Imperceptible Adversarial Attack via Invertible Neural Networks

Zihan Chen*, Ziyue Wang*, Jun-Jie Huang†, Wentao Zhao, Xiao Liu, Dejian Guan

College of Computer Science, National University of Defense Technology, Changsha, Hunan, China
{chenzihan21, wangzy13, jjhuang, wtzhao, liuxiao13a, guandejian20}@nudt.edu.cn

Abstract

Adding perturbations via utilizing auxiliary gradient information or discarding existing details of the benign images are two common approaches for generating adversarial examples. Though visual imperceptibility is the desired property of adversarial examples, conventional adversarial attacks still generate traceable adversarial perturbations. In this paper, we introduce a novel Adversarial Attack via Invertible Neural Networks (AdvINN) method to produce robust and imperceptible adversarial examples. Specifically, AdvINN fully takes advantage of the information preservation property of Invertible Neural Networks and thereby generates adversarial examples by simultaneously adding class-specific semantic information of the target class and dropping discriminant information of the original class. Extensive experiments on CIFAR-10, CIFAR-100, and ImageNet-1K demonstrate that the proposed AdvINN method can produce less imperceptible adversarial images than the state-of-the-art methods and AdvINN yields more robust adversarial examples with high confidence compared to other adversarial attacks. Code is available at https://github.com/jjhuangcs/AdvINN.

Introduction

Deep Neural Networks (DNNs) have achieved outstanding performance in a wide range of applications, however, have shown to be vulnerable to adversarial examples (Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2014; Akhtar and Mian 2018; Hendrycks et al. 2021). By adding mild adversarial noise to a benign image, classification DNNs can be easily deceived and misclassify this adversarial example to an erroneous class label. Though the existence of adversarial examples may hinder the applications of DNNs to risk sensitive domains, it further promotes investigation on robustness of DNNs.

Adversarial examples can be generated by either adding or dropping certain information with respect to the input benign images. Adding adversarial perturbations (Szegedy et al. 2014; Moosavi-Dezfooli, Fawzi, and Frossard 2016; Carlini and Wagner 2017) to clean images is the most common approach to craft adversarial examples. Fast Gradient Sign Method (FGSM) (Szegedy et al. 2014) and its variations (Kurakin, Goodfellow, and Bengio 2016; Dong et al. 2018; Lin et al. 2019) add adversarial noise to the benign image according to the sign of the gradients of the loss function with respect to the input image. An alternative is to mix a sequence of images to make the classifier output erroneous predictions and improve the transferability of generated adversarial examples (Wang et al. 2021), since information from images of other classes could disturb the prediction of DNNs. Recently, dropping existing information from the original images has also shown to be an effective way to generate adversaries (Duan et al. 2021). Compared to methods of adding adversarial perturbations to the benign images, AdvDrop (Duan et al. 2021) shows stronger robustness against denoising-based defence methods and will not lead to suspicious increase of image storage size.

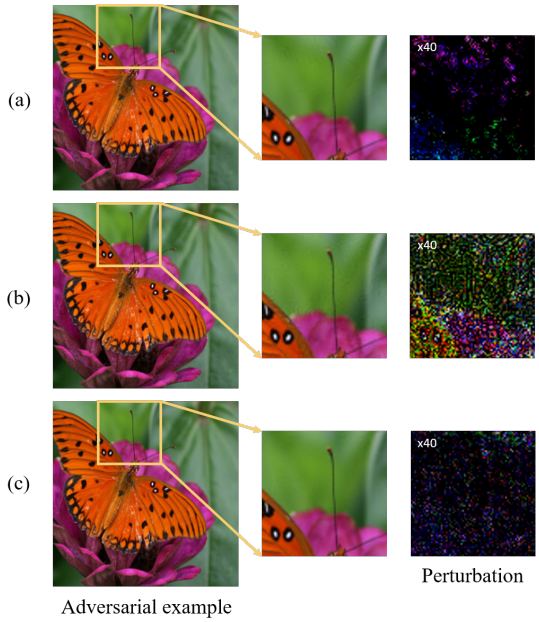


Figure 1: Adversarial examples generated by (a) PerCAL (Zhao, Liu, and Larson 2020), (b) SSAH (Luo et al. 2022), and (c) the proposed AdvINN, respectively. All perturbations are enlarged for better visualization.

*These authors contributed equally.
†Corresponding author

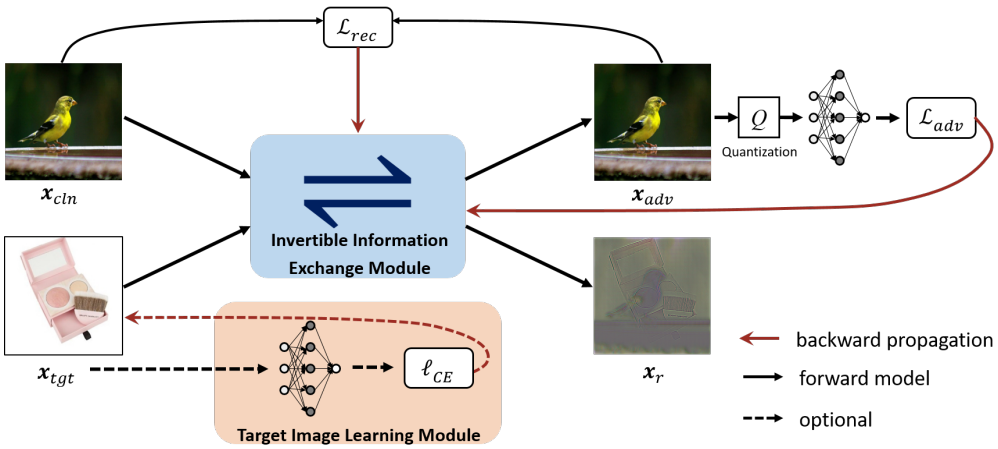


Figure 2: The overview architecture of our proposed Adversarial Attack using Invertible Neural Networks (AdvINN) method. The Invertible Information Exchange Module, which is with the information preservation property, non-linearly exchanges information between the input benign image and the target image. The Target Image Learning Module is used to update the learnable target image x_{tgt}. The quantization module is set to round the pixel values of the generated adversarial examples x_{adv} to be integers and within the range of [0,255].

Adversarial examples crafted by adding or dropping information are both able to deceive DNNs with incorrect prediction of image contents, however, both approaches have their limitations. The methods based on adding adversarial perturbations may lead to perceptible noise patterns and noticeable increase of image storage size, while the method of dropping existing information has limited performance on targeted attacks. Therefore, it is of great potential to make an attempt to combine the best features from two perspectives by simultaneously adding semantic information from the target image and dropping semantic information of the original class to craft adversarial examples.

In this paper, we propose a novel Adversarial attack method using Invertible Neural Networks, termed AdvINN, by leveraging the information preservation property of Invertible Neural Networks (INNs) to achieve simultaneously adding extra information and dropping existing details. Specifically, given a clean image, a target image is selected or learned as the source of information for adding adversarial perturbations. The clean image and the target image are inputs to an Invertible Information Exchange Module (IIEM) for alternating update. The amount of information within the input and output of IIEM keeps the same due to its information preservation property. Therefore, driven by an adversarial loss and a reconstruction loss, the generated adversarial image will gradually transfer discriminant features of the clean image to the residual image and at the same time add class-specific semantic features from the target image to form an adversarial example.

The contribution of this paper is three-fold:

- We propose a novel Adversarial attack method using Invertible Neural Networks (AdvINN) which exploits the information preservation property of Invertible Neural Networks and is able to achieve simultaneously adding class-specific information from a target image and dropping semantic information of the original class.
- We propose three approaches to choose the target image, including highest confidence image, universal adversarial perturbation, and learnable classifier guided target image. With the proposed AdvINN, class-specific features can be effectively transferred to the input image leading to highly interpretable and imperceptible results.
- With comprehensive experiments and analysis, we have demonstrated the effectiveness and robustness of the proposed AdvINN method, and shown that the adversarial examples generated by AdvINN are more imperceptible and with high attacking success rates.

Related Works

Adversarial Attack

Adversarial attacks (Szegedy et al. 2014) aim to deceive DNNs with adversarial examples whose difference to the input benign image is bounded by l_∞-norm. That is, an adversarial example should be able to fool DNNs and at the same time be as imperceptible as possible. In general, adversarial examples can be crafted by adding disturbing adversarial perturbations to clean images or dropping crucial information from the clean images.

Adding adversarial perturbations to clean images is a predominant way to generate adversarial examples in recent works. FSGM (Szegedy et al. 2014) proposes to add adversarial perturbation in the direction of sign of gradient. BIM (Kurakin, Goodfellow, and Bengio 2016) increases the number of iterations and updates with smaller steps to improve the attacking success rate. StepLL (Kurakin, Goodfellow, and Bengio 2016) proposes to choose the least-likely class as the target class and can generate

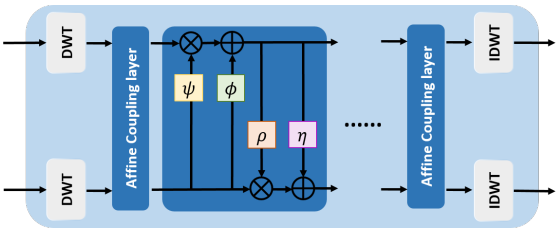


Figure 3: The network architecture of the Invertible Information Exchange Module.

adversarial examples which are highly misclassified by Inceptionv3 (Szegedy et al. 2016). PGD (Madry et al. 2018) is similar to BIM, while with the randomly initialized starting point in the neighborhood of ground-truth image. DeepFool (Moosavi-Dezfooli, Fawzi, and Frossard 2016) proposes to add the minimal norm adversarial perturbation around decision boundary to make false predictions. C&W (Carlini and Wagner 2017) attempts to find a balance between imperceptible perturbations and adversarial attacks with l_0, l_2 and l_∞-norm regularizations. (Wang et al. 2021) propose Admix method to generate more transferable adversarial examples by mixing the input image with a small portion of images sampled from other categories.

Dropping existing information has also proven to be able to successfully craft adversarial examples, which provides a new perspective for generating adversarial examples. (Duan et al. 2021) propose AdvDrop method which learns the quantization table in the JPEG compression framework leading to dropping information in frequency domain. Compared with traditional adversarial attack methods (*e.g.* PGD (Madry et al. 2018)), adversarial examples generated by AdvDrop have fewer details, is with decreased image size and possess a higher robustness with respect to denoising-based adversarial defense methods. However, the confidence of prediction could hardly be improved due to limited and reduced information in the generated adversarial examples. Moreover, there are visible quantization artifacts since the input image is splitted into blocks before transformation.

Imperceptibility of adversarial examples is an important criterion for adversarial attacks, however, it is not well attained by many well-known adversarial attack methods and there usually contains noticeable adversarial perturbations to human-beings. For wider applications, imperceptible adversaries can be applied in privacy protection *e.g.* face recognition. Recently, there are an increasing number of works aiming to improve the imperceptibility of adversarial perturbations (Croce and Hein 2019; Jia et al. 2022; Tian et al. 2022). Zhao *et al.* (Zhao, Liu, and Larson 2020) introduce PerC-AL in which adversarial perturbations are optimized in terms of perceptual color distance leading to improve visual imperceptibility. Luo *et al.* (Luo et al. 2022) propose a semantic similarity attack and introduce a new constraint on low-frequency sub-bands between benign images and adversaries, which encourages to add distortions on the high-frequency sub-bands.

Invertible Neural Networks

Invertible Neural Networks (INNs) (Dinh, Krueger, and Bengio 2014; Dinh, Sohl-Dickstein, and Bengio 2016; Kingma and Dhariwal 2018; Jacobsen, Smeulders, and Oyallon 2018) are bijective function approximators due to their mathematically induced network architecture. Given an intermediate feature, INNs are able to explicitly perfect reconstruct features of other layers. That is, the information of INNs' input is preserved throughout all its layers and there is no extra information injected or lost.

INNs are able to explicitly construct inverse mapping, therefore are suitable candidates to perform mappings between two domains. INNs have been applied in many computer vision tasks, including image rescaling (Xiao et al. 2022; Zhang et al. 2022), image colorization (Ardizzone et al. 2019; Zhao et al. 2021), video super-resolution (Zhu et al. 2019; Huang et al. 2021), image denoising (Liu et al. 2021; Huang and Dragotti 2022, 2021), image separation (Huang et al. 2022), image steganography (Lu et al. 2021; Jing et al. 2021; Guan et al. 2022), and invertible image conversion (Cheng, Xie, and Chen 2021), etc. The most relevant to our work is AdvFlow (Mohaghegh Dolatabadi, Erfani, and Leckie 2020) which utilizes the normalizing flows to model the density of adversarial examples for black-box adversarial attack. However, to the best of our knowledge, there is no work using the information preservation property of Invertible Neural Networks for generating adversarial examples.

Proposed Method

Adding bounded class-specific adversarial information to a benign image and dropping existing discriminant information of the original class are two distinctive perspectives for generating adversarial examples and with their own strengths. In this paper, we aim to combine the best features of two paradigms. That is, crafting imperceptible and robust adversarial examples by simultaneously adding and dropping semantic information in an unified framework.

Overview

Given a benign image $\boldsymbol{x}_{cln}$ with class label c, our objective is to generate a corresponding adversarial image $\boldsymbol{x}_{adv}$ by dropping discriminate information of class c while adding adversarial details from a target image $\boldsymbol{x}_{tgt}$, and at the same time to be able to interpret the features that have been added or dropped through a residual image $\boldsymbol{x}_r$.

Fig. 2 shows the overview of the proposed Adversarial attack using Invertible Neural Networks (AdvINN) method. The proposed AdvINN $f_{\boldsymbol{\theta}}(\cdot,\cdot)$ is parameterized by $\boldsymbol{\theta}$ and with $(\boldsymbol{x}_{adv}, \boldsymbol{x}_r) = f_{\boldsymbol{\theta}}(\boldsymbol{x}_{cln}, \boldsymbol{x}_{tgt})$, where $\boldsymbol{\theta}$ represents the parameters of AdvINN. It consists of an Invertible Information Exchange Module (IIEM), a Target Image Learning Module (TILM) and loss functions for optimization. As the source of adversarial information, a target image $\boldsymbol{x}_{tgt}$ can be chosen as the highest confidence target image (HCT), an universal adversarial perturbation (UAP), or an online learned classifier guided target image (CGT) using TILM.

With $(\boldsymbol{x}_{cln}, \boldsymbol{x}_{tgt})$, IIEM driven by the loss functions generates the adversarial image $\boldsymbol{x}_{adv}$ by performing information exchange between the two images. Owing to the information preservation property of IIEM, the amount of information within input images ($\boldsymbol{x}_{cln}$, $\boldsymbol{x}_{tgt}$) and output images ($\boldsymbol{x}_{adv}$, $\boldsymbol{x}_r$) is the same and there explicitly exists an inverse mapping with $(\boldsymbol{x}_{cln}, \boldsymbol{x}_{tgt}) = f_{\boldsymbol{\theta}}^{-1}(\boldsymbol{x}_{adv}, \boldsymbol{x}_r)$.

The learning objective of the proposed AdvINN method can be expressed as:

$$\begin{aligned} \boldsymbol{x}_{adv} = &\arg\min_{\boldsymbol{\theta}} \lambda_{adv}\mathcal{L}_{adv}\left(\boldsymbol{x}_{adv}, c\right) + \mathcal{L}_{rec}(\boldsymbol{x}_{adv}, \boldsymbol{x}_{cln}), \\ &\text{s.t.} \left\|\boldsymbol{x}_{adv} - \boldsymbol{x}_{cln}\right\|_\infty \leq \epsilon, \end{aligned} \tag{1}$$

where $\boldsymbol{\theta}$ denotes the parameters of AdvINN, $\mathcal{L}_{adv}(\cdot,\cdot)$ denotes the adversarial loss, $\mathcal{L}_{rec}(\cdot,\cdot)$ denotes the reconstruction loss, λ_{adv} is the regularization parameter and ϵ denotes the budget of adversarial perturbation.

In the following, we will introduce the details of Invertible Information Exchange Module, target image selection and learning, and the loss functions for optimizations.

Invertible Information Exchange Module

To achieve simultaneously adding and dropping semantic information for adversarial example generation, an Invertible Information Exchange Module (IIEM) is proposed as a nonlinear transform with information preservation property to interchange information between the clean image and the target image.

Discrete Wavelet Transform. In order to disentangle the input clean and target images into low-frequency and high-frequency components, Discrete Wavelet Transform (DWT) (Mallat 1989) has been applied to the inputs for decomposition. The separation of low- and high-frequency features will facilitate modifications to the input image applied on the high-frequency components and therefore results in less perceptible adversarial examples.

With DWT $\mathcal{T}(\cdot)$, an input image $\boldsymbol{x} \in \mathbb{R}^{C\times H\times W}$ will transformed into wavelet domain $\mathcal{T}(\boldsymbol{x}) \in \mathbb{R}^{4C\times H/2\times W/2}$. It contains one low-frequency sub-band feature and three high-frequency sub-band features. At the output end of IIEM, Inverse Discrete Wavelet Transform (IDWT) $\mathcal{T}^{-1}(\cdot)$ has been applied to reconstruct the features back to image domain.

Affine Coupling Blocks. Invertible Information Exchange Module is composed of M Affine Coupling Blocks. Let us denote with $\boldsymbol{w}_{cln}^i$ and $\boldsymbol{w}_{tgt}^i$ the input features of the i-th Affine Coupling Block, and with $\boldsymbol{w}_{cln}^0 = \mathcal{T}(\boldsymbol{x}_{cln})$ and $\boldsymbol{w}_{tgt}^0 = \mathcal{T}(\boldsymbol{x}_{tgt})$. Then, the forward process of the i-th Affine Coupling Block can be expressed as:

$$\begin{aligned} \boldsymbol{w}_{cln}^i =& \boldsymbol{w}_{cln}^{i-1} \odot \exp\left(\alpha\left(\psi\left(\boldsymbol{w}_{tgt}^{i-1}\right)\right)\right) + \phi\left(\boldsymbol{w}_{tgt}^{i-1}\right), \\ \boldsymbol{w}_{tgt}^i =& \boldsymbol{w}_{tgt}^{i-1} \odot \exp\left(\alpha\left(\rho\left(\boldsymbol{w}_{cln}^{i}\right)\right)\right) + \eta\left(\boldsymbol{w}_{cln}^{i}\right), \end{aligned} \tag{2}$$

where $\odot$ denotes element-wise multiplication, $\boldsymbol{\alpha}$ is a Sigmoid function multiplied by a constant factor, and $\psi(\cdot), \phi(\cdot), \rho(\cdot), \eta(\cdot)$ denote dense network architectures as in (Wang et al. 2018).

Given the output of M-th Affine Coupling Block, the adversarial image and the residual image can be reconstructed using IDWT with $\boldsymbol{x}_{adv} = \mathcal{T}^{-1}(\boldsymbol{w}_{cln}^M)$ and $\boldsymbol{x}_r = \mathcal{T}^{-1}(\boldsymbol{w}_{tgt}^M)$.

By default, for DWT/IDWT, Haar wavelet transform is used, and the number of Affine Coupling Blocks is set to 2.

Information Preservation Property. Due to the invertibility of DWT and IDWT, $\boldsymbol{w}_{cln}^M$ and $\boldsymbol{w}_{tgt}^M$ can be restored from $(\boldsymbol{x}_{adv}, \boldsymbol{x}_r)$. In IIEM, only the forward process of the Affine Coupling Blocks is used for generating adversarial images, and it's worth to note that $(\boldsymbol{w}_{cln}^{i-1}, \boldsymbol{w}_{tgt}^{i-1})$ can be perfectly recovered from $(\boldsymbol{w}_{cln}^{i}, \boldsymbol{w}_{tgt}^{i})$:

$$\begin{aligned} \boldsymbol{w}_{tgt}^{i-1} &= \left(\boldsymbol{w}_{tgt}^{i} - \eta\left(\boldsymbol{w}_{cln}^{i}\right)\right) \odot \exp\left(-\alpha(\rho(\boldsymbol{w}_{cln}^{i}))\right), \\ \boldsymbol{w}_{cln}^{i-1} &= \left(\boldsymbol{w}_{cln}^{i} - \phi\left(\boldsymbol{w}_{tgt}^{i-1}\right)\right) \odot \exp\left(-\alpha(\psi(\boldsymbol{w}_{tgt}^{i-1}))\right). \end{aligned} \tag{3}$$

Therefore, IIEM is fully invertible and the output images $(\boldsymbol{x}_{adv}, \boldsymbol{x}_r)$ contain the same amount of information as the input images $(\boldsymbol{x}_{cln}, \boldsymbol{x}_{tgt})$. Their relationship can be represented as:

$$\begin{cases} \boldsymbol{x}_{adv} &= \boldsymbol{x}_{cln} - \boldsymbol{\sigma} + \boldsymbol{\delta}, \\ \boldsymbol{x}_r &= \boldsymbol{x}_{tgt} + \boldsymbol{\sigma} - \boldsymbol{\delta}. \end{cases} \tag{4}$$

where $\boldsymbol{\sigma}$ denotes the dropped existing information of the clean image, and $\boldsymbol{\delta}$ denotes the added discriminant information from the target image to the clean image.

In the case of $\boldsymbol{x}_{tgt}$ being a constant image, there will be no information added from the target image to the clean image, *i.e.*, $\boldsymbol{\delta} = 0$. The residual image $\boldsymbol{x}_r$ will then only correspond to the dropped information $\boldsymbol{\sigma}$ and can be used to interpret the results of AdvINN.

Target Image Selection and Learning

The target image in the proposed AdvINN method plays an essential role and determines the information to be added to the clean image for generating the adversarial image. In this section, we introduce three options for selecting or learning the target image.

Highest Confidence Target Image (HCT). The most intuitive idea is to select the image with the highest confidence in each class as the target image as StepLL (Kurakin, Goodfellow, and Bengio 2016), since the higher confidence of the target image to the classifier is, the more discriminant information the images may possess. However, selecting a natural image as the target image may not be the best option, since a natural image often carries a considerable amount of information unrelated to the target class, such as background texture and the details of other classes. This could hinder the optimization speed as well as the success rates of attacks.

UAP as Target Image (UAP). Universal Adversarial Perturbations (Moosavi-Dezfooli et al. 2017; Poursaeed et al. 2018; Khrulkov and Oseledets 2018; Zhang et al. 2020; Benz et al. 2020) aggregate dominant information of images of the target class and minimize the interference of irrelevant details. Therefore, it could be better option for the target image. Zhao *et al.*(Zhao, Liu, and Larson 2021) propose that targeted adversarial perturbations are optimized in

Dataset	Methods	$l_2 \downarrow$	$l_\infty \downarrow$	SSIM↑	LPIPS↓	FID↓	ASR(%)↑
ImageNet-1K	StepLL	26.90	0.04	0.948	0.1443	25.176	98.5
	C&W	10.33	0.07	0.977	0.0617	11.515	91.7
	PGD	64.42	0.04	0.881	0.2155	35.012	90.2
	PerC-AL	**1.93**	0.10	0.995	0.0339	5.118	**100.0**
	AdvDrop	18.47	0.07	0.977	0.0639	9.687	**100.0**
	SSAH	6.97	**0.03**	0.991	0.0352	5.221	99.8
	AdvINN-HCT	5.73	**0.03**	0.991	0.0206	3.661	**100.0**
	AdvINN-UAP	5.84	**0.03**	0.990	0.0212	2.900	**100.0**
	AdvINN-CGT	2.66	**0.03**	**0.996**	**0.0118**	**1.594**	**100.0**
CIFAR-100	StepLL	0.73	0.04	0.923	0.0411	11.608	94.3
	C&W	1.24	0.09	0.943	0.0706	12.507	97.7
	PGD	1.59	**0.03**	0.954	0.0793	23.899	99.2
	PerC-AL	3.09	0.27	0.961	0.0426	6.035	97.2
	AdvDrop	87.09	0.61	0.774	0.2549	14.722	90.7
	SSAH	0.43	0.04	0.992	0.0200	4.508	99.4
	AdvINN-HCT	0.28	**0.03**	0.991	**0.0035**	**3.413**	98.3
	AdvINN-UAP	0.27	**0.03**	**0.993**	0.0037	3.982	**99.6**
	AdvINN-CGT	**0.23**	**0.03**	**0.993**	0.0037	3.921	99.5
CIFAR-10	StepLL	0.77	0.04	0.982	0.0462	10.997	98.2
	C&W	1.06	0.09	0.970	0.0667	10.510	99.3
	PGD	1.61	**0.03**	0.956	0.0861	24.014	**100.0**
	PerC-AL	0.52	0.13	0.990	0.0134	**1.518**	**100.0**
	AdvDrop	70.10	0.46	0.570	0.4483	122.950	97.7
	SSAH	0.38	**0.03**	0.993	0.0180	3.654	99.9
	AvdINN-HCT	0.18	**0.03**	**0.995**	0.0033	2.627	99.9
	AdvINN-UAP	0.19	**0.03**	**0.995**	0.0031	2.791	99.9
	AdvINN-CGT	**0.17**	**0.03**	**0.995**	**0.0030**	2.480	99.9

Table 1: Accuracy and evaluation metrics on different methods. All methods use $\epsilon = 8/255$ as the adversarial budget. ASR donates the accuracy of adversarial attacks. ↑ means the value is higher the better, and vice versa. (The best and the second best result in each column is in bold and underline.)

a data-free manner. We follow this method and utilize the optimized universal adversarial perturbation as target images, which is able to moderately accelerate convergence speed.

Classifier Guided Target Image (CGT). In order to further improve the optimization speed as well as attacking success rate, we propose a Target Image Learning Module (TILM) which learns a classifier guided target image rather than using a fixed image as the target image. Inspired by targeted UAPs (Zhao, Liu, and Larson 2021), the target image is set to be a learnable variable which is initialized with a constant image (*i.e.,* all pixels are set to 0.5) and then updated according to the gradient from the attacking classifier. In this way, an adaptively generated target image can embed more discriminant information of the target class assisting the generation of adversarial examples. Detailed experimental results on the target image selection will be discussed and presented in Experiments section.

Learning Details

We set up a reconstruction loss $\mathcal{L}_{rec}$ and an adversarial loss $\mathcal{L}_{adv}$ to locate the correct optimized direction and accelerate convergence speed. The total loss can be expressed as:

$$\mathcal{L}_{total} = \lambda_{adv}\mathcal{L}_{adv} + \mathcal{L}_{rec}, \tag{5}$$

where λ_{adv} is the regularization parameter.

Reconstruction loss is utilized to constrain the optimized adversarial image being close to the input clean image, while imposing the modifications to be applied mainly on the high-frequency and less perceptible components leading to less visible adversarial examples:

$$\begin{aligned}\mathcal{L}_{rec} = &\sum_{i\in\{ll,lh,hl,hh\}} w_i \|\mathcal{T}(\boldsymbol{x}_{cln})_i, \mathcal{T}(\boldsymbol{x}_{adv})_i\|_2^2 \\ &+ \lambda_{perp}\|\rho(\boldsymbol{x}_{cln}), \rho(\boldsymbol{x}_{adv})\|_2^2,\end{aligned} \tag{6}$$

where ll, lh, hl, hh denote the low- and high-frequency components of the wavelet transform, w_i is the weight of the corresponding wavelet component, λ_{perp} is the weight of perceptual loss and $\rho(\cdot)$ denotes the features of the VGG-16 model pretrained on ImageNet dataset.

Adversarial loss evaluates dissimilarity of prediction logits and target label. The cross entropy loss $\ell_{CE}(\cdot)$ is used to measure the difference.

$$\mathcal{L}_{adv} = \ell_{CE}\left(g_\phi(\boldsymbol{x}_{adv}), c_{tgt}\right), \tag{7}$$

where $g_\phi(\cdot)$ denotes the target classifier, and c_{tgt} is the label of the target class.

We set a classifier guided loss $\mathcal{L}_{cgt}$ for learning $\boldsymbol{x}_{cgt}$. It is also a cross entropy loss similar to (7).

The optimizer for optimizing the learning objective of AdvINN in (1) is set to Adam (Kingma and Ba 2014) optimizer with initial learning rate $1e^{-4}$ which is decayed every 200 iterations with decay rate 0.9 and is lower bounded by $1e^{-5}$. We empirically set the regularization parameters $\lambda_{adv} = 3$, $w_{ll} = 2$, $w_{lh,hl,hh} = 1$ and $\lambda_{perp} = 0.001$.

Experiments

Experimental Setup

Dataset and models. We evaluate the performance of the comparison methods on ImageNet-1K dataset which contains 1000 images sampled from the ImageNet-1K validation set (Russakovsky et al. 2015). The benign images are all correctly classified by the target classifier. All experiments were performed on a computer with a NVIDIA RTX 3090 GPU with 24 GB memory.

Comparison methods. Six comparison methods have been included for evaluation, with three well-known adversarial attack methods as our baselines including PGD (Madry et al. 2018) under l_∞-norm, StepLL (Kurakin, Goodfellow, and Bengio 2016), and C&W (Carlini and Wagner 2017), and three recent state-of-the-art methods, including AdvDrop (Duan et al. 2021), PerC-AL (Zhao, Liu, and Larson 2020) and SSAH (Luo et al. 2022).

Evaluation metrics. We use attacking success rate (ASR) to evaluate the attacking performance, and five popular metrics to evaluate the quality of the generated adversarial images, including: l_2-norm, l_∞-norm, Structural Similarity Index (SSIM) (Wang et al. 2004), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018) and Fréchet Inception Distance (FID) (Heusel et al. 2017). Specifically, l_2-norm measures the average energy of the adversarial perturbations, l_∞-norm evaluates the maximum perturbation intensity, SSIM assesses the structural similarity between two images, and LPIPS and FID both measure the perceptual similarity.

Attack setting. For fair comparisons, all comparison methods perform targeted attacks with the least-likely objective (except SSAH) to avoid choosing closely related classes which is less meaningful in real applications. For the target classifier, we use pre-trained ResNet50[1] which is with 23.85% top-1 error on ImageNet-1K.

Evaluation on Targeted Attacks

Table 1 shows the white-box targeted attack performance of different methods on ImageNet-1K as well as the quality of the adversarial images evaluated using l_2-norm, l_∞-norm, SSIM, LPIPS, and FID. We can see that under the same perturbation budget $\epsilon = 8/255$ with respect to l_∞-norm, the proposed AdvINN method with HCT, UAP, and CGT achieve better image quality, especially on perceptual metrics, than the state-of-the-art methods. Specifically, AdvINN-CGT achieves the best results in terms of SSIM, LPIPS and FID and is with 100% ASR. In terms of FID, AdvINN-CGT achieves 8.093, 3.627, 3.524 lower FID score compared to AdvDrop (Duan et al. 2021), SSAH (Luo et al. 2022), PerC-AL (Zhao, Liu, and Larson 2020), respectively. PerC-AL achieves lower l_2-norm but the largest l_∞-norm, because it modifies a smaller number of pixels but with significantly larger values among all comparison methods. Therefore, PerC-AL has unsatisfactory perceptual scores. AdvINN-CGT achieves the 2^{nd} lowest l_2-norm, and the best scores in all other metrics. This indicates that the adversarial examples generated by AdvINN have higher structural and perceptual similarity to the ground-truth images than those of comparison methods.

Fig. 4 shows the exemplar adversarial examples crafted by different methods on ImageNet-1K. We can see that the proposed AdvINN method is able to generate visually less perceptible adversarial examples than comparison methods. The possible reason is that AdvINN performs information exchange on feature domain via Invertible Neural Networks. While traditional methods that generate adversarial samples based on the image domain usually restrict the number of perturbations by l_2-norm and l_∞-norm leading to uniformly distributed adversarial perturbations over the whole image.

We have also evaluated the performance of all comparison methods on the testing set of CIFAR-10 and CIFAR-100. The benign images are all correctly classified by the target classifier. For the target classifier, we use pre-trained ResNet-20[2] with 7.4% and 30.4% top-1 error on CIFAR-10 and CIFAR-100, respectively.

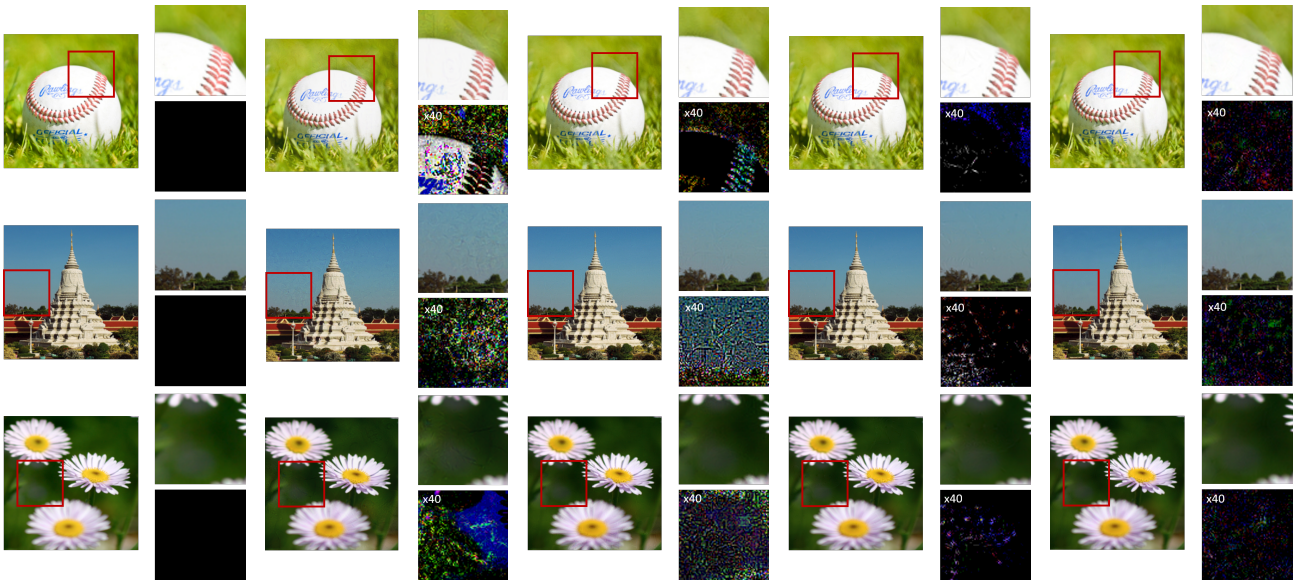
The parameter setting of AdvINN for CIFAR-10 and CIFAR-100 is slightly different from those used in ImageNet-1K due to the large difference on testing image size. The optimal setting for AdvINN is to use 1 Affine Coupling Block and 2 DWT layers. And the hyper-parameter λ_{adv} is set to 0.25 for CIFAR-100 and 0.2 for CIFAR-10 in order to craft more imperceptible adversarial examples.

Table 1 shows the white-box targeted attack performance of different methods on CIFAR-100 and CIFAR-10 as well as the quality of the adversarial images evaluated using l_2-norm, l_∞-norm, SSIM, LPIPS, and FID. We can observe that the proposed AdvINN method achieves less perceptible adversarial examples and guarantees high attacking success rate. Fig. 5 and Fig. 6 show the adversarial examples generated on CIFAR-100 and CIFAR-10, which can also support that our method generates more imperceptible examples in human visual system than other methods.

Table 1 also depicts the performance of AdvINN with different choices of target images. With highest confidence target images, AdvINN-HCT produces satisfactory image quality, however natural images usually contain irrelevant details to the target class that may slow down the optimization speed. By utilizing UAPs which effectively excludes unrelated information, AdvINN-UAP achieves around 20% speed-up in optimization speed and is with similar performance to AdvINN-HCT. However, the highest confidence images and targeted UAPs both need to prepare before-

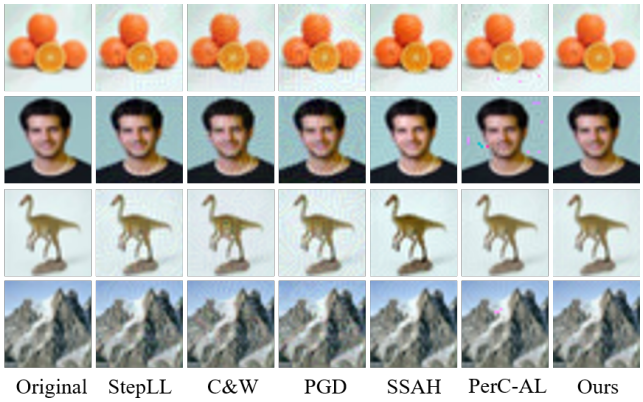
[1] https://download.pytorch.org/models

[2] https://github.com/chenyaofo/pytorch-cifar-models



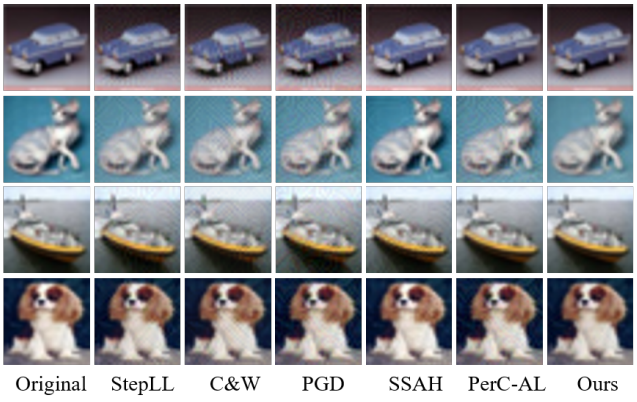
(a) Clean (b) C&W (c) SSAH (d) PerC-AL (e) AdvINN-CGT

Figure 4: Exemplar adversarial examples and the corresponding adversarial perturbations generated by C&W, SSAH, PerC-AL, and the proposed AdvINN method on ImageNet-1K.



Original StepLL C&W PGD SSAH PerC-AL Ours

Figure 5: More adversarial examples crafted by different methods on CIFAR-100.



Original StepLL C&W PGD SSAH PerC-AL Ours

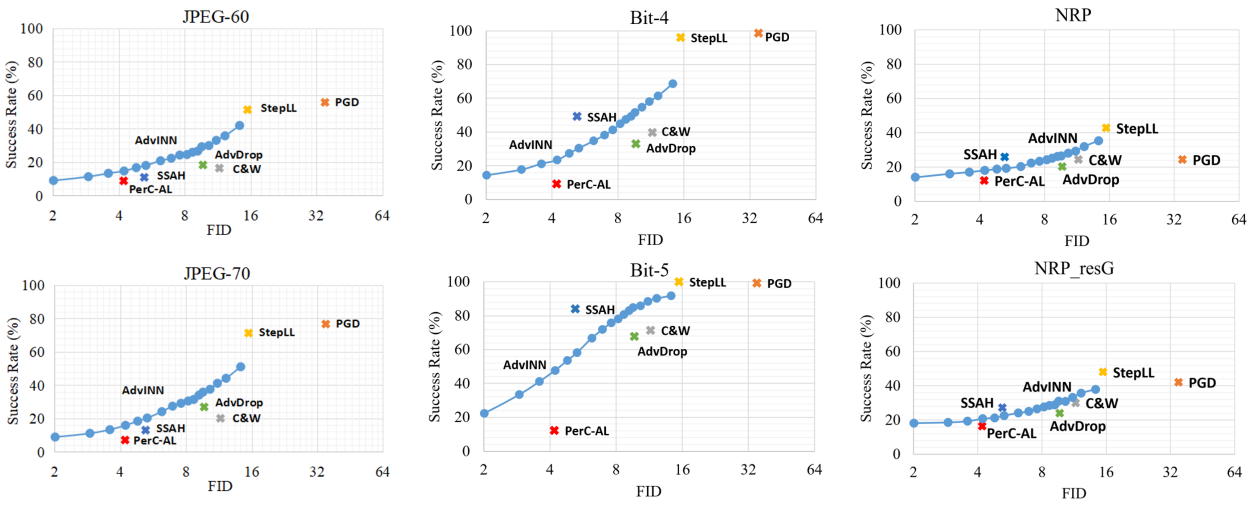
Figure 6: More adversarial examples crafted by different methods on CIFAR-10.

hand and may not contain suitable features to be transferred. AdvINN-CGT instead uses a learnable target image which is generated through the guidance of the classifier and learns essential details to the objective. We can see that AdvINN-CGT achieves the best image quality, and moreover, it only takes around 10% iterations to converge compared to AdvINN-UAP. Unless otherwise specified, we refer AdvINN as AdvINN-CGT.

Robustness

We follow robustness evaluation settings in AdvDrop (Duan et al. 2021) and PerC-AL (Zhao, Liu, and Larson 2020) and choose two common defense methods based on image transformation, *i.e.*, JPEG compression (Das et al. 2018) and bit-depth reduction (Guo et al. 2017). Except that, we have added purification-based methods named NRP and NRP_resG (Naseer et al. 2020) to investigate the robustness of the adversarial examples. Fig. 7 illustrates the imperceptibility and robustness of the adversarial examples generated by different methods. Specifically, the horizontal and vertical axis represents the FID score and the attacking success rate, respectively.

By varying the regularization parameter λ_{adv} in (5) within the range of $[3, 400]$, the proposed AdvINN method can achieve a trade-off between model robustness and imperceptibility and at the same time ensure 100% attacking success rate. We can see that the adversarial examples generated by PGD and StepLL are more robust against defense



(a) JPEG Compression (b) Bit-Depth Reduction (c) NRP

Figure 7: Evaluation on robustness of adversarial examples. All other methods use its recommended parameter settings.

Generator	SSIM↑	l_2 ↓	LPIPS↓	FID↓	Iter↓
w/ IIEM	0.996	2.66	0.0118	1.594	321
w/o IIEM	0.973	39.61	0.0325	6.141	272
w/ CNN	0.901	46.04	0.0360	4.345	1800

Table 2: Ablation study on the effectiveness of IIEM. Iter represents the average iterations on the whole dataset.

methods, however, the significantly higher FID scores indicate that these adversarial examples are too visually perceptible to fool human-beings. When FID score is in the range of [2,16], AdvINN outperforms all comparison methods against JPEG compression, and outperforms most comparison methods against bit-depth reduction and NRP except SSAH. With the same FID constraint, AdvINN can generate more robust adversarial examples, and with the same attacking success rate, AdvINN can achieve a lower visual perceptibility.

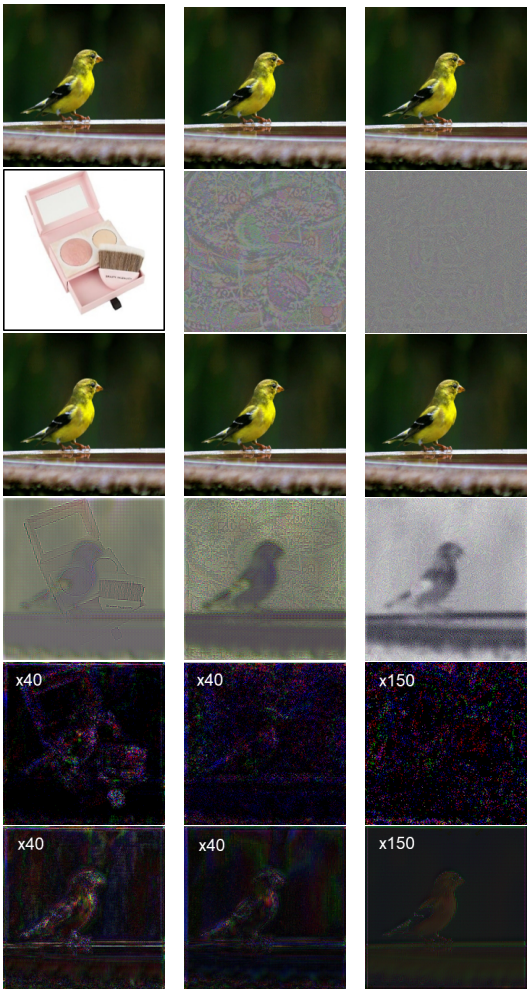
Effectiveness of IIEM

The Invertible Information Exchange Module (IIEM) is with the information preservation property, and performs feature-level information exchanging between the input clean image and the target image. Table 2 shows the performance of AdvINN with IIEM, without IIEM, and using a CNN to replace IIEM. When IIEM is not used in AdvINN, the adversarial examples are generated by directly combining of the benign image and the learned target image. From Table 2, we can observe that the results of w/o IIEM are with a significant deterioration compared to those of w/ IIEM. This indicates that IIEM is an indispensable component in AdvINN and can improve the image quality and accelerate convergence. In Table 2, w/ CNN denotes that IIEM is replaced by a CNN (Xiao et al. 2018). We can see that the scores of all metrics further deteriorate except the FID score, moreover it takes much more iterations to converge. This result confirms that the information preservation property of IIEM is essential to the success of AdvINN.

Visualization and Analysis

Fig. 8 visualizes input images, output images, adversarial perturbations and the estimated dropped information when using different target images. The 1-st row shows the input clean images with the class label *goldfinch*. In the 2-nd row, we show the target images with the highest confidence of the *face powder*, the targeted UAP generated by (Zhao, Liu, and Larson 2021) and the classifier guided target image. From the output adversarial examples $\boldsymbol{x}_{adv}$ in the third row, we cannot see noticeable visual differences compared to the input clean image in all cases. This further verifies the effectiveness of the proposed method. In order to further interpret the results of AdvINN, we visualizes the residual images $\boldsymbol{x}_r$ ($\boldsymbol{x}_r$ is normalized for clearer perception), the absolute difference between $\boldsymbol{x}_{cln}$ and $\boldsymbol{x}_{tgt}$, and the estimated dropped information[3] in row 4 to row 6, respectively. We can see that the adversarial example generated by AdvINN-HCT contains the boundary information of the target image and discards the discriminant high-frequency features of the goldfinch; the adversarial example crafted by AdvINN-UAP includes some universal adversarial perturbation patterns and drops certain key features corresponding to head, chest and tail of the goldfinch; the adversarial example generated by AdvINN-CGT only adds minor modification to the clean images which is enlarged by 150 times for better visualization, and drops slight information corresponding to the shape of the goldfinch.

[3]The dropped information is estimated by replacing the target image with a constant image (with no information) while keeping parameters of AdvINN fixed, therefore the generated residual image only contains the dropped information from the clean image.



(a) HCT (b) UAP (c) CGT

Figure 8: Visualization of $\boldsymbol{x}_{cln}$, $\boldsymbol{x}_{tgt}$, $\boldsymbol{x}_{adv}$, $\boldsymbol{x}_r$, $|\boldsymbol{x}_{adv} - \boldsymbol{x}_{cln}|$, and the estimated dropped information with different target images.

In summary, we can observe that AdvINN drops discriminant information (high-frequency details or shape information) of clean images and adds class-specific information from the target images simultaneously.

Ablation Study

Adversarial Budget ϵ

The adversarial budget controls the maximum amplitude of the perturbation allowed on the generated adversarial examples. The performance of certain adversarial attack methods would be limited if a smaller adversarial budget is required. Table 3 shows the performance of AdvINN with three different adversarial budgets, *i.e.*, 4/255, 8/255, and 16/255. We can see that there is no significant difference on the performance of AdvINN under the different constraints. This indicates that the quality of the adversarial examples generated by AdvINN does not limited by the maximum perturbation constraint, and AdvINN method maintains a stable convergence speed and achieves high attacking success rates even under a stricter perturbation budget. For fair comparisons, the adversarial budget ϵ is still set to 8/255 by l_∞-norm which is consistent with other comparison methods.

ϵ	$l_\infty \downarrow$	LPIPS↓	FID↓	Iter↓	ASR(%)↑
4/255	0.0172	0.0118	1.575	341	100.0
8/255	0.0281	0.0118	1.594	321	100.0
16/255	0.0332	0.0119	1.568	325	100.0

Table 3: Ablation study: the performance of AdvINN under different adversarial budget constraints.

Classifier	$l_2 \downarrow$	LPIPS↓	FID↓	ASR(%)↑
Inception_v3	4.57	0.0155	2.600	100.0
Densenet121	2.51	0.0114	1.604	100.0

Table 4: The performance of AdvINN on different classifiers. The adversarial weights λ_{adv} are set to 10 and 3 on Inception_v3 and Densenet121, respectively.

Results on Other Classifiers

We have also tested AdvINN on the other two classifiers: Densenet121 (Huang et al. 2017) and Inception_v3 (Szegedy et al. 2016). Densenet121 fully utilizes the features by dense connection, which largely reduces the number of parameters. But the reduction in parameters leads to weaker robustness against adversarial attacks. Inception_v3 utilizes convolution kernels of different sizes and is more complex in structure, but more robust against adversarial attacks.

We adjust the adversarial weights λ_{adv} on different classifiers for better performance. Specifically, λ_{adv} is set to 10 on Inception_v3 and 3 on Densenet121, respectively. Table 4 shows the experimental results and with both conditions, AdvINN achieves 100% success attacking rate. We observe that AdvINN can be simply applied to other classifiers. Even with a more robust classifier, AdvINN succeeds in generating more imperceptible adversarial examples.

Conclusion

In this paper, we propose a novel adversarial attack framework, termed as AdvINN, to generate adversarial examples based on Invertible Neural Networks (INNs). By utilizing the information preservation property of INNs, the proposed Invertible Information Exchange Module, driven by the loss functions, performs information exchanging at the feature level and achieves simultaneously dropping discriminant information of clean images and adding class-specific features of the target images to craft adversaries. Moreover, three target image selection and learning approaches have been carefully investigated and analyzed. Extensive experimental results have shown that the proposed AdvINN method can generate visually less perceptible and more robust adversarial examples compared to the state-of-the-art methods.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Project 62201600 and U1811462, and NUDT Research Project ZK22-56.

References

Akhtar, N.; and Mian, A. S. 2018. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access*, 6: 14410–14430.

Ardizzone, L.; Lüth, C.; Kruse, J.; Rother, C.; and Köthe, U. 2019. Guided image generation with conditional invertible neural networks. *arXiv preprint arXiv:1907.02392*.

Benz, P.; Zhang, C.; Imtiaz, T.; and Kweon, I. S. 2020. Double targeted universal adversarial perturbations. In *Proceedings of the Asian Conference on Computer Vision*.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. IEEE.

Cheng, K. L.; Xie, Y.; and Chen, Q. 2021. Iicnet: A generic framework for reversible image conversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1991–2000.

Croce, F.; and Hein, M. 2019. Sparse and imperceivable adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4724–4732.

Das, N.; Shanbhogue, M.; Chen, S. T.; Hohman, F.; Li, S.; Chen, L.; Kounavis, M. E.; and Chau, D. H. 2018. Shield: Fast, Practical Defense and Vaccination for Deep Learning using JPEG Compression. In *the 24th ACM SIGKDD International Conference*.

Dinh, L.; Krueger, D.; and Bengio, Y. 2014. Nice: Nonlinear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2016. Density estimation using real NVP. *arXiv preprint arXiv:1605.08803*.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Duan, R.; Chen, Y.; Niu, D.; Yang, Y.; Qin, A. K.; and He, Y. 2021. AdvDrop: Adversarial Attack to DNNs by Dropping Information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7506–7515.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Guan, Z.; Jing, J.; Deng, X.; Xu, M.; Jiang, L.; Zhang, Z.; and Li, Y. 2022. DeepMIH: Deep Invertible Network for Multiple Image Hiding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Guo, C.; Rana, M.; Cisse, M.; and Van Der Maaten, L. 2017. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. X. 2021. Natural Adversarial Examples. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15257–15266.

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.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Huang, J.-J.; and Dragotti, P. L. 2021. LINN: Lifting inspired invertible neural network for image denoising. In *2021 29th European Signal Processing Conference (EUSIPCO)*, 636–640. IEEE.

Huang, J.-J.; and Dragotti, P. L. 2022. WINNet: Wavelet-Inspired Invertible Network for Image Denoising. *IEEE Transactions on Image Processing*, 31: 4377–4392.

Huang, J.-J.; Liu, T.; Yang, Z.; Fu, S.; Zhao, W.; and Dragotti, P. L. 2022. DURRNet: Deep Unfolded Single Image Reflection Removal Network. *arXiv preprint arXiv:2203.06306*.

Huang, Y.-C.; Chen, Y.-H.; Lu, C.-Y.; Wang, H.-P.; Peng, W.-H.; and Huang, C.-C. 2021. Video Rescaling Networks with Joint Optimization Strategies for Downscaling and Upscaling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3527–3536.

Jacobsen, J.-H.; Smeulders, A.; and Oyallon, E. 2018. i-RevNet: Deep Invertible Networks. In *ICLR 2018-International Conference on Learning Representations*.

Jia, S.; Ma, C.; Yao, T.; Yin, B.; Ding, S.; and Yang, X. 2022. Exploring Frequency Adversarial Attacks for Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4103–4112.

Jing, J.; Deng, X.; Xu, M.; Wang, J.; and Guan, Z. 2021. HiNet: Deep Image Hiding by Invertible Network. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4713–4722.

Khrulkov, V.; and Oseledets, I. 2018. Art of Singular Vectors and Universal Adversarial Perturbations. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8562–8570.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Lin, J.; Song, C.; He, K.; Wang, L.; and Hopcroft, J. E. 2019. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In *International Conference on Learning Representations*.

Liu, Y.; Qin, Z.; Anwar, S.; Ji, P.; Kim, D.; Caldwell, S.; and Gedeon, T. 2021. Invertible Denoising Network: A Light Solution for Real Noise Removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13365–13374.

Lu, S.-P.; Wang, R.; Zhong, T.; and Rosin, P. L. 2021. Large-capacity image steganography based on invertible neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10816–10825.

Luo, C.; Lin, Q.; Xie, W.; Wu, B.; Xie, J.; and Shen, L. 2022. Frequency-driven Imperceptible Adversarial Attack on Semantic Similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15315–15324.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Mallat, S. G. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7): 674–693.

Mohaghegh Dolatabadi, H.; Erfani, S.; and Leckie, C. 2020. Advflow: Inconspicuous black-box adversarial attacks using normalizing flows. *Advances in Neural Information Processing Systems*, 33: 15871–15884.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal Adversarial Perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2574–2582.

Naseer, M.; Khan, S.; Hayat, M.; Khan, F. S.; and Porikli, F. 2020. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 262–271.

Poursaeed, O.; Katsman, I.; Gao, B.; and Belongie, S. 2018. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4422–4431.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Tian, Y.; Pan, J.; Yang, S.; Zhang, X.; He, S.; and Jin, Y. 2022. Imperceptible and Sparse Adversarial Attacks via a Dual-Population Based Constrained Evolutionary Algorithm. *IEEE Transactions on Artificial Intelligence*.

Wang, X.; He, X.; Wang, J.; and He, K. 2021. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16158–16167.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, 0–0.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xiao, C.; Li, B.; Zhu, J.-Y.; He, W.; Liu, M.; and Song, D. 2018. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*.

Xiao, M.; Zheng, S.; Liu, C.; Lin, Z.; and Liu, T.-Y. 2022. Invertible Rescaling Network and Its Extensions. *International Journal of Computer Vision*, 1–26.

Zhang, C.; Benz, P.; Imtiaz, T.; and Kweon, I. S. 2020. Understanding adversarial examples from the mutual influence of images and perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14521–14530.

Zhang, M.; Pan, Z.; Zhou, X.; and Kuo, C.-C. J. 2022. Enhancing Image Rescaling using Dual Latent Variables in Invertible Neural Network. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5602–5610.

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.

Zhao, R.; Liu, T.; Xiao, J.; Lun, D. P.; and Lam, K.-M. 2021. Invertible image decolorization. *IEEE Transactions on Image Processing*, 30: 6081–6095.

Zhao, Z.; Liu, Z.; and Larson, M. 2020. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1039–1048.

Zhao, Z.; Liu, Z.; and Larson, M. 2021. On Success and Simplicity: A Second Look at Transferable Targeted Attacks. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 6115–6128. Curran Associates, Inc.

Zhu, X.; Li, Z.; Zhang, X.-Y.; Li, C.; Liu, Y.; and Xue, Z. 2019. Residual invertible spatio-temporal network for video super-resolution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 5981–5988.