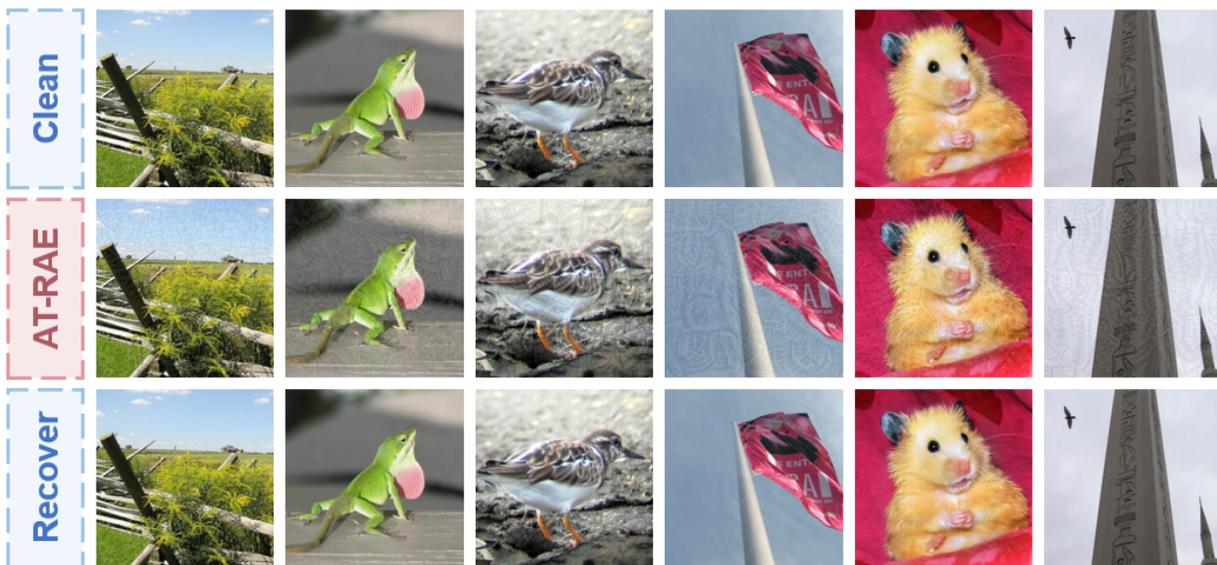


Fig. 3. Visual results of DP-TRAE, including clean images, attacked images, and the corresponding recovery images.

resulting in suboptimal performance when transferred to different target models. The main reason for the poor transferability is that these methods fail to effectively leverage the common features across different models, leading to a significant degradation in attack performance during cross-model transfer. In contrast, DP-TRAE enhances transferability by introducing preprocessing noise generated during the white-box phase. The introduction of this noise provides a more stable perturbation pattern for subsequent attacks, allowing DP-TRAE to achieve more consistent attack performance across different models. In addition, the MA-EA method further enhances the attack effect on the target model by optimizing the perturbation through the historical query results. Compared to the previous version DP-RAE, DP-TRAE produces black-box attack with better transferability due to the improved attack performance of the white-box attack phase.

C. Robustness evaluation

In practical applications, networks often preprocess input data and apply defense techniques to reduce the impact of adversarial attacks, improving overall performance. These defenses aim to reduce the effectiveness of adversarial examples, ensuring the model remains stable and accurate under malicious perturbations. Therefore, the robustness of adversarial examples against these defenses is crucial. In this study, we employ several preprocessing and defense methods, including Spatial Squeezing (Spatial) [66], Random Resizing and Padding (Random) [67], Gaussian Blurring (Gaussian) [68], JPEG Compression (JPEG) [66], and Super-resolution (Super) [69].

As shown in Table IV, adversarial perturbations from black-box attacks lose their effectiveness when subjected to these defensive techniques. This is because such preprocessing introduces uncertainty, weakening the impact of query-based attacks. In contrast, DP-TRAE utilized the white-box preprocessing approach to identify and exploit shared vulnerabilities

TABLE IV
THE ASR (%) OF ADVERSARIAL ATTACKS WHEN AGAINSTING DIFFERENT DEFENCE METHODS.

Attack method	Defense method				
	Spatial	Random	Gaussian	JPEG	Super
Simba [49]	4.5	16.8	20.3	20.3	2.7
Simba-DCT [49]	7.8	32.6	32.4	23.5	8.0
Surfree [50]	20.0	23.4	23.7	54.0	13.6
DP-RAE [42]	63.1	52.0	39.6	52.3	74.0
DP-TRAE (Ours)	89.9	54.5	50.7	82.3	88.3

across different models, thereby reducing the effectiveness of defense strategies. This approach enhances robustness by addressing vulnerabilities common to multiple model architectures.

D. Recover ability

RAEs are often evaluated based on their recovery performance. Notably, previous studies overlooked the assessment of this key metric. To demonstrate the recovery capability of DP-TRAE, we compared the images restored by DP-TRAE with the clean samples. As shown in Figure 3, the perturbations introduced by DP-TRAE cause only a slight degradation in image quality, which is imperceptible to human observers but can be devastating to adversarial models. The restored images effectively eliminate these perturbations without loss, making them indistinguishable from the original images.

Table V presents an evaluation of the restored images using several quality metrics. The PSNR of the restored images exceeds 45 dB, and the SSIM approaches 1 and the recovered images successfully recover the correct classifications in the model. In addition, we observed that the DP-TAE and the recoverable DP-TRAE are almost identical in all respects, this is because of the threshold-based perturbation compressing and Huffman coding compression, the RDH approach does

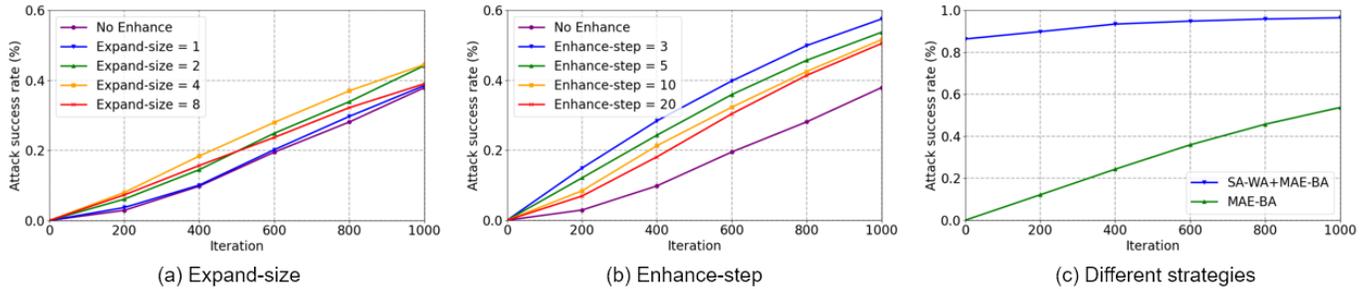


Fig. 4. (a) reported the impact of different expand sizes on attack performance; (b) reported the impact of different enhance step sizes on attack performance; (c) reported the impact of dual-phase strategie on attack performance.

TABLE V

WE REPORTED THE PSNR AND SSIM OF THE DP-TAE, DP-TRAE AND ITS RECOVER EXAMPLES, AND WE ALSO EVALUATED THEIR CLASSIFICATION SUCCESS RATE (SR) IN THE TARGET MODEL. "↑" MEANS THE BIGGER THE BETTER.

	PSNR (dB) ↑	SSIM ↑	SR (%) ↑
DP-TAE	31.14	0.8256	1.1
DP-TRAE	31.10	0.8242	1.1
Recover (DP-TRAE)	48.94	0.9913	100

not have a large impact on the adversarial examples. These results demonstrate the effectiveness of DP-TRAE is not only maintaining the adversarial performance but also ensuring that the restored images retain both high visual fidelity and model accuracy.

E. Ablation study

This section presents an ablation study on DP-TRAE to evaluate the impact of different parameters and strategies on attack performance.

SA-WA: First, we assessed the impact of the SA-WA on attack performance. To this end, we employed various gradient calculation strategies, including the IFGSM, MI-FGSM, DI-MI-FGSM (DI-MI), and DI-TI-MI-FGSM (DTMI). The core idea of these methods is to improve the efficiency of gradient calculation and the effectiveness of perturbations through different strategies, thereby enhancing the attack performance of adversarial examples.

As shown in Table VI, the SA-WA consistently demonstrated significant improvements in attack performance across different iteration counts. This indicates that SA, by applying additional perturbations to gradient-sensitive regions, can more effectively exploit model vulnerabilities, thereby increasing attack success rates and robustness, while accelerating the generation process of adversarial examples. Moreover, input diversity also played a critical role in gradient calculation, particularly in the DI-MI and DTMI methods. The input diversity effectively prevented gradients from falling into local optima, making the computed gradients more general and effective. This input diversification strategy enables adversarial examples to maintain a high success rate when facing different target models and defense mechanisms.

Overall, the SA-WA can be flexibly integrated with most gradient-based update methods, demonstrating its versatility

and adaptability. This flexibility allows DP-TRAE to seamlessly incorporate different gradient calculation techniques as they evolve, enabling it to adopt more effective strategies to further improve attack performance.

MAE-BA: Subsequently, we conducted an ablation study on the MAE-BA method to investigate the impact of varying enhancement frequency and expansion size on attack performance. Specifically, we examined how different settings for expanding perturbation regions and increasing enhancement frequency affected the adversarial attack efficacy. Notably, it can be seen as our previous work DP-RAE when the expand size is 0. For clarity, we only utilized clean perturbations without employing white-box preprocessing techniques. The experimental results, as illustrated in Figure Figure 4 (a), indicate a significant improvement in attack performance with an increase in the expansion size.

The underlying mechanism can be attributed to the targeted amplification of perturbations around the points that historically demonstrated the highest effectiveness. By selectively enhancing perturbations in these key regions, the attack gains a more precise impact, thereby effectively leveraging the model’s vulnerabilities. However, when the expansion size was allowed to grow without restriction, a decline in attack performance was observed. This deterioration can be explained by the fact that an overly extensive expansion range tends to blur the accurate gradient update direction, leading to the inclusion of numerous irrelevant points and resulting in incorrect gradient updates. Consequently, the efficacy of the attack DI-Minishes as the focus on key perturbation areas becomes diluted.

Moreover, we also observed a notable increase in attack success rate as the historical enhancement frequency was increased in Figure 4 (b). This suggests that frequent reinforcement of previously effective perturbations allows for more persistent and accumulative exploitation of model weaknesses. However, considering the computational overhead associated with frequent updates, we opted to update the perturbations every five iterations. This approach strikes a balance between maintaining a high attack success rate and minimizing the additional computational burden. By selectively tuning both the expansion size and enhancement frequency, the MAE-BA method demonstrated an effective trade-off, achieving robust attack results while managing computational efficiency.

Finally, we tested the impact of different strategy combi-

TABLE VI
ASR (%) ON SEVERAL MODELS UNDER ATTACK SCENARIOS USING RES-50, VGG-16, AND INC-V3 AS THE ENSEMBLE MODELS. WE CONDUCT THESE EXPERIMENTS UNDER TWO METHODS AND REPORT THE ASR WITH ORIGINAL/SA-WA METHOD.

steps = 5	Test model							Ensemble model		
	Attack	RN-34	RN-152	DN-121	VGG-19	AlexNet	Mob-v2	Mob-v3	RN-50	VGG16
BIM	45.0/ 48.3	40.7/ 42.5	47.8 /47.2	73.4/ 75.9	18.9/ 21.9	44.7/ 48.9	14.7/ 16.1	98.6 / 98.6	99.3/99.3	98.8/98.6
MI-FGSM	56.7/ 58.3	51.1/ 52.2	58.6/ 58.9	82.1/ 82.6	27.7/ 30.8	57.2/ 57.8	21.0/ 21.9	99.3 /99.2	99.4 /99.3	98.6/ 98.8
DI-MI	60.2/ 62.1	55.2/ 57.0	64.1/ 65.4	86.0/ 87.9	33.0/ 34.9	64.0/ 67.2	25.8/ 26.7	90.1/ 92.1	98.4/ 99.7	88.7/ 91.5
DTMI	67.9/ 70.3	63.5/ 64.8	71.9/ 75.0	89.7 /89.4	39.9/ 40.3	71.4/ 75.6	38.6/ 40.6	91.1/ 92.9	99.2 /98.0	90.8/ 94.1

steps = 10	Test model							Ensemble model		
	Attack	RN-34	RN-152	DN-121	VGG-19	AlexNet	Mob-v2	Mob-v3	RN-50	VGG16
BIM	53.2/ 54.9	47.3/ 52.2	51.8/ 56.5	81.7/ 82.8	21.4/ 21.7	53.1/ 54.6	16.3/ 17.9	99.5 / 99.5	99.6 /99.4	99.0/ 99.2
MI-FGSM	58.5/ 60.9	55.0/ 56.6	61.2/ 63.4	85.2/ 86.2	29.4/ 30.1	59.2/ 61.3	21.5/ 23.1	99.5/ 99.6	99.5/ 99.6	99.3 / 99.3
DI-MI	72.8/ 76.6	69.7/ 74.5	78.0/ 80.4	93.9/ 95.0	36.5/ 36.7	77.3/ 79.2	29.9/ 30.9	96.8/ 98.6	99.8 / 99.8	96.2/ 97.9
DTMI	80.5/ 84.9	78.3/ 82.8	83.9/ 86.6	95.1/ 96.2	43.1/ 43.4	84.7/ 87.9	43.6/ 44.7	98.3/ 99.3	99.8/ 100	98.7 /98.6

nations on the attack performance. As shown in Figure 4 (c), when the SA-WA is combined with MAE, the attack efficiency improves significantly. This demonstrates that using adversarial perturbations as the initial disturbance can notably enhance performance in black-box model attacks. The combination of these strategies not only improves the precision of the attack but also allows for more effective exploitation of model vulnerabilities, leading to a higher success rate in bypassing the defenses of the black-box models. This result highlights the potential of adversarial perturbations in strengthening the performance of attacks under challenging black-box scenarios.

F. Commercial model attack

To evaluate the effectiveness of our RAE on real-world systems, we conducted tests on Baidu’s cloud vision API¹, an object recognition service. The objective of the attack was to mislead the top-3 categories returned by the API, all while adhering to the constraints of limited queries and perturbations. We selected 50 images for testing and achieved a 92% success rate. Notably, a significant portion of the images were misclassified even before the queries were completed. This underscores the efficiency of the attack strategy, as the white-box perturbations applied at the outset were sufficiently powerful to influence the model’s decision boundaries early in the querying process. Such early-stage perturbations highlight the potential effectiveness of adversarial attacks, particularly in scenarios with constrained query budgets.

As shown in Figure 5, DP-TRAE successfully misclassified the original labels, highlighting the potential threat of our method to commercial black-box models. Considering the limited number of queries allowed by commercial black-box models, we believe that increasing the number of queries can effectively enhance the success rate of the attack.

V. CONCLUSION

In this paper, we introduced the DP-TRAE method, which effectively combines the characteristics of different types of

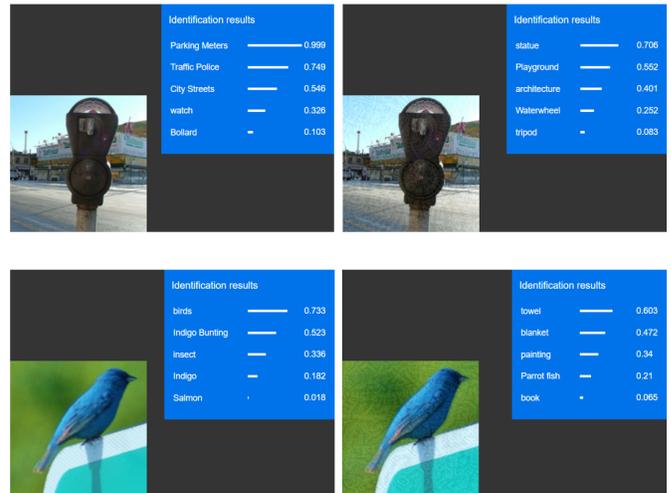


Fig. 5. Results of DP-TRAЕ attacks on a commercial model.

attacks to enhance the protection of sensitive data in complex environments. By leveraging the perturbations generated through white-box attacks, DP-TRAE significantly improves the transferability of adversarial examples, while the black-box attack component ensures targeted attacks on unknown models. Experimental results further demonstrate the superiority of our method, maintaining a high attack success rate even under various defense strategies. Notably, to the best of our knowledge, DP-TRAE is the first method to successfully perform reversible adversarial attacks on commercial black-box models. As a potential future direction, we are looking forward to extending our method to improve the performance of various applications such as large language models [4], [70], [71] and distributed learning system [9], [72]–[74].

ACKNOWLEDGMENTS

This work was supported in part by the Xiamen Research Project for the Natural Science Foundation of Xiamen, China

¹<https://ai.baidu.com/tech/imagerecognition/general>

(3502Z202472028), the Xiamen Science and Technology Plan Project (3502Z20231042), the Xiamen University of Technology High-Level Talent Launch Project (YKJ22041R) and the Fundamental Research Funds for the Central Universities(1082204112364).

REFERENCES

- [1] Xianhao Tian, Peijia Zheng, and Jiwu Huang. Secure deep learning framework for moving object detection in compressed video. *IEEE Transactions on Dependable and Secure Computing*, 21(4):2836–2851, 2023.
- [2] Chunpeng Ge, Zhe Liu, Willy Susilo, Liming Fang, and Hao Wang. Attribute-based encryption with reliable outsourced decryption in cloud computing using smart contract. *IEEE Transactions on Dependable and Secure Computing*, 21(2):937–948, 2023.
- [3] Zheng Lin, Zhe Chen, Zihan Fang, Xianhao Chen, Xiong Wang, and Yue Gao. FedSN: A Federated Learning Framework over Heterogeneous LEO Satellite Networks. *IEEE Trans. Mobile Comput.*, 2024.
- [4] Zihan Fang, Zheng Lin, Zhe Chen, Xianhao Chen, Yue Gao, and Yuguang Fang. Automated Federated Pipeline for Parameter-efficient Fine-tuning of Large Language Models. *arXiv preprint arXiv:2404.06448*, 2024.
- [5] Yuxin Zhang, Haoyu Chen, Zheng Lin, Zhe Chen, and Jin Zhao. Lcfed: An efficient clustered federated learning framework for heterogeneous data. *arXiv preprint arXiv:2501.01850*, 2025.
- [6] Senkang Hu, Zhengru Fang, Zihan Fang, Yiqin Deng, Xianhao Chen, and Yuguang Fang. AgentsCoDriver: Large Language Model Empowered Collaborative Driving with Lifelong Learning. *arXiv preprint arXiv:2404.06345*, 2024.
- [7] Zihan Fang, Zheng Lin, Senkang Hu, Hangcheng Cao, Yiqin Deng, Xianhao Chen, and Yuguang Fang. IC3M: In-Car Multimodal Multi-Object Monitoring for Abnormal Status of Both Driver and Passengers. *arXiv preprint arXiv:2410.02592*, 2024.
- [8] Zheng Lin, Wei Wei, Zhe Chen, Chan-Tong Lam, Xianhao Chen, Yue Gao, and Jun Luo. Hierarchical Split Federated Learning: Convergence Analysis and System Optimization. *IEEE Trans. Mobile Comput.*, 2025.
- [9] Mingda Hu, Jingjing Zhang, Xiong Wang, Shengyun Liu, and Zheng Lin. Accelerating Federated Learning with Model Segmentation for Edge Networks. *IEEE Trans. Green Commun. Netw.*, 2024.
- [10] Haoxuan Yuan, Zhe Chen, Zheng Lin, Jinbo Peng, Yuhang Zhong, Zihang Song, Xiong Wang, and Yue Gao. Graph Learning for Multi-Satellite Based Spectrum Sensing. In *Proc. IEEE ICCT*, pages 1112–1116, 2023.
- [11] Haoxuan Yuan, Zhe Chen, Zheng Lin, Jinbo Peng, Yuhang Zhong, Xuanjie Hu, Songyan Xue, Wei Li, and Yue Gao. Constructing 4D Radio Map in LEO Satellite Networks with Limited Samples. *arXiv preprint arXiv:2501.02775*, 2025.
- [12] Jinbo Peng, Junwen Duan, Zheng Lin, Haoxuan Yuan, Yue Gao, and Zhe Chen. SigChord: Sniffing Wide Non-sparse Multiband Signals for Terrestrial and Non-terrestrial Wireless Networks. *arXiv preprint arXiv:2504.06587*, 2025.
- [13] Zheng Lin, Lifeng Wang, Jie Ding, Bo Tan, and Shi Jin. Channel Power Gain Estimation for Terahertz Vehicle-to-Infrastructure Networks. *IEEE Commun. Lett.*, 27(1):155–159, 2022.
- [14] Yongyang Tang, Zhe Chen, Ang Li, Tianyue Zheng, Zheng Lin, Jia Xu, Pin Lv, Zhe Sun, and Yue Gao. MERIT: Multimodal Wearable Vital Sign Waveform Monitoring. *arXiv preprint arXiv:2410.00392*, 2024.
- [15] Jinshan Liu and Jung-Min Park. “seeing is not always believing”: Detecting perception error attacks against autonomous vehicles. *IEEE Transactions on Dependable and Secure Computing*, 18(5):2209–2223, 2021.
- [16] James Curzon, Tracy Ann Kosa, Rajen Akalu, and Khalil El-Khatib. Privacy and artificial intelligence. *IEEE Transactions on Artificial Intelligence*, 2(2):96–108, 2021.
- [17] Mark Huasong Meng, Guangdong Bai, Sin Gee Teo, Zhe Hou, Yan Xiao, Yun Lin, and Jin Song Dong. Adversarial robustness of deep neural networks: A survey from a formal verification perspective. *IEEE Transactions on Dependable and Secure Computing*, 2022.
- [18] Yazan Otoum, Navya Gottimukkala, Neeraj Kumar, and Amiya Nayak. Machine learning in metaverse security: Current solutions and future challenges. *ACM Computing Surveys*, 2024.
- [19] Tao Ni, Yongliang Chen, Weitao Xu, Lei Xue, and Qingchuan Zhao. Xporter: A study of the multi-port charger security on privacy leakage and voice injection. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pages 1–15, 2023.
- [20] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284, 2019.
- [21] Tao Ni, Jianfeng Li, Xiaokuan Zhang, Chaoshun Zuo, Wubing Wang, Weitao Xu, Xiapu Luo, and Qingchuan Zhao. Exploiting contactless side channels in wireless charging power banks for user privacy inference via few-shot learning. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pages 1–15, 2023.
- [22] Tianyang Duan, Zongyuan Zhang, Zheng Lin, Yue Gao, Ling Xiong, Yong Cui, Hongbin Liang, Xianhao Chen, Heming Cui, and Dong Huang. Rethinking Adversarial Attacks in Reinforcement Learning from Policy Distribution Perspective. *arXiv preprint arXiv:2501.03562*, 2025.
- [23] Tao Ni, Xiaokuan Zhang, Chaoshun Zuo, Jianfeng Li, Zhenyu Yan, Wubing Wang, Weitao Xu, Xiapu Luo, and Qingchuan Zhao. Uncovering user interactions on smartphones via contactless wireless charging side channels. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 3399–3415. IEEE, 2023.
- [24] Carole Cadwalladr and Emma Graham-Harrison. Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach. *The guardian*, 17(1):22, 2018.
- [25] Ju Jia, Yueming Wu, Anran Li, Siqi Ma, and Yang Liu. Subnetwork-lossless robust watermarking for hostile theft attacks in deep transfer learning models. *IEEE transactions on dependable and secure computing*, 2022.
- [26] Zhibo Wang, Hongshan Yang, Yunhe Feng, Peng Sun, Hengchang Guo, Zhifei Zhang, and Kui Ren. Towards transferable targeted adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20534–20543, 2023.
- [27] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. On generating transferable targeted perturbations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2021.
- [28] Yi Ding, Fuyuan Tan, Ji Geng, Zhen Qin, Mingsheng Cao, Kim-Kwang Raymond Choo, and Zhiguang Qin. Interpreting universal adversarial example attacks on image classification models. *IEEE Transactions on Dependable and Secure Computing*, 20(4):3392–3407, 2022.
- [29] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. *Advances in Neural Information Processing Systems*, 32, 2019.
- [30] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7639–7648, 2021.
- [31] Shuai Yuan, Hongwei Li, Xingshuo Han, Guowen Xu, Wenbo Jiang, Tao Ni, Qingchuan Zhao, and Yuguang Fang. Itpatch: An invisible and triggered physical adversarial patch against traffic sign recognition. *arXiv preprint arXiv:2409.12394*, 2024.
- [32] Yao Zhu, Yufeng Chen, Xiaodan Li, Kejiang Chen, Yuan He, Xiang Tian, Bolun Zheng, Yaowu Chen, and Qingming Huang. Toward understanding and boosting adversarial transferability from a distribution perspective. *IEEE Transactions on Image Processing*, 31:6487–6501, 2022.
- [33] C Szegedy. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [34] Seong Joon Oh, Mario Fritz, and Bernt Schiele. Adversarial image perturbation for privacy protection a game theory perspective. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1491–1500. IEEE, 2017.
- [35] Jiang Liu, Chun Pong Lau, and Rama Chellappa. Diffprotect: Generate adversarial examples with diffusion models for facial privacy protection. *arXiv preprint arXiv:2305.13625*, 2023.
- [36] Mingfu Xue, Shichang Sun, Zhiyu Wu, Can He, Jian Wang, and Weiqiang Liu. Socialguard: An adversarial example based privacy-preserving technique for social images. *Journal of Information Security and Applications*, 63:102993, 2021.
- [37] Jiayang Liu, Weiming Zhang, Kazuto Fukuchi, Youhei Akimoto, and Jun Sakuma. Unauthorized ai cannot recognize me: Reversible adversarial example. *Pattern Recognition*, 134:109048, 2023.
- [38] Lizhi Xiong, Yue Wu, Peipeng Yu, and Yuhui Zheng. A black-box reversible adversarial example for authorizable recognition to shared images. *Pattern Recognition*, 140:109549, 2023.
- [39] Jiawei Zhang, Jinwei Wang, Hao Wang, and Xiangyang Luo. Self-recoverable adversarial examples: A new effective protection mechanism

- in social networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2):562–574, 2022.
- [40] Zhaoxia Yin, Yinyin Peng, and Youzhi Xiang. Reversible data hiding in encrypted images based on pixel prediction and bit-plane compression. *IEEE Transactions on Dependable and Secure Computing*, 19(2):992–1002, 2020.
- [41] Zhenxing Qian, Hang Zhou, Xinpeng Zhang, and Weiming Zhang. Separable reversible data hiding in encrypted jpeg bitstreams. *IEEE Transactions on Dependable and Secure Computing*, 15(6):1055–1067, 2016.
- [42] Jiajie Zhu, Xia Du, Jizhe Zhou, Chi-Man Pun, Qizhen Xu, and Xiaoyuan Liu. Dp-rae: A dual-phase merging reversible adversarial example for image privacy protection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 671–680, 2024.
- [43] Zhaoxia Yin, Hua Wang, Li Chen, Jie Wang, and Weiming Zhang. Reversible adversarial attack based on reversible image transformation. *arXiv preprint arXiv:1911.02360*, 2019.
- [44] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [45] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.
- [46] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.
- [47] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019.
- [48] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4312–4321, 2019.
- [49] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International conference on machine learning*, pages 2484–2493. PMLR, 2019.
- [50] Thibault Maho, Teddy Furon, and Erwan Le Merrer. Surfree: a fast surrogate-free black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10430–10439, 2021.
- [51] Rajarathnam Chandramouli and Nasir Memon. Analysis of lsb based image steganography techniques. In *Proceedings 2001 international conference on image processing (Cat. No. 01CH37205)*, volume 3, pages 1019–1022. IEEE, 2001.
- [52] Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. *Digital watermarking and steganography*. Morgan kaufmann, 2007.
- [53] Dongdong Hou, Weiming Zhang, Kejiang Chen, Sian-Jheng Lin, and Nenghai Yu. Reversible data hiding in color image with grayscale invariance. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(2):363–374, 2018.
- [54] Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. Hinet: Deep image hiding by invertible network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4733–4742, 2021.
- [55] J Zhu. Hidden: hiding data with deep networks. *arXiv preprint arXiv:1807.09937*, 2018.
- [56] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.
- [57] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.
- [58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [59] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [60] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [61] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [62] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [63] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [64] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.
- [65] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [66] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017.
- [67] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.
- [68] Haya Brama and Tal Grinshpoun. Heat and blur: an effective and fast defense against adversarial examples. *arXiv preprint arXiv:2003.07573*, 2020.
- [69] Aamir Mustafa, Salman H Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29:1711–1724, 2019.
- [70] Senkang Hu, Zhengru Fang, Zihan Fang, Yiqin Deng, Xianhao Chen, Yuguang Fang, and Sam Kwong. AgentsCoMerge: Large Language Model Empowered Collaborative Decision Making for Ramp Merging. *arXiv preprint arXiv:2408.03624*, 2024.
- [71] Zheng Lin, Yuxin Zhang, Zhe Chen, Zihan Fang, Xianhao Chen, Praneeth Vepakomma, Wei Ni, Jun Luo, and Yue Gao. HSPLITLoRA: A Heterogeneous Split Parameter-Efficient Fine-Tuning Framework for Large Language Models. *arXiv preprint arXiv:2505.02795*, 2025.
- [72] Yuxin Zhang, Haoyu Chen, Zheng Lin, Zhe Chen, and Jin Zhao. FedAC: A Adaptive Clustered Federated Learning Framework for Heterogeneous Data. *arXiv preprint arXiv:2403.16460*, 2024.
- [73] Zheng Lin, Yuxin Zhang, Zhe Chen, Zihan Fang, Cong Wu, Xianhao Chen, Yue Gao, and Jun Luo. Leo-Split: A Semi-Supervised Split Learning Framework over LEO Satellite Networks. *arXiv preprint arXiv:2501.01293*, 2025.
- [74] Yuxin Zhang, Zheng Lin, Zhe Chen, Zihan Fang, Wenjun Zhu, Xianhao Chen, Jin Zhao, and Yue Gao. SatFed: A Resource-Efficient LEO Satellite-Assisted Heterogeneous Federated Learning Framework. *arXiv preprint arXiv:2409.13503*, 2024.