Towards Model Extraction Attacks in GAN-Based Image Translation via Domain Shift Mitigation

Di Mi[1], **Yanjun Zhang**[2], **Leo Yu Zhang**[3],
Shengshan Hu[4], **Qi Zhong**[5], **Haizhuan Yuan**[1], **Shirui Pan**[3]
[1]Xiangtan University
[2]University of Technology Sydney
[3]Griffith University
[4]Huazhong University of Science and Technology
[5]City University of Macau
midi_xtu@163.com, yanjun.zhang@uts.edu.au, leo.zhang@griffith.edu.au,
hushengshan@hust.edu.cn, qizhong@cityu.edu.mo, yhz@xtu.edu.cn, s.pan@griffith.edu.au

Abstract

Model extraction attacks (MEAs) enable an attacker to replicate the functionality of a victim deep neural network (DNN) model by only querying its API service remotely, posing a severe threat to the security and integrity of pay-per-query DNN-based services. Although the majority of current research on MEAs has primarily concentrated on neural classifiers, there is a growing prevalence of image-to-image translation (I2IT) tasks in our everyday activities. However, techniques developed for MEA of DNN classifiers cannot be directly transferred to the case of I2IT, rendering the vulnerability of I2IT models to MEA attacks often underestimated. This paper unveils the threat of MEA in I2IT tasks from a new perspective. Diverging from the traditional approach of bridging the distribution gap between attacker queries and victim training samples, we opt to mitigate the effect caused by the different distributions, known as the domain shift. This is achieved by introducing a new regularization term that penalizes high-frequency noise, and seeking a flatter minimum to avoid overfitting to the shifted distribution. Extensive experiments on different image translation tasks, including image super-resolution and style transfer, are performed on different backbone victim models, and the new design consistently outperforms the baseline by a large margin across all metrics. A few real-life I2IT APIs are also verified to be extremely vulnerable to our attack, emphasizing the need for enhanced defenses and potentially revised API publishing policies.

1 Introduction

Deep neural networks (DNNs) have exhibited remarkable success in diverse domains, driven by substantial investments in data processing, computational power, and expertise knowledge (Pouyanfar et al. 2018). This success has been capitalized through the introduction of pay-per-query API services[4]. However, recent research has uncovered a notable vulnerability: model extraction attacks (MEAs). These attacks empower an adversary to replicate the remote DNN's functionality by crafting surrogate models (Tramèr et al. 2016). Consequently, this allows unauthorized access to the service, enabling the adversary to launch adversarial attacks or privacy attacks on the service provider (Papernot et al. 2017; Zhang et al. 2022b; Ma et al. 2023; Ye et al. 2022). Such vulnerabilities not only undermine the security of the entire model supply chain but also pose a critical challenge to the integrity of DNN-based services.

Despite the apparent simplicity of this attack process, a significant challenge persists for the adversary: the queried samples are unlikely to perfectly match the secret training dataset utilized for training the victim model, then how can the MEA be executed in a manner that ensures both effectiveness and efficiency, given this discrepancy between queried and training samples?

To close the distribution gap between query and training samples, numerous works have been proposed for various classification tasks, including image processing (Orekondy, Schiele, and Fritz 2019; Pal et al. 2020; Yuan et al. 2022; Barbalau et al. 2020), NLP processing (Krishna et al. 2019; Xu et al. 2021), and graph data processing (Wu et al. 2022; Shen et al. 2022). For instance, in image classification, (Orekondy, Schiele, and Fritz 2019) suggested using reinforcement learning, and (Pal et al. 2020) suggested using active learning as a means to identify better query samples. Alternatively, (Yuan et al. 2022; Barbalau et al. 2020) advocated aligning the distribution of the secret training dataset with Generative Adversarial Networks (GANs). In NLP tasks, (Xu et al. 2021) suggested targeting an ensemble of victim APIs simultaneously, all providing the same service, to make them implicitly vote for good query samples.

Concurrent with classification tasks, image-to-image translation (I2IT) tasks, such as image super-resolution or restoration, constitute another significant application domain of DNNs. As a testament to this prevalence, Table 1 makes a comparison of the pay-per-use price between classification and I2IT tasks on different platforms[1,2]. Nonetheless, the MEA vulnerabilities in I2IT are rarely explored, making their risk largely underestimated.

We attribute this to the natural difference between study-

[1]Providing image translation service: https://cloud.baidu.com, https://imglarger.com, https://vanceai.com.

[2]Providing classification service: https://cloud.baidu.com, https://aws.amazon.com, https://cloud.google.com.

Image translation		
Baidu AI Cloud	Imglarger	VanceAI
\$0.0064-0.0614	\$0.09	\$0.035
Classification		
Amazon	Google Cloud	Baidu AI Cloud
\$0.0008-0.001	\$0.001-0.0015	\$0.00029-0.00041

Table 1: Pay-per-use price (/image) comparison on different APIs[1,2].

ing MEA on classification models and I2IT models. In particular, extracting a classification model corresponds to effectively identifying its decision boundary (Pal et al. 2020). Hence, any alteration in the label or confidence scores of a queried sample (results yielded by the victim API) implies the sample is crossing the decision boundary or changing its proximity to the boundary, which can be exploited directly by the adversary to select better queries. In contrast, I2IT models take images as inputs and produce images as their outputs. It remains unknown what kind of information regarding the victim model the output images carry, and which output image carries more.

Considering this discrepancy, this paper introduces an innovative approach to initiate MEA on I2IT models. The core idea is to directly mitigate the domain shift problem when training the surrogate. To achieve this, we design two complemented components: one regulates the behavior of the surrogate to suppress noisy components in translation outcomes while reducing model complexity, and the other pursues a flatter optimum to avoid overfitting to the shifted distribution. To our best knowledge, this is the first time that domain shift mitigation techniques is investigated in the context of MEAs. The contribution of this work is twofold:

- Besides the traditional wisdom of closing the distribution discrepancy in MEA, we, for the first time, highlight that mitigating the domain shift constitutes another angle for launching MEA attacks. This approach proves especially advantageous in scenarios where how to select better queries is not clear. This fresh angle on MEA attacks is of independent research interest.
- We apply concrete domain shift mitigation strategies (*i.e.*, wavelet regularization and sharpness-aware minimization) to extract GAN-based models in I2IT tasks. Extensive experimental results in controlled laboratory conditions and real-world scenarios corroborate that MEA is a real threat to image translation systems.

2 Background and Related Work

GAN-Based Image Translation

GAN functions by training two competing models to ultimately learn the unknown true distribution, $P_{\text{data}}(x)$, of the training data X. The generator model G, which creates a synthetic image $G(z)$ from a random variable z, and the discriminator model D, which operates as a binary classifier to distinguish $G(z)$ from true image $x \sim P_{\text{data}}(x)$ are obtained by solving

$$\mathcal{L}_{\text{GAN}} = \min_G \max_D \; \mathbb{E}_{x\sim P_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim P_z}[\log(1 - D(G(z)))], \tag{1}$$

where P_z is a random distribution. Upon convergence (*i.e.*, when $G(z)$ approximates $P_{\text{data}}(x)$), G can produce high-quality and photo-realistic samples, rendering it valuable for many image processing tasks.

By expanding the role of D to differentiate between images from a source domain and those from a target domain (*e.g.*, image super-resolution or style transfer), GAN and its variants have gained widespread application in I2IT, which is the primary concern of this work. According to (Pang et al. 2021), I2IT can be divided into supervised and unsupervised based on whether the training samples in the source and the target domain are paired or not. This work considers the widely used supervised framework Pix2Pix (Wang et al. 2018a) and the unsupervised framework CycleGAN (Kim et al. 2019). Details can be found in Supp.-A.

Model Extraction Attacks

As mentioned in Sec. 1, by taking the victim model F_V as a labeling oracle, the goal of MEA involves creating an attack model $F_{\mathcal{A}}$ that emulates the functionalities of the victim model F_V. In classification tasks, this goal translates to finding the decision boundary while reducing the query budget under the assumption of attackers' knowledge. Strategies include selecting better query samples with reinforcement learning (Orekondy, Schiele, and Fritz 2019) or active learning (Pal et al. 2020), or even harnessing GAN to generate images that are close to the secret training data of F_V (Yuan et al. 2022; Barbalau et al. 2020).

It is crucial to note that such strategies cannot be readily transferred to MEAs of I2IT models. In classification tasks, the outputs of the victim model F_V are labels or confidence scores, which inherently carry rich information about the decision boundary of F_V. In contrast, I2IT victim models merely return translated images upon query and do not expose their latent embeddings to attackers. Moreover, it remains unknown (Hu and Pang 2021; Szyller et al. 2021) what kind of translated images contain more information about the victim F_V, and this effect is exacerbated when the attacker's query data distribution deviates from the secret training data of F_V (*i.e.*, the domain shift problem). As such, instead of looking for better query samples, we adopt an alternative route that directly mitigates the effect caused by domain shift when extracting the underlying models.

3 Threat Model

Adversary's Knowledge

We consider a victim model $F_V : X \to Y$ that translates images from a source domain X to a target domain Y has been well-trained on a secret dataset $D_V = (D_V^X, D_V^Y)$. The victim F_V has been employed as the backbone to provide API service to remote users. The attacker $\mathcal{A}$ is knowledgeable of the general service domains (*e.g.*, translating horse pictures to zebra), but he cannot access the structure, parameters, hyperparameters, and the secret training dataset of F_V.

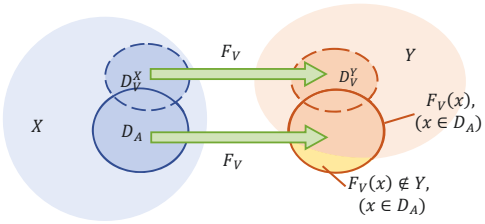

Figure 1: An illustration of the domain shift problem: The problem arises when there is a disparity between the domains of the victim model's training data and the attack data. This mismatch causes certain attack data to be incorrectly mapped to the target domain.

The attacker first constructs his own training dataset $D_{\mathcal{A}}$ by querying the API service with public samples $x \in D_{\mathcal{A}}^X$ and collecting labeled pairs $(x, F_V(x))$. Then $\mathcal{A}$ develops a local attack model $F_{\mathcal{A}}$ that mimics the functionality of F_V by solving

$$\min_{F_{\mathcal{A}}} \mathbb{E}_{x \sim p_{\mathcal{A}}(x)}\, d(F_{\mathcal{A}}(x), F_V(x)), \tag{2}$$

where $p_{\mathcal{A}}(x)$ represents the data distribution of $D_{\mathcal{A}}^X$ and d is a distance metric.

Adversary's Goals

We consider two typical adversary's goals following the existing literature of MEA (Jagielski et al. 2020).

Functionality Extraction aims to obtain a local replica that is capable of completing the intended function of the victim model. Here, we define the Functional Completion Degree ($R_{\text{capability}}$) to assess the capability of $F_{\mathcal{A}}$ in accomplishing the I2IT task. $R_{\text{capability}}$ represents the distance between the region into which the source domain falls after being mapped by the attack model and the target domain. It can be written as

$$R_{\text{capability}} = \mathbb{E}_{x \sim p_{\text{test}}(x)}\, d(F_{\mathcal{A}}(x), Y),$$

where $p_{\text{test}}(x)$ is the distribution of the test data. It's essential to note that $F_{\mathcal{A}}$ might even outperform F_V.

Fidelity Extraction is to minimize the discrepancy between the output distribution of the attack model and that of the victim model. We define the Output Fidelity (R_{fidelity}) as

$$R_{\text{fidelity}} = \mathbb{E}_{x \sim p_{\text{test}}(x)}\, d(F_{\mathcal{A}}(x), F_V(x)).$$

It is noted that $R_{\text{capability}}$ and R_{fidelity} can vary independently. In the situation that the attacker lacks access to the target domain Y, R_{fidelity} is a practical metric to evaluate the effectiveness of MEA.

4 Method

Domain Shift Problem

The domain shift problem, as discussed in (Glorot, Bordes, and Bengio 2011), typically pertains to the decline in model performance caused by disparities between the distribution of training data and that of test data. In this work, we identify

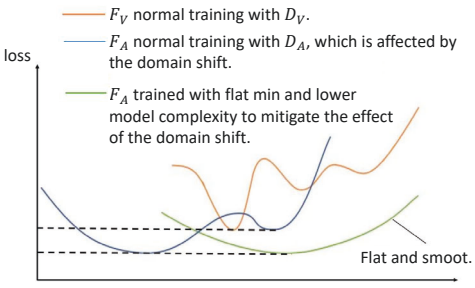

Figure 2: Illustration of the effect of domain shift mitigation.

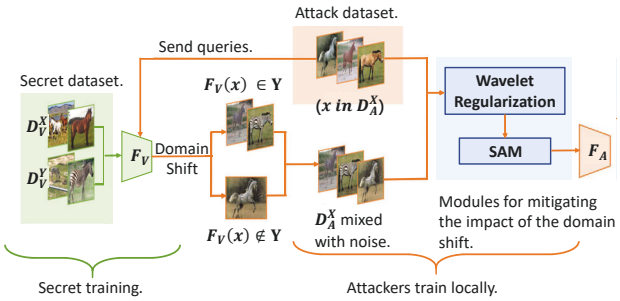

Figure 3: Approach overview.

this phenomenon as a fundamental factor affecting the performance of MEA. In MEAs, attackers lack access to the secret data used to train the victim model and they usually only use public datasets as attack data. When a disparity exists between the distribution of this public data and the distribution of the secret data, it causes the attack model's output quality to decrease and can even result in deviations from the target domain. Such data that fails to effectively represent the mapping relationship between the source and target domains can be considered noise data, which hinders the training of subsequent attack models.

Fig. 1 demonstrates such a scenario. The victim model F_V is trained to transfer images from the source domain X to the target domain Y. However, the knowledge it has learned is solely based on the data distributions of D_V^X and D_V^Y in its training dataset D_V. Consequently, the domain shift between D_V^X and $D_{\mathcal{A}}^X$ can lead to situations where the model's output may not necessarily correspond to the target domain, meaning that $F_V(x)$ might not be a valid representation in Y for images $x \in D_{\mathcal{A}}^X$. The situation becomes worse when the attacker utilizes a publicly available dataset. In such cases, the discrepancies between the data distribution in D_V^X and $D_{\mathcal{A}}^X$ could be significant, exacerbating the challenges of mapping the model's output to the target domain accurately. We visualize this discrepancy with real examples from style transfer tasks in Supp.-B.

Attack Overview

For MEA of classification tasks, traditional wisdom aims at closing the distributional gap between the victim dataset and the attack dataset (Orekondy, Schiele, and Fritz 2019; Pal et al. 2020; Yuan et al. 2022). Due to the unique challenge

of I2IT tasks, this work aims to address the impact of the domain shift problem from an orthogonal perspective by resorting to a flatter and smoother loss landscape for the attack model. Fig. 2 illustrates our insight. In the sketch, we briefly illustrate the effect of different factors on the model's test loss. Under vanilla training of the attack model, the obtained $F_{\mathcal{A}}$ will be fitted to the distribution of the attack's training set $D_{\mathcal{A}}$, hence deviating from the original optima of F_V. In contrast, when the model is enforced to be with lower complexity (*i.e.*, smooth) around a flat minimum area, such local overfitting introduced by domain shift can be mitigated.

To achieve this, we introduce the following two tailored components (as shown in Fig. 3). We first construct a wavelet regularization term from a frequency perspective. This concept draws inspiration from recent investigations (Zhang et al. 2022a) which highlight a particular property of GANs' behavior within the frequency domain. Specifically, GANs tend to exhibit low errors within the low-frequency range but usually fail to produce high-quality results in the high-frequency bands. We therefore apply the discrete wavelet transform (DWT) to decompose images in $D_{\mathcal{A}}^X$ (i.e., the attackers' input images) and $D_{\mathcal{A}}^Y$ (i.e., the images generated by the victim model's from the attacker's images) into different frequency bands, and penalize the L1 distance on the high-frequency band. The wavelet regularization term can effectively reduce the complexity of the I2IT network. This, in turn, encourages consistent outputs between F_V and $F_{\mathcal{A}}$, particularly enhancing finer image details within the high-frequency band.

We then train the attack model using sharpness-aware minimization (SAM). Due to the noise caused by domain shifting (*i.e.*, $D_V \rightarrow D_{\mathcal{A}}$), the inherent issue of mode collapse during GAN training can be exacerbated, which leads the model to overly specialize in certain patterns during the generation process, resulting in an overfitting to partial data (d'Ascoli, Sagun, and Biroli 2020). Recent studies have demonstrated that SAM (Foret et al. 2020) optimizer has held the promise of seeking out flatter minima by simultaneously minimizing loss value and loss sharpness. However, to the best of our knowledge, there have been no prior works reporting the process of training GANs using SAM. We therefore introduce a GAN-specific SAM variant towards a wide minimum and further improve the effectiveness of MEA against the I2IT network.

Wavelet Regularization

To construct a wavelet regularization term, we first decompose the output image using the DWT into four distinct subbands: low frequency (LL), low-high frequency (LH), high-low frequency (HL), and high-high frequency (HH). We assume the image intended for DWT as c. This can be represented as $\Psi(c) = \{LL(c), LH(c), HL(c), HH(c)\}$. We denote $\Psi_p^H(c) = \{LH(c), HL(c), HH(c)\}$ as high-frequency terms, where p represents the times DWT applied. We then use the L1 distance to construct the wavelet regularization term as

$$L_w^p = \mathbb{E}_{x \sim p_{\mathcal{A}}(x)} \left\| \left(\Psi_p^H \circ F_{\mathcal{A}}\right)(x) - \left(\Psi_p^H \circ F_V\right)(x) \right\|_1 .$$

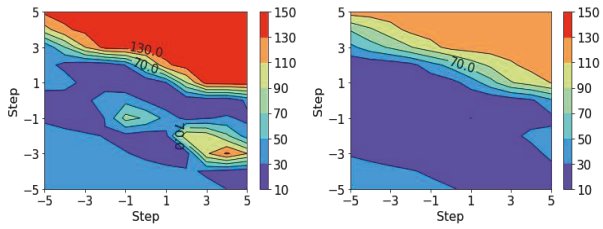

Figure 4: Comparison of model loss landscape. Left and right are losses trained with Adam and SAM, respectively.

As a result, the overall loss function, denoted as L, becomes

$$L = L_o + \alpha L_w^p, \tag{3}$$

where L_o is the vanilla loss function of the model backbone (*e.g.*, Pix2Pix or CycleGAN) in I2IT tasks, and α is the coefficient used to balance between L_o and L_w^p.

By incorporating L_w^p into the training process, our objective is to minimize the disparity in high-frequency information between the victim model's output ($F_V(x)$) and the attack model's output ($F_{\mathcal{A}}(x)$). The regularization term can also penalize the model during training, thereby reducing model complexity and alleviating the overfitting to the noise.

SAM for GAN-Based I2IT

Besides designing a new regularization term (*i.e.*, Eq. 3), we also investigate domain shift mitigation from the view of optimizers. Assume that the loss function of the model to be optimized is L and the model parameters are $\mathbf{w}$, loss sharpness is defined as

$$\max_{\|\epsilon\|_2 \leq \rho} L(\mathbf{w} + \epsilon) - L(\mathbf{w}),$$

where ρ ($\rho \geq 0$) is the neighborhood size. SAM (Foret et al. 2020) optimizes both the loss function and the loss sharpness by solving

$$\min_{\mathbf{w}} \left(\max_{\|\epsilon\|_2 \leq \rho} L(\mathbf{w} + \epsilon) - L(\mathbf{w}) \right) + L(\mathbf{w}).$$

With the first-order Taylor expansion, the value ϵ that solves the inner maximization is

$$\epsilon(\mathbf{w}) = \rho \cdot \frac{\nabla_{\mathbf{w}} L(\mathbf{w})}{\|\nabla_{\mathbf{w}} L(\mathbf{w})\|_2}.$$

Substitute $\epsilon(\mathbf{w})$ back, the minimization can be approximately solved as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha_t \cdot \nabla_{\mathbf{w}} L(\mathbf{w})|_{\mathbf{w}+\epsilon(\mathbf{w})}, \tag{4}$$

where α_t is the learning rate at time step t. Note that solving Eq. 4 requires solving $\epsilon(\mathbf{w})$ first, and we use Adam twice in our experiments.

Without loss of generality, consider the backbone model used for I2IT consists of N generators $G = \{G_1, ..., G_N\}$ and M discriminators $D = \{D_1, ..., D_M\}$, with their respective parameters being $\mathbf{w}^G = \{\mathbf{w}^{G_1}, \ldots, \mathbf{w}^{G_N}\}$ and

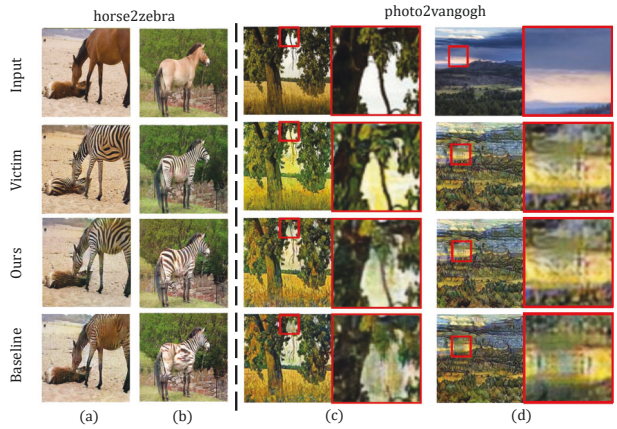

Figure 5: Qualitative results of our attack against the style transfer task. (a-b). horse2zebra, (c-d). photo2vangogh.

$\mathbf{w}^D = \{\mathbf{w}^{D_1}, \ldots, \mathbf{w}^{D_N}\}$. Referring to Eq. 1, we alternatively optimize G and D with SAM optimizer by

$$\begin{cases} \epsilon(\mathbf{w}^{G_i}) = \rho_{G_i} \cdot \dfrac{\nabla_{\mathbf{w}^{G_i}} L(\mathbf{w})}{\|\nabla_{\mathbf{w}^{G_i}} L(\mathbf{w})\|_2}, \\ g(\mathbf{w}^{G_i}) = \nabla_{\mathbf{w}^{G_i}} L(\mathbf{w}^{G_i})|_{\mathbf{w}^{G_i}+\epsilon(\mathbf{w}^{G_i})}, \\ \mathbf{w}^{G_i}_{t+1} = \mathbf{w}^{G_i}_t - \alpha_t \cdot g(\mathbf{w}^{G_i}), \end{cases}$$

for each $G_i \in G$ $(i \in [1, N])$ and

$$\begin{cases} \epsilon(\mathbf{w}^{D_j}) = \rho_{D_j} \cdot \dfrac{\nabla_{\mathbf{w}^{D_j}} L(\mathbf{w})}{\|\nabla_{\mathbf{w}^{D_j}} L(\mathbf{w})\|_2}, \\ g(\mathbf{w}^{D_j}) = \nabla_{\mathbf{w}^{D_j}} L(\mathbf{w}^{D_j})|_{\mathbf{w}^{D_j}+\epsilon(\mathbf{w}^{D_j})}, \\ \mathbf{w}^{D_j}_{t+1} = \mathbf{w}^{D_j}_t + \alpha_t \cdot g(\mathbf{w}^{D_j}), \end{cases}$$

for each $D_j \in D$ $(j \in [1, M])$. Here, ρ_{G_i} and ρ_{D_j} are the loss sharpness hyper-parameters of G_i and D_j, respectively.

We summarize the process in algorithm in the Supp.-C. It is imperative to highlight that the scenario under consideration resembles that of the CycleGAN model, where two generators and two discriminators are used and the generators iterate concurrently. However, if certain generator parameters require individual iteration, they should have respective optimization directions computed separately. As gradients need to be computed twice, the SAM requires forward propagation to be performed twice in each iteration.

We conduct experiments in the horse2zebra task, and visualize the loss landscape of the Pix2Pix backbone by training models with different optimizers in Fig. 4. After applying SAM, the loss landscape of F_A becomes flatter.

5 Experimental Analyses

Experimental Setup

Dataset and Models. We assess the performance of our attack on typical I2IT tasks, *i.e.*, style transfer tasks including horse2zebra (converting horse images to zebras) and photo2vangogh (converting photos to Van Gogh style), as well as the super-resolution task (enhancing anime images resolution).

To build the victim model, we use CycleGAN (Liu, Breuel, and Kautz 2017) to train the style transfer tasks, and we directly employ the pre-trained real-ESRGAN model (Wang et al. 2021) for the super-resolution task. The datasets employed to train the victim style transfer models are horse2zebra and photo2vangogh in (Zhu and et al. 2021). The datasets used for training the victim super-resolution model are DIV2K (Agustsson and Timofte 2017), Flickr2K (Timofte and et al. 2017), and OutdoorSceneTraining (Wang et al. 2018b). All of these are identical to the original settings (Liu, Breuel, and Kautz 2017; Wang et al. 2021). All victim models have demonstrated compelling performance in their respective tasks.

To train the model, we use Pix2Pix (Isola et al. 2017) and CycleGAN (Liu, Breuel, and Kautz 2017) as attack modeling backbone frameworks for style transfer and super-resolution tasks, respectively.

Since attacker $\mathcal{A}$ is knowledgeable of the general service domains, we construct $\mathcal{A}$'s training dataset $D^X_{\mathcal{A}}$ as follows. For the horse2zebra task, we employ horse images from the Animal10 dataset (Zhu et al. 2019). For the photo2vangogh task, we utilize a subset of two thousand landscape images from the Landscape dataset (Rougetet 2020). Regarding the super-resolution task, we use the Anime dataset (Chen 2020), which comprises images sourced from anime films created by Makoto Shinkai, Hayao Miyazaki and Kon Satoshi.

We employ the test set of the victim model to evaluate the performance on style transfer tasks. We select 200 images from Anime dataset (Chen 2020) (ensuring they are disjoint with the training dataset used for the attack) as the test set for the super-resolution task. Further details about the setup can be found in Supp.-D.

Evaluation Metrics. We assess the image quality of our attack using widely adopted metrics: Fréchet Inception Distance (FID) (Heusel et al. 2017) and Kernel Inception Distance (KID) (Bińkowski et al. 2018). Both FID and KID measure the divergence between image distributions, with lower scores implying higher similarity. Additionally, we employ PSNR (Peak Signal-to-Noise Ratio) and LIPIPS (Zhang et al. 2018) at both pixel and perceptual levels. PSNR calculates pixel-wise differences between images, with higher values denoting greater similarity. Whereas, LPIPS incorporates perception-based features, with lower values indicating increased similarity.

Baseline. We consider the Artist-Copy (Szyller et al. 2021) as the baseline since it is currently the only known MEA in GAN-based I2IT.

Attack Performance

Style Transfer. Compared to the baseline, our method demonstrates significant performance improvement in all comparative experiments.

As shown in Table 2, in the horse2zebra task, the FID/KID scores of our attack reach 82.63/2.55 for $R_{\text{capability}}$ and 57.87/0.32 for R_{fidelity}, significantly surpassing the Artist-Copy method which achieves 115.09/5.59 and 75.65/1.93

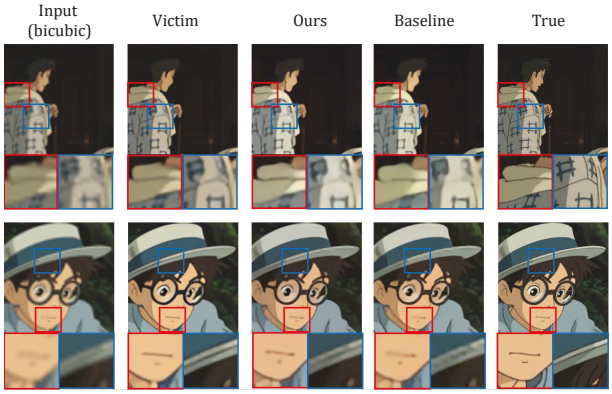

Figure 6: Qualitative results of our attack against the super-resolution task (4x upscaling).

Task	Method	$R_{\text{capability}}$		R_{fidelity}	
		FID↓	KID↓	FID↓	KID↓
h2z†	Victim	63.08	1.21±0.09	NA	NA
	A-C	115.09	5.59±0.39	75.65	1.93±0.25
	Ours	**82.63**	**2.55±0.24**	**57.87**	**0.32±0.09**
p2v†	Victim	109.78	2.73±0.43	NA	NA
	A-C	112.31	3.01±0.41	59.83	1.70±0.27
	Ours	**110.00**	**2.66±0.37**	**59.09**	**1.45±0.25**

‡ Since the consistency between the victim model's output and itself is not required, the R_{fidelity} values are represented as NA. KID is reported as KID×100±std.×100, the same hereinafter. † We use h2z and p2v to represent horse2zebra and photo2vangogh.

Table 2: Quantitative results of our attack against the style transfer tasks‡ (A-C represents Artist-Copy).

FID/KID scores for $R_{\text{capability}}$ and R_{fidelity}, respectively. In the photo2vangogh task, our attack achieves 110.00/2.66 and 59.09/1.45 FID/KID scores for $R_{\text{capability}}$ and R_{fidelity}. In contrast, the Artist-Copy method only attains FID/KID values of 112.31/3.01 for $R_{\text{capability}}$ and 59.83/1.70 for R_{fidelity}. It is worth mentioning that in the horse2zebra task, the attack model shows the highest improvement, with a decrease of 32.46 in $R_{\text{capability}}$ FID. Moreover, in the photo2vangogh task, our model achieves performance on $R_{\text{capability}}$ that is almost as good as the victim model.

Fig. 5 demonstrates the effect of our attack in style transfer tasks. As shown in Fig. 5 (a-b), our attack successfully extracts the functionality of the victim model, enabling the transformation of horses into zebras while the Artist-Copy method struggles to properly add zebra patterns to the horses in the horse2zebra task. Fig. 5 (c-d) illustrates the attack performance against the photo2vangogh task with zoomed-in views of the outputs in order to demonstrate our attack's capacity in generating fine details of the images. We can observe that our attack has achieved closer outputs to the victim model and it handles the details better due to its ability in generating higher-quality high-frequency information. However, the Artist-Copy method shows significant abnormal noise and artifacts in the output due to overfitting on noisy data induced by domain shift problems. Additional results can be found in Supp.-E.

Method	$R_{\text{capability}}$		R_{fidelity}	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
Victim	34.56	0.067	NA	NA
Artist-Copy	27.67	0.177	27.89	0.156
Ours	**29.73**	**0.145**	**29.73**	**0.141**

Table 3: Quantitative results of our attack against the super-resolution task (4x upscaling).

Notably, our attack necessitates a significantly smaller number of queries compared to the baseline. For example, the performance of our attack using half the number of queries (2-3k queries) surpasses that of the baseline (4-5k queries). Results can be found in Supp.-E).These findings highlight the impact of the domain shift issue, a factor that significantly limits the effectiveness of traditional MEAs which rely on increasing query counts to enhance attack capabilities. In contrast, our MEA, incorporating methods that address the domain shift problem, achieves favorable performance with a substantially reduced number of queries.

Super-resolution. Our method also achieves better attack performance in the super-resolution task (4x upscaling). In Table 3, we present the experimental results. The metrics in the table are computed as the averages of the test set images.

Compared to the Artist-Copy, our method shows a 2.06 increase in PSNR for $R_{\text{capability}}$ and a decrease of 0.032 in LIPIS. Moreover, our method exhibits a 2.16 increase in PSNR for R_{fidelity} and a decrease of 0.015 in LIPIS.

In Fig. 6, we compare the performance of bicubic interpolation (i.e., a common method for image upscaling), victim, Artist-Copy, and our method in the super-resolution task. We can observe that our method's output shows a significant improvement in image clarity compared to the bicubic method, demonstrating the success of the attack. Furthermore, through an analysis of the performance on the detailed parts of Fig. 6, it can be noted that our method's output presents texture details with greater clarity compared to the Artist-Copy. For example, in the second row of Fig. 6, it becomes evident that our method's output exhibits significantly sharper edges on the hat and more distinct lines around the mouth. This observation further implies that our approach consistently aligns with the performance of the victim model, particularly in capturing high-frequency elements within the frequency domain.

Ablation Study

We now evaluate the effectiveness of the wavlet regularization and SAM, respectively, in model extraction attacks against the horse2zebra task. Table 4 shows that both the FID and KID scores of $R_{\text{capability}}$ and R_{fidelity} undergo a significant increase when either the SAM or the wavelet regularization term is removed, which demonstrates the necessity and effectiveness of each component in our method.

The analysis reveals that the use of the wavelet regularization term alone on top of the baseline gives a better performance in terms of both R_{fidelity} and $R_{\text{capbility}}$ because it promotes the consistency of the outputs of the victim model and the attack model in the frequency domain. On top of it,

Component		$R_{\text{capability}}$		R_{fidelity}	
L_w^p	SAM	FID↓	KID↓	FID↓	KID↓
×	×	115.09	5.59±0.39	75.65	1.93±0.25
√	×	100.69	4.25±0.30	68.01	1.06±0.17
×	√	104.04	4.37±0.30	70.72	1.25±0.22
√	√	**82.63**	**2.55±0.24**	**57.87**	**0.32±0.09**

Table 4: Ablation Experiment Results.

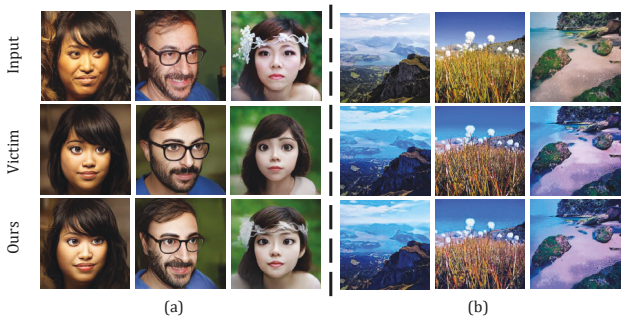

Figure 7: The performance of our attack against commercial I2IT services. (a). face2cartoon (3d), (b). landscape2cartoon.

Task	Method	FID↓	KID↓
face2cartoon (2d)	Artist-Copy	63.20	1.79±0.27
	Ours	**59.61**	**1.36±0.2**
face2cartoon (3d)	Artist-Copy	60.89	1.43±0.18
	Ours	**51.67**	**0.61±0.13**
landscape2cartoon	Artist-Copy	68.55	1.47±0.21
	Ours	**64.65**	**1.21±0.20**

Table 5: The performance of our attack against commercial I2IT services.

SAM can further improve the attack performance due to its ability in finding a much flatter optimum.

6 MEA Againsnt I2IT in Real Life

In this section, we conduct MEA against real-world commercial I2IT services to verify our attack's ability. We select two widely used I2IT service platforms as our target. The outcomes of our attack substantiate its impressive efficacy in successfully extracting victim models.

I2IT Services

The mainstream I2IT services can be divided into two categories: style transfer, which facilitate the conversion of input image styles, and image enhancement services, which focus on restoring and improving the quality of degraded images.

Typically, users can access the I2IT services through two primary ways. The first involves utilizing an API in the cloud, following a pay-as-you-go model. The other is to access the model features locally through buyout purchases, allowing for an unlimited number of interactions. Nonetheless, the local model is typically encapsulated or encrypted within a black box, which constrains users from accessing its internal components, such as the model's architecture and parameters. In both scenarios, MEAs can be performed by sending queries to the black-box target model.

Attack Process

Attack Settings We select two popular third-party I2IT service providers in the market: Imglarger[2] and Baidu AI Cloud[1]. We conduct MEAs against Imglarger's human face cartoonization functions, including face2cartoon (2d) and face2cartoon (3d), as well as Baidu AI Cloud's picture style cartoonization function (i.e., landscape2cartoon).

We randomly select 2,000 images from the face dataset FFHQ (512×512) (Marinez 2020) and Landscape dataset (Rougetet 2020) as the attack dataset, respectively. The face dataset FFHQ is composed of 52,000 high-quality PNG images with 512×512 resolution crawled from Flickr. These images vary considerably in age, race, and image background, and are automatically aligned and cropped using dlib. We use the images in this dataset as the attack dataset for both face2cartoon (2d) and face2cartoon (3d) tasks and choose CycleGAN as the attack model backbone. The Landscape dataset is the same we used in the photo2vangog task, and Pix2Pix is used as the attack model backbone for landscape2cartoon task. To evaluate the extraction performance, we take 200 and 150 images from the remaining parts of the FFHQ and landscape datasets, respectively, as the test set.

Results Comparison

We now analyze the results in both quantitative and qualitative ways. Since the D_V cannot be obtained from the commercial services, we cannot measure the distance from the attack model output to the target domain (*i.e.*, the $R_{\text{capability}}$). As a result, we only evaluate the attack model using the R_{fidelity} metric in this section. Fig. 7 and Table 5 illustrate the performance of our attack. The results show that our approach successfully replicates the function of the victim model, surpassing the performance of the baseline method. Additional results can be found in Supp.-E.

7 Conclusion

In this paper, we have presented a novel MEA attack to extract models in I2IT tasks. We identify that the impact of the domain shift problem is a fundamental factor affecting the performance of MEA. This is particularly notable for I2IT tasks where the optimal selection of queries is not apparent. Our approach addresses the issue from a new angle by resorting to a flatter and smoother loss landscape for the attack model. By incorporating wavelet regularization and sharpness-aware minimization, our attack exhibits significant performance improvement in all comparative experiments. We also conduct our attack against real-world commercial services. The outcomes of our attack substantiate its impressive efficacy in successfully extracting victim models.

Future work will include further investigation on dedicated defense mechanisms. We also see new research opportunities in extending the attack approach against diverse GAN-based models.

Ethics Statement

Given the nature of Sec. 6 which involves discussing real-world attacks, we do not publicly release our code of this section in the usual manner. Instead, we intend to make it accessible only to legitimate researchers as per request. This approach aims to facilitate the reproducibility of our findings while also maintaining a responsible handling of potentially sensitive information.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 12271464) and the Hunan Provincial Natural Science Foundation of China (No. 2023JJ10038).

Correspondence of this work should be directed to L. Zhang (leo.zhang@griffith.edu.au) and H. Yuan (yhz@xtu.edu.cn).

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 126–135.

Barbalau, A.; Cosma, A.; Ionescu, R. T.; and Popescu, M. 2020. Black-Box Ripper: Copying black-box models using generative evolutionary algorithms. *Advances in Neural Information Processing Systems*, 33: 20120–20129.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Chen, X. 2020. Hayao and Shinkai datasets. https://github.com/TachibanaYoshino/AnimeGANv2/releases. Online; accessed 1 May 2023.

d'Ascoli, S.; Sagun, L.; and Biroli, G. 2020. Triple descent and the two kinds of overfitting: Where & why do they appear? *Advances in Neural Information Processing Systems*, 33: 3058–3069.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2020. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning*, 513–520.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

Hu, H.; and Pang, J. 2021. Stealing machine learning models: Attacks and countermeasures for generative adversarial networks. In *ACSAC*, 1–16.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1125–1134.

Jagielski, M.; Carlini, N.; Berthelot, D.; Kurakin, A.; and Papernot, N. 2020. High accuracy and high fidelity extraction of neural networks. In *Proceedings of the 29th USENIX Conference on Security Symposium*, 1345–1362.

Kim, J.; Kim, M.; Kang, H.; and Lee, K. 2019. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*.

Krishna, K.; Tomar, G. S.; Parikh, A. P.; Papernot, N.; and Iyyer, M. 2019. Thieves on sesame street! model extraction of BERT-based APIs. *arXiv preprint arXiv:1910.12366*.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30.

Ma, M.; Zhang, Y.; Arachchige, P. C. M.; Zhang, L. Y.; Chhetri, M. B.; and Bai, G. 2023. LoDen: Making Every Client in Federated Learning a Defender Against the Poisoning Membership Inference Attacks. In *ACM AsiaCCS*, 122–135.

Marinez, H. 2020. FFHQ datasets. https://www.kaggle.com/datasets/arnaud58/flickrfaceshq-dataset-ffhq. Online; accessed 1 May 2023.

Orekondy, T.; Schiele, B.; and Fritz, M. 2019. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4954–4963.

Pal, S.; Gupta, Y.; Shukla, A.; Kanade, A.; Shevade, S.; and Ganapathy, V. 2020. ACTIVETHIEF: Model extraction using active learning and unannotated public data. In *AAAI 2020-34th AAAI Conference on Artificial Intelligence*, 865–872. AAAI press.

Pang, Y.; Lin, J.; Qin, T.; and Chen, Z. 2021. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, 24: 3859–3881.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 506–519.

Pouyanfar, S.; Sadiq, S.; Yan, Y.; Tian, H.; Tao, Y.; Reyes, M. P.; Shyu, M.-L.; Chen, S.-C.; and Iyengar, S. S. 2018. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5): 1–36.

Rougetet, A. 2020. Landscape datasets. https://www.kaggle.com/datasets/arnaud58/landscape-pictures. Accessed 1 May 2023.

Shen, Y.; He, X.; Han, Y.; and Zhang, Y. 2022. Model stealing attacks against inductive graph neural networks. In *SP*, 1175–1192. IEEE.

Szyller, S.; Duddu, V.; Gröndahl, T.; and Asokan, N. 2021. Good artists copy, great artists steal: Model extraction attacks against image translation generative adversarial networks. *arXiv preprint arXiv:2104.12623*.

Timofte, R.; and et al. 2017. *NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results*, volume 2, 6. IEEE.

Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M. K.; and Ristenpart, T. 2016. Stealing machine learning models via prediction APIs. In *USENIX Security 16*, 601–618.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018a. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 8798–8807.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1905–1914.

Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018b. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 606–615.

Wu, B.; Yang, X.; Pan, S.; and Yuan, X. 2022. Model Extraction Attacks on Graph Neural Networks: Taxonomy and Realisation. In *Asia CCS*, 337–350.

Xu, Q.; He, X.; Lyu, L.; Qu, L.; and Haffari, G. 2021. Student Surpasses Teacher: Imitation attack for black-box NLP APIs. *arXiv preprint arXiv:2108.13873*.

Ye, J.; Maddi, A.; Murakonda, S. K.; Bindschaedler, V.; and Shokri, R. 2022. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 3093–3106.

Yuan, X.; Ding, L.; Zhang, L.; Li, X.; and Wu, D. O. 2022. ES attack: Model stealing against deep neural networks without hurdles. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Zhang, L.; Chen, X.; Tu, X.; Wan, P.; Xu, N.; and Ma, K. 2022a. Wavelet knowledge distillation: Towards efficient image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12464–12474.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, Z.; Zhang, L. Y.; Zheng, X.; Hussain Abbasi, B.; and Hu, S. 2022b. Evaluating Membership Inference Through Adversarial Robustness. *The Computer Journal*, 65(11): 2969–2978.

Zhu, J.-Y.; and et al. 2021. CycleGAN datasets. http://efrosgans.eecs.berkeley.edu/cyclegan/datasets/. Accessed 1 May 2023.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. 2019. Horse datasets. https://www.kaggle.com/datasets/alessiocorrado99/animals10. Accessed 1 May 2023.