Imperceptible Adversarial Attack via Invertible Neural Networks

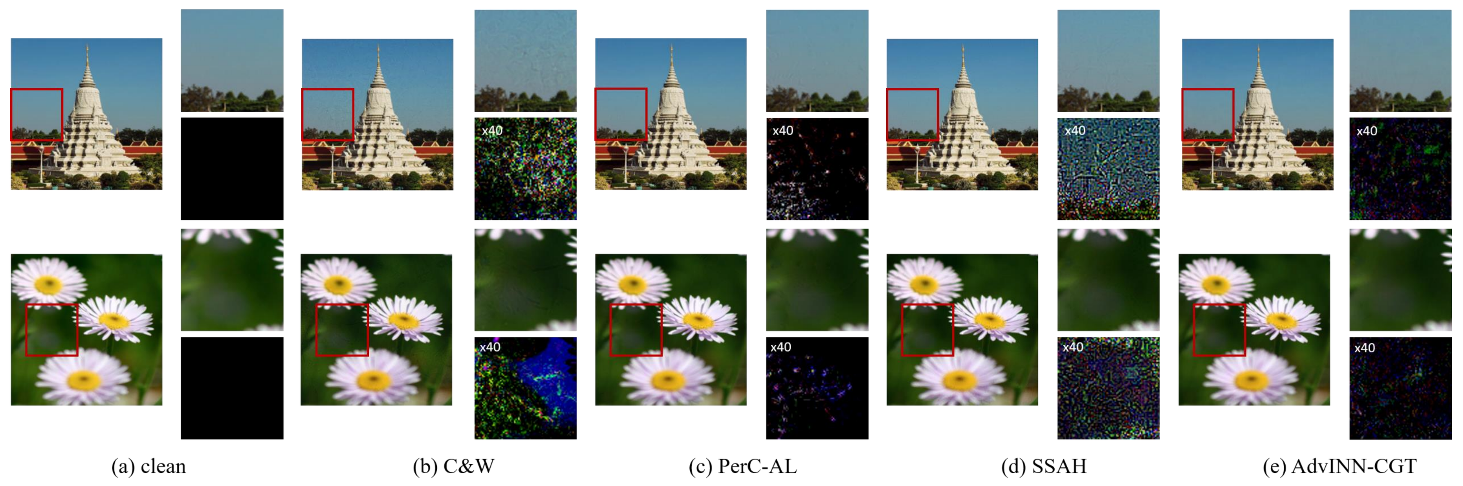
Zihan Chen# Ziyue Wang# Junjie Huang* Wentao Zhao Xiao Liu Dejian Guan

National University of Defense Technology, China

Introduction & Motivation

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.

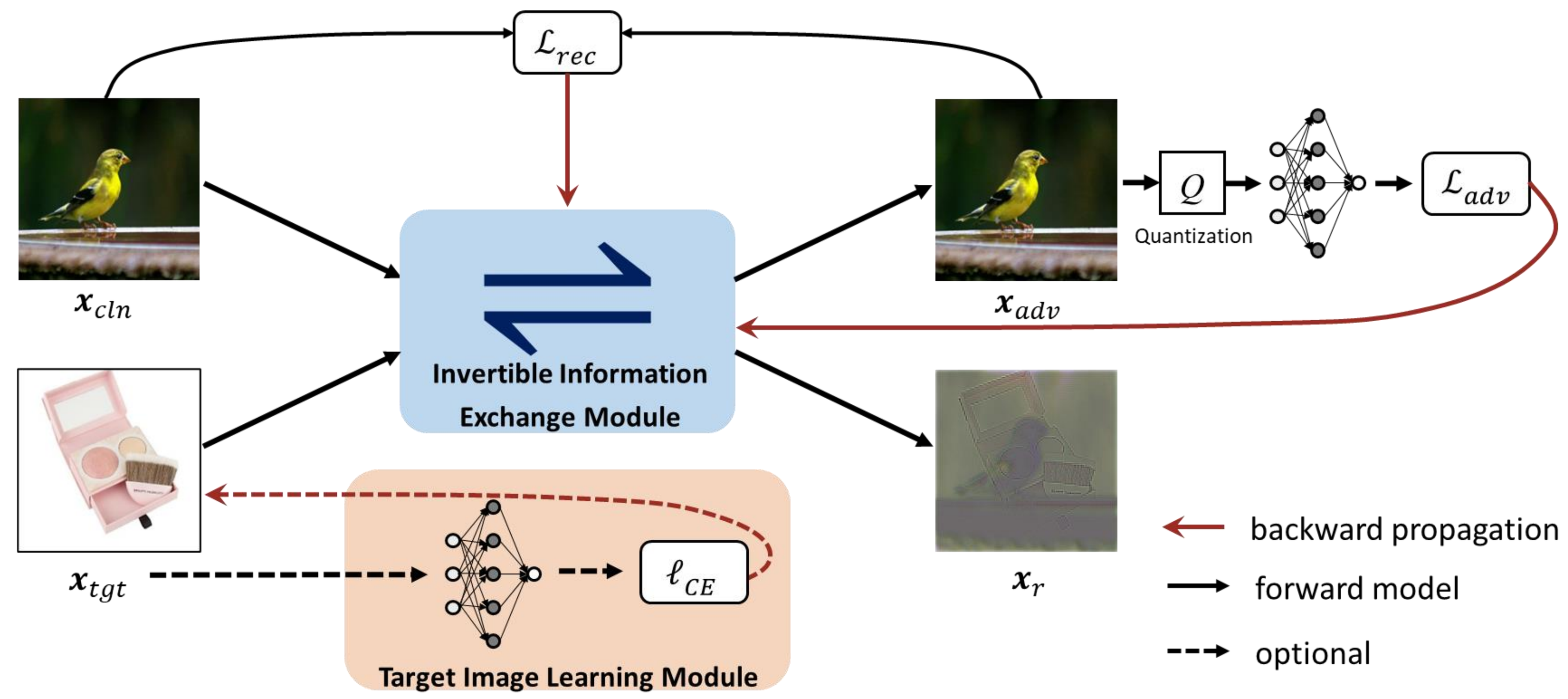
- ✓ The methods based on adding adversarial perturbations may lead to perceptible noise patterns and noticeable increase of image storage size.
- ✓ The method of dropping existing information has limited performance on targeted attacks.

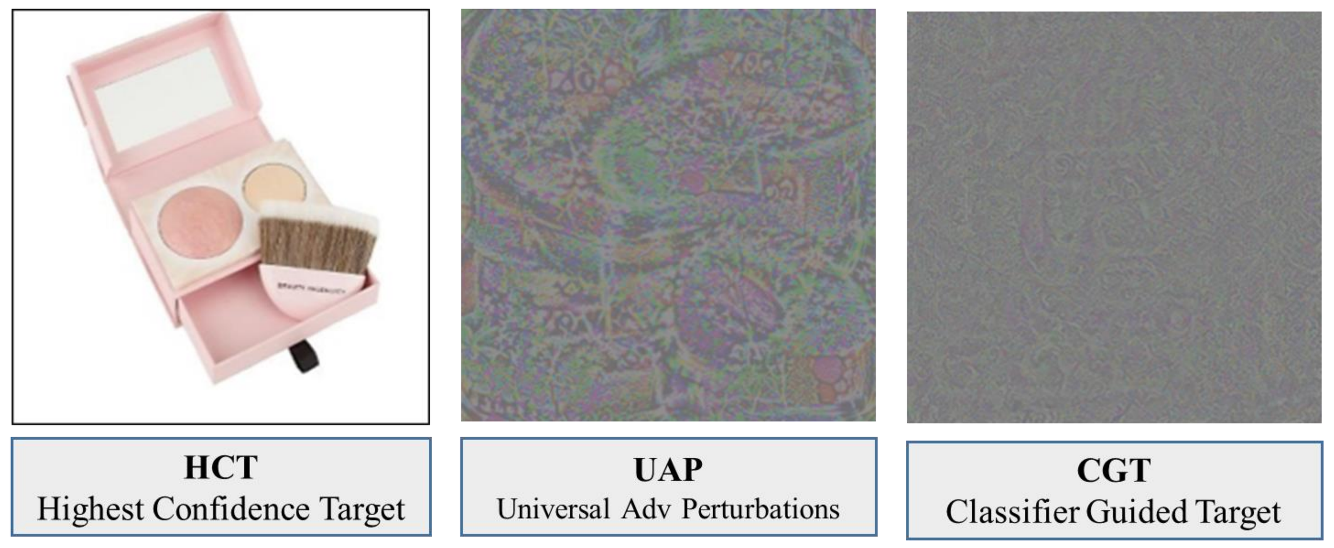


We propose a novel Adversarial attack method using Invertible Neural Networks, termed AdvINN, by leveraging the information preservation property of Invertible Neural Networks (INNs) to achieve simultaneously adding extra information and dropping existing details.

Proposed AdvINN & Target Images Selection and Learning

- ➢ 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].





- ➢ **Target Image Selection and Learning**
  - ✓ **Highest Confidence Target Image (HCT):** select the image with the highest confidence in each class as the target image.
  - ✓ **UAP as Target Image (UAP):** utilize the targeted universal adversarial perturbation as target images.
  - ✓ **Classifier Guided Target Image (CGT):** 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.

Experiments and Visualization Results

Table 1. Accuracy and evaluation metrics on different methods.

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

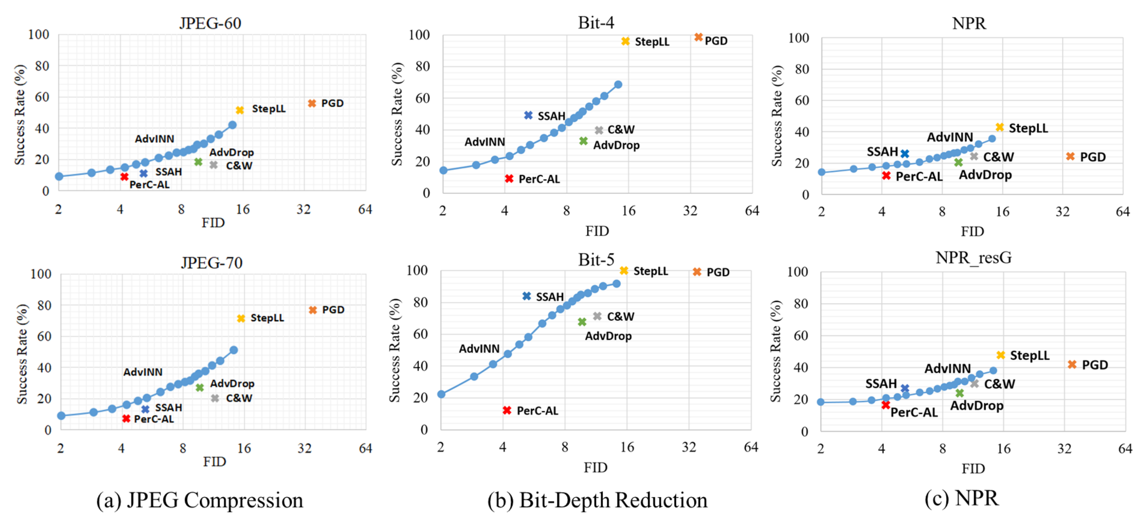


Figure 1. Evaluation on robustness of adversarial examples.

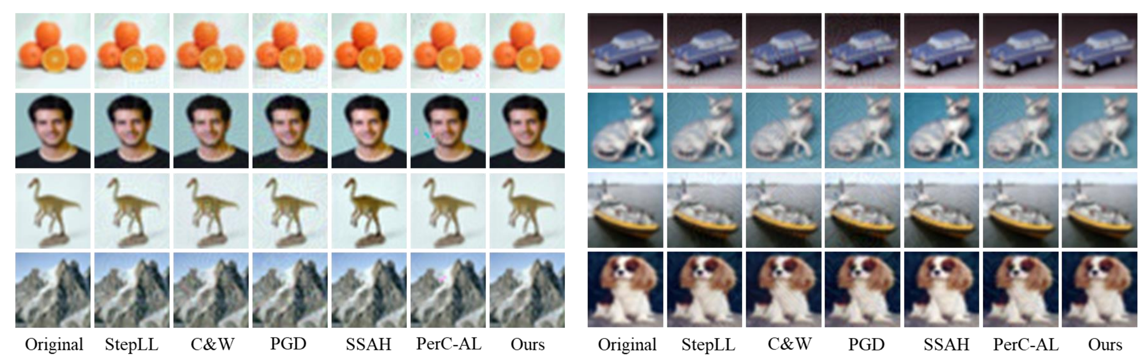


Figure 2. More adversarial examples crafted by different methods on CIFAR-100 and CIFAR-10.

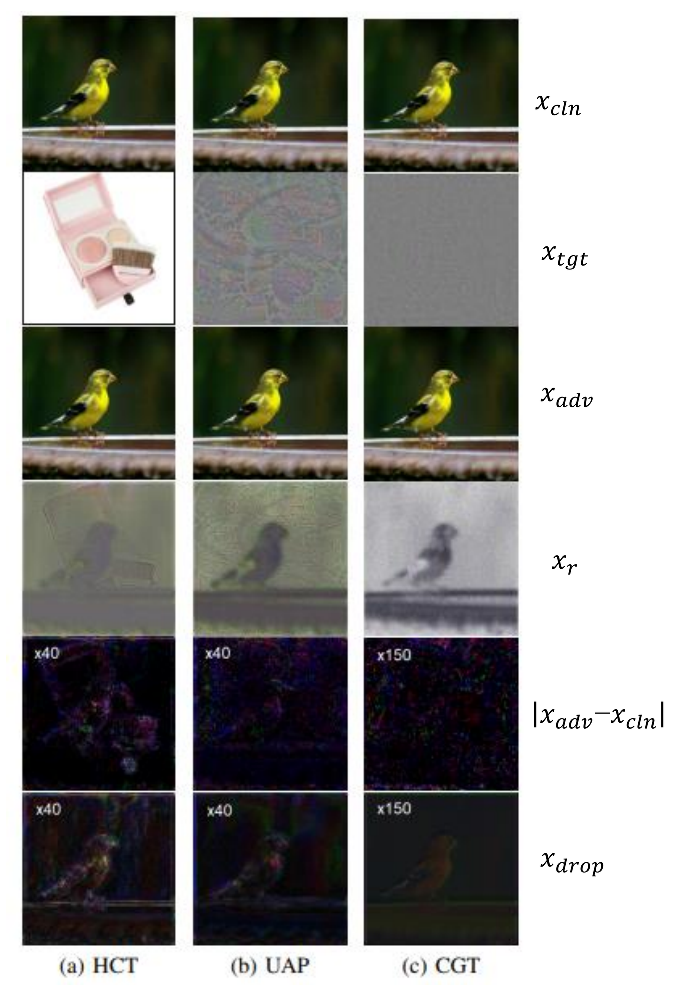


Figure 3. Visualization results with different target images.

Contributions & Conclusions

- ✓ 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.

website: https://advinn.github.io/ code: https://github.com/jjhuangcs/AdvINN