国防科技大学
National University of Defense Technology

Imperceptible Adversarial Attack via Invertible Neural Networks

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

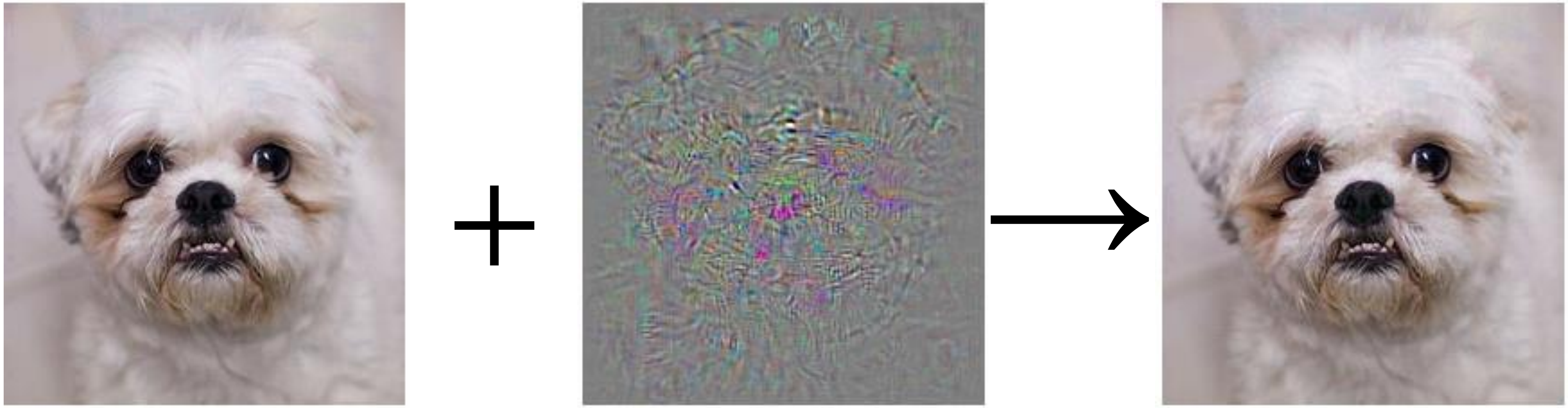
College of Computer Science
National University of Defense Techology

Website Paper Code

1. Background

Adversarial Attacks

- Deep neural networks are vulnerable to adversarial examples, and can misclassify the adversarial example to an erroneous class label



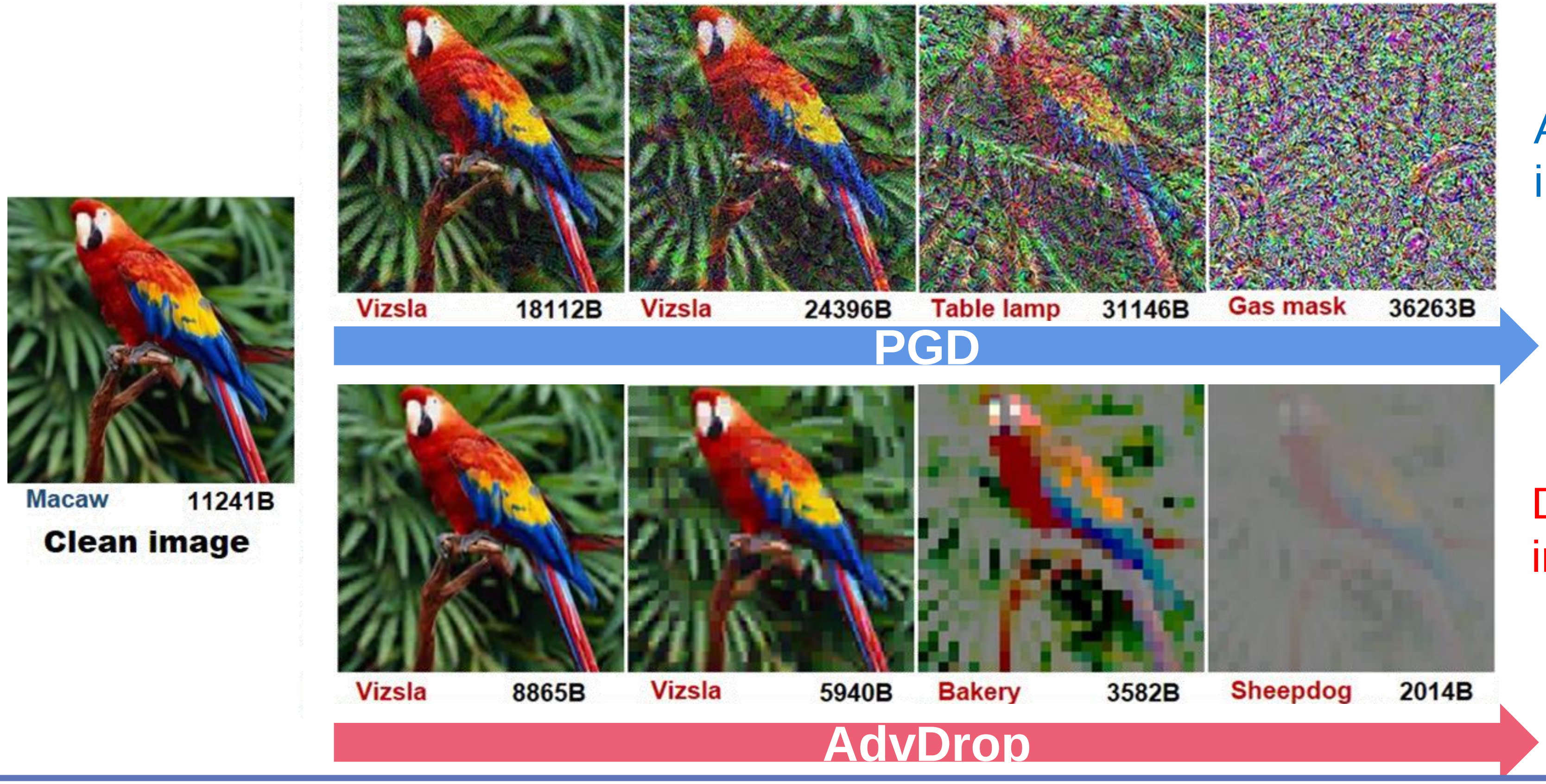
Dog Perturbation Ostrich

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." ICLR 2015.

1. Background

Adversarial Attacks

- Adversarial examples can be generated by **adding** or **dropping** information



Adding class-specific information of the target class

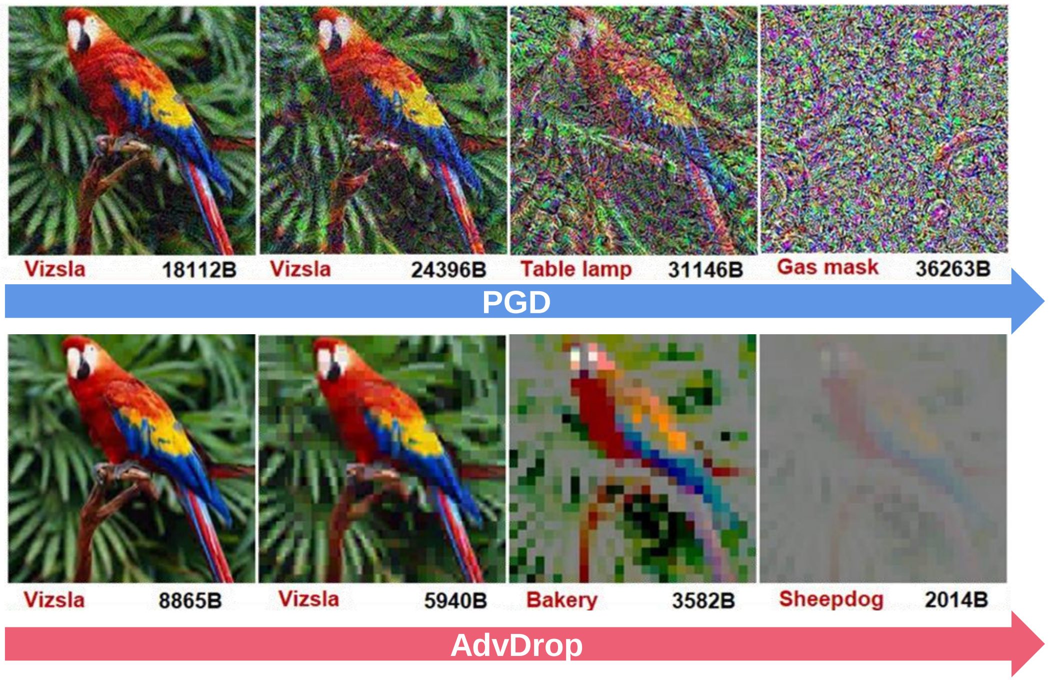
Dropping discriminant information of the original class

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." ICLR 2015.
Duan, Ranjie, et al. "Advdrop: Adversarial attack to DNNs by dropping information." CVPR 2021.

1. Background

Adversarial Attacks

- Adversarial examples can be generated by **adding** or **dropping** information



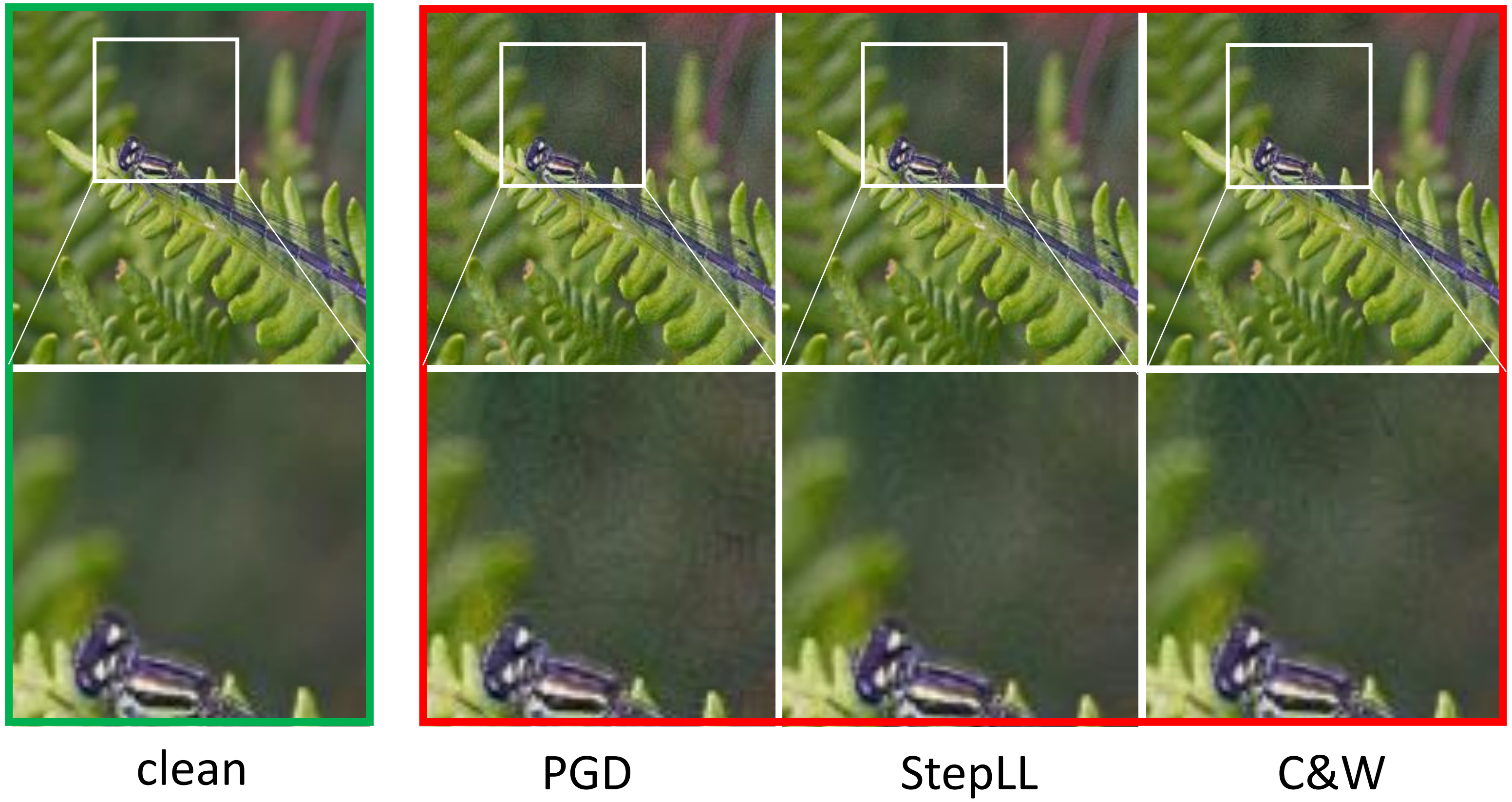
Advantages	Limitations
• Flexible in targeted/untargeted attacks	• Perceptible noise patterns • Noticeable increase of image size
• Robust to denoising-based defense • Higher imperceptible	• Limited performance on targeted attack • Blocking artifacts

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." ICLR 2015.
Duan, Ranjie, et al. "Advdrop: Adversarial attack to DNNs by dropping information." CVPR 2021.

1. Background

Imperceptible Adversarial Attacks

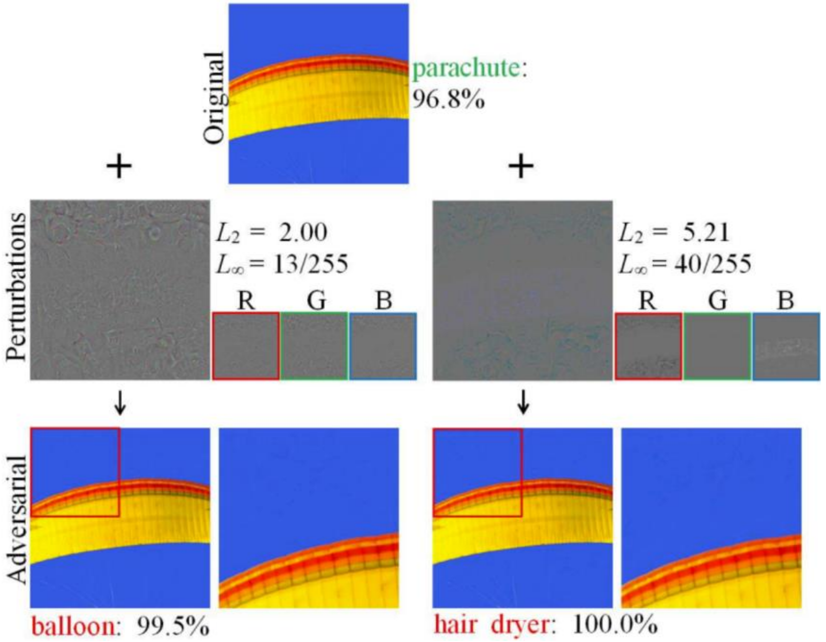
- Imperceptibility is an important criterion for adversarial attacks, however, it is not well attained by many well-known adversarial attack methods

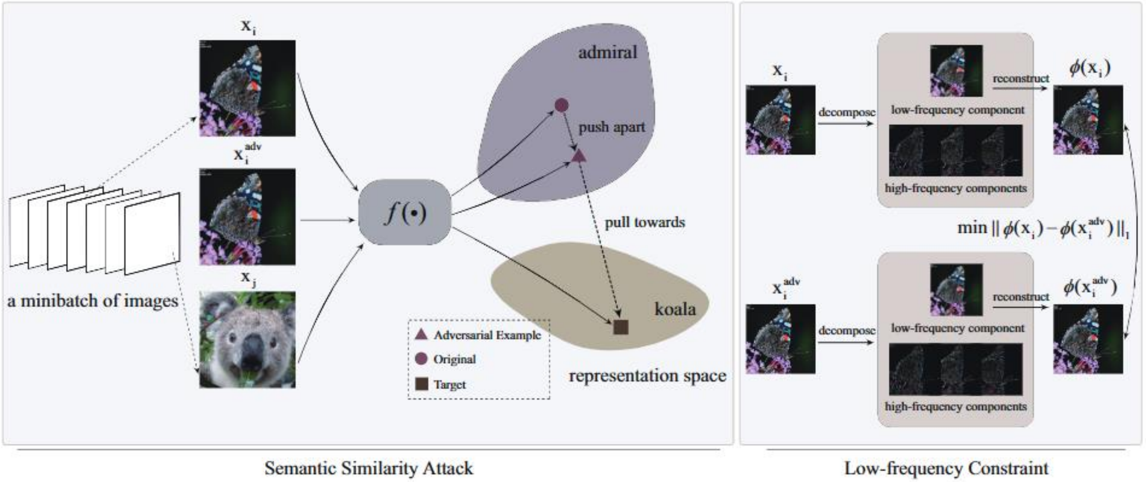


1. Background

Imperceptible Adversarial Attacks

PerC-AL [Zhao et al.2020]: adversarial perturbations are optimized in terms of perceptual color distance leading to improve visual imperceptibility.





SSAH [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.

Zhao, Zhengyu et al. "Towards Large Yet Imperceptible Adversarial Image Perturbations with Perceptual Color Distance." CVPR 2020.
Luo, Cheng, et al. "Frequency-driven Imperceptible Adversarial Attack on Semantic Similarity." CVPR 2022.

2. Adversarial Attack via Invertible Neural Networks

Motivation:

- Whether it is possible to craft **imperceptible and robust** adversarial examples by simultaneously **Adding** and **Dropping** information **in an unified framework**?

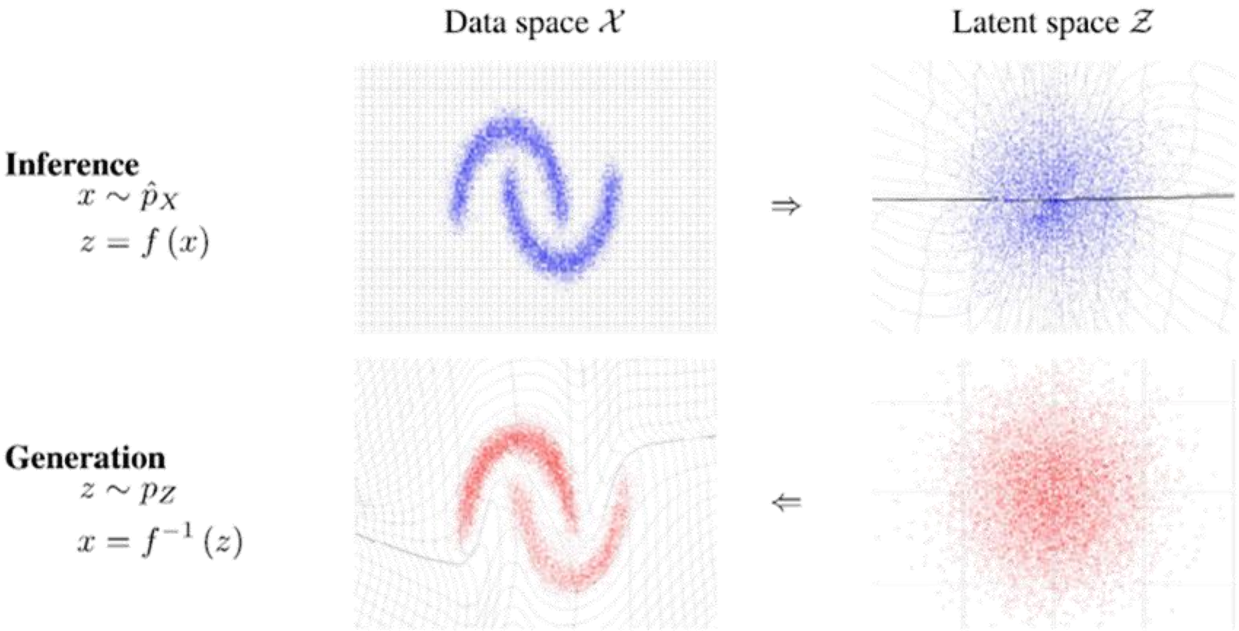
Idea:

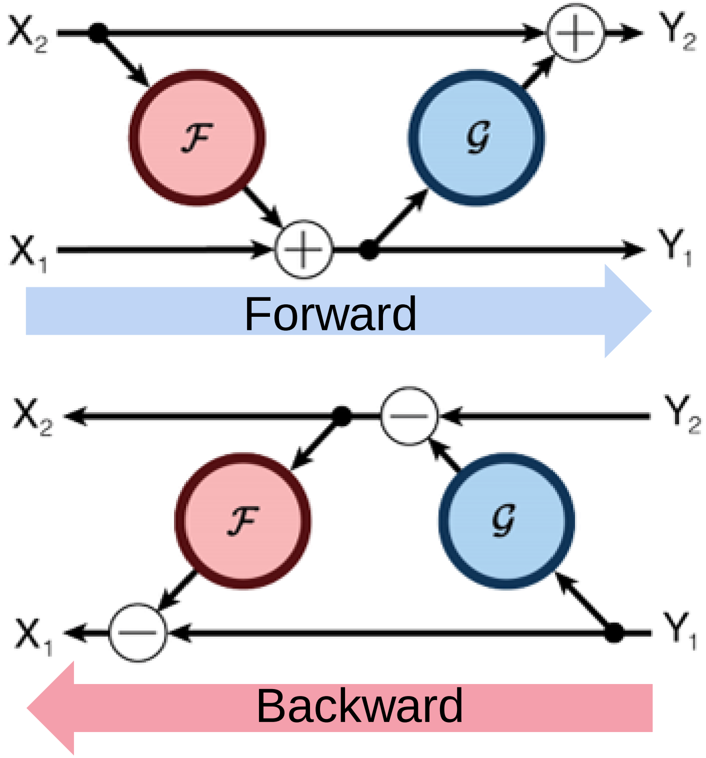
- Learning a non-linear transform with information preservation property to interchange information between the clean image and the target image
    - ✓**Invertible Neural Networks!**

2. Adversarial Attack via Invertible Neural Networks

Invertible Neural Networks

- INNs are bijective function approximators





Gomez, Aidan N., Mengye Ren, Raquel Urtasun, and Roger B. Grosse. "The Reversible Residual Network: Backpropagation without Storing Activations." NeurIPS 2017.
Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using real nvp." ICLR 2017.

2. Adversarial Attack via Invertible Neural Networks

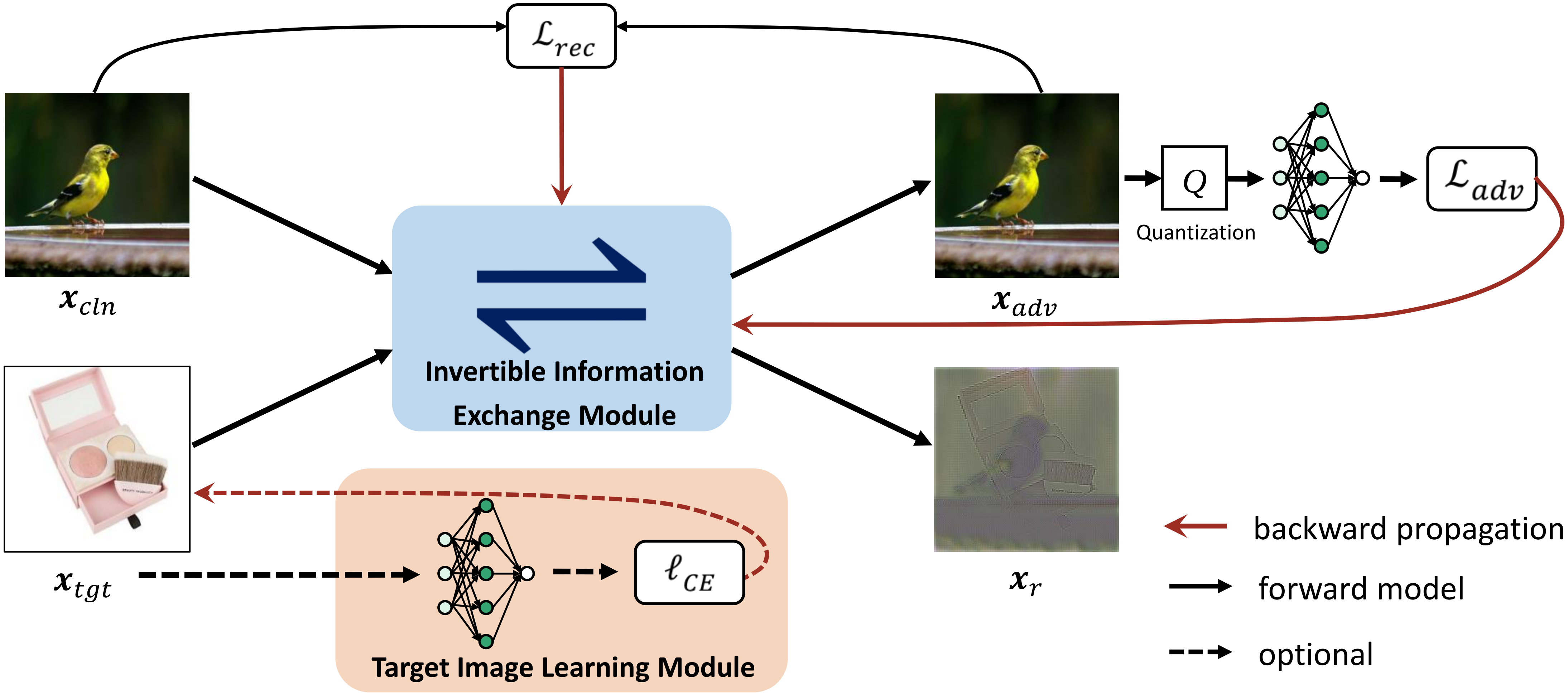
Overview

- **Invertible Information Exchange Module:** generate an adversarial image $\boldsymbol{x}_{adv}$ by dropping discriminate information of the original class while adding adversarial details from a target image $\boldsymbol{x}_{tgt}$

- **Target Image Learning Module:** selecting or learning the target image $\boldsymbol{x}_{tgt}$ as the source information for adding adversarial perturbation
  - Highest Confidence Target Image
  - Universal Adversarial Perturbation as Target Image
  - Classifier Guided Target Image

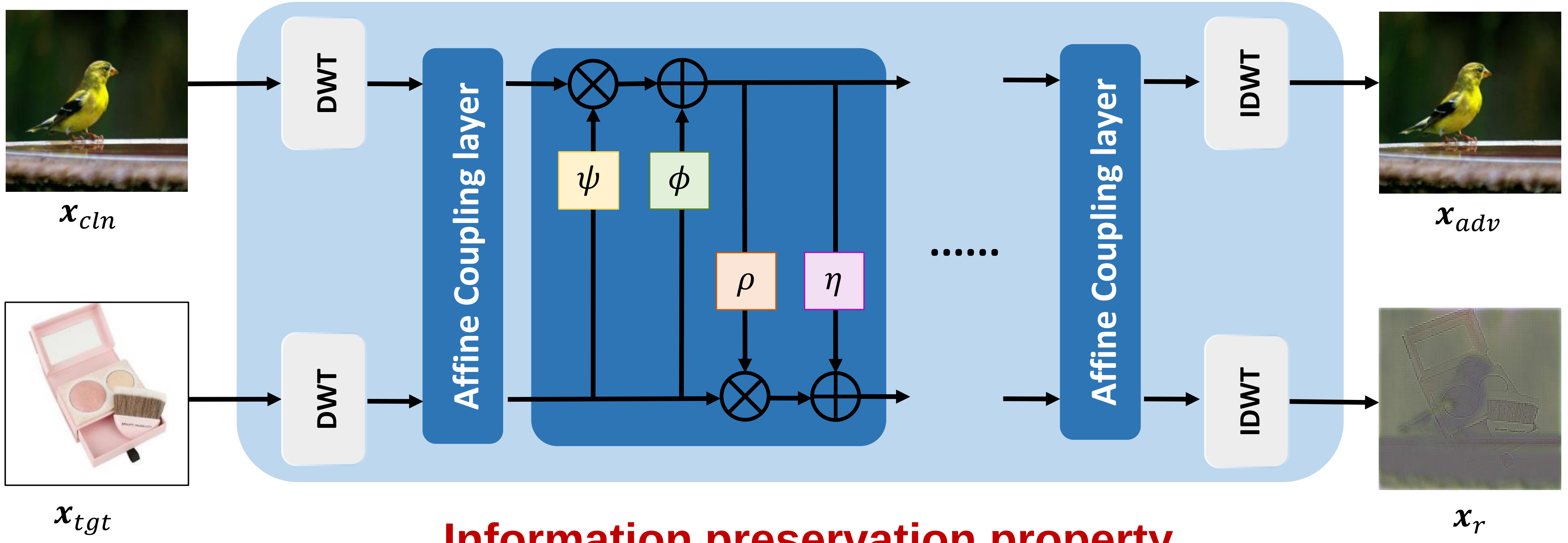
2. Adversarial Attack via Invertible Neural Networks

Overview



2. **Adv**ersarial Attack via **I**nvertible **N**eural **N**etworks

Invertible Information Exchange Module

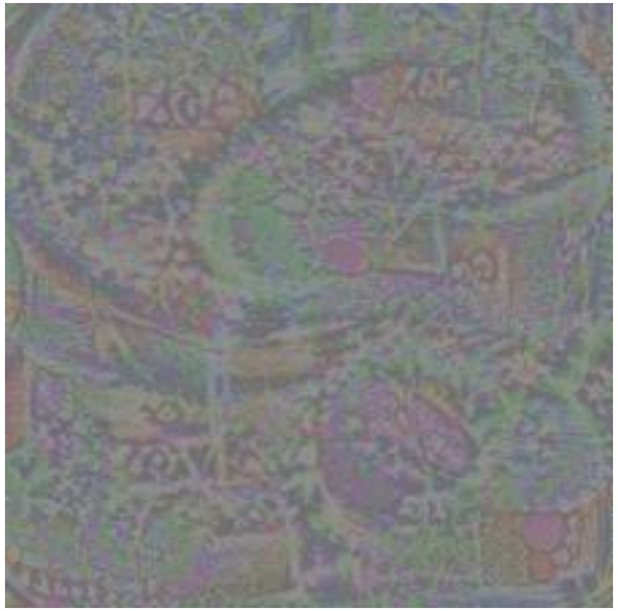


Information preservation property

$$\begin{cases} x_{adv} = x_{cln} - \sigma + \delta, \\ x_r \quad = x_{tgt} + \sigma - \delta. \end{cases}$$

2. **Adv**ersarial Attack via **I**nvertible **N**eural **N**etworks

Target Image Learning Module



Highest Confidence Target (HCT)

- The highest confidence in each class
- Contain a considerable amount of information unrelated to the target class

Universal Adversarial Puturbation (UAP)

- Universal Adversarial Puturbation is optimized in a data-free manner
- Need to prepared first

Zhao, Zhengyu, Zhuoran Liu, and Martha Larson. "On success and simplicity: A second look at transferable targeted attacks." NeurIPS, 2021

2. Adversarial Attack via Invertible Neural Networks

Target Image Learning Module

Classifier Guided Target (CGT)

- Learnable and embed more discriminant information of the target class
- Online optimized

Algorithm 2: AdvINN-CGT

Input : clean image $\boldsymbol{x}_{cln}$, classifier guided image $\boldsymbol{x}_{cgt}$, adversarial budget ϵ, confidence κ, learning rate lr_1, learning rate lr_2;
Output: Adversarial image $\boldsymbol{x}_{adv}$;

Initialize the parameters of AdvINN: $\boldsymbol{\theta}$;
Initialize $\boldsymbol{x}_{cgt}$ with all 0.5;
while $\boldsymbol{x}_{adv}$ *is not adversarial* **do**
    $(\boldsymbol{x}_{adv}, \boldsymbol{x}_r) \leftarrow f_{\boldsymbol{\theta}}(\boldsymbol{x}_{cln}, \boldsymbol{x}_{cgt})$
    $\boldsymbol{x}_{adv} \leftarrow \min(\boldsymbol{x}_{cln} + \epsilon, \max(\boldsymbol{x}_{adv}, \boldsymbol{x}_{cln} - \epsilon))$;
    $p_{tgt} \leftarrow g(\boldsymbol{x}_{adv})$;
    if $p_{tgt} \leq \kappa$ **then**
        Update loss function $\mathcal{L}_{total}$;
        Update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + lr_1 \cdot \text{Adam}(\mathcal{L}_{total})$;
        Update loss function $\mathcal{L}_{cgt}$;
        Update $\boldsymbol{x}_{cgt} \leftarrow \boldsymbol{x}_{cgt} + lr_2 \cdot \text{Adam}(\mathcal{L}_{cgt})$;
    else
        break;
    end
end
return: $\boldsymbol{x}_{adv}$.

---

3. Experimental Results

Experiment Settings

- Comparison methods: PGD, StepLL, C&W and AdvDrop, PerC-AL, SSAH
- Least-likely objective: avoid choosing closely related classes
- Target classifier: ResNet50
- Adversarial budget: 8/255 with respect to l_∞-norm
- Evaluation metrics: l_2-norm, l_∞-norm, SSIM, LPIPS, FID
- Defense methods: JPEG compression, bit-rate reduction, Neural Representation Purifier (NRP)

3. Experimental Results

Quantitive Comparison on ImageNet-1K

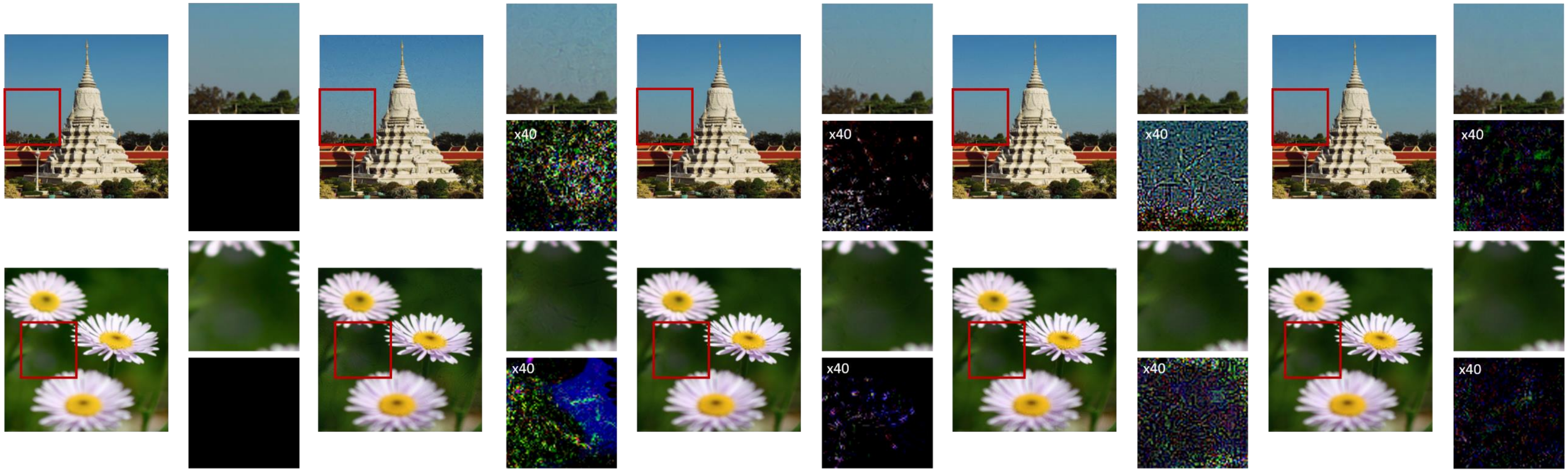
Table 1: Accuracy and evaluation metrics on different methods. All methods use $\epsilon = 8/255$ as the adversarial budget. ASR donates the accuracy of adversarial attacks. ↑ means the value is higher the better, and vice versa. (The best and the second best result in each column is in bold and underline.)

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	<u>0.995</u>	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	<u>99.8</u>
	AdvINN-HCT	5.73	**0.03**	0.991	<u>0.0206</u>	3.661	**100.0**
	AdvINN-UAP	5.84	**0.03**	0.990	0.0212	<u>2.900</u>	**100.0**
	AdvINN-CGT	<u>2.66</u>	**0.03**	**0.996**	**0.0118**	**1.594**	**100.0**

Less perceptible adversarial examples with 100% attacking success rate!

3. Experimental Results

Visual Comparison on ImageNet-1K



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

Our results are more imperceptible on both smooth region and edge region.

3. Experimental Results

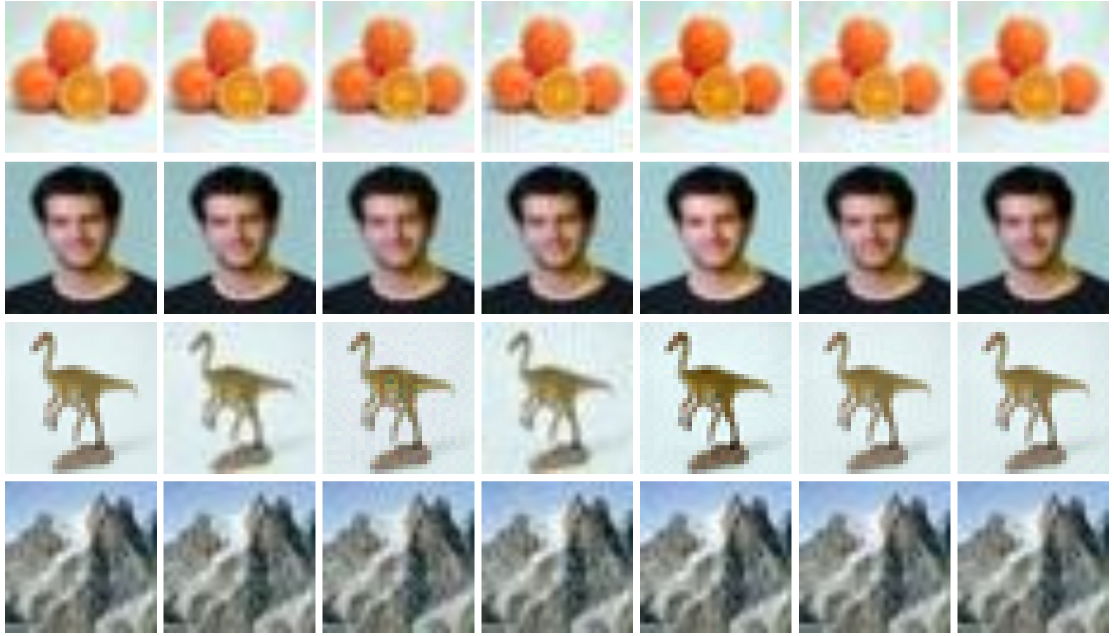
Quantitive Comparison on CIFAR-100 and CIFAR-10

Dataset	Methods	$l_2 \downarrow$	$l_\infty \downarrow$	SSIM↑	LPIPS↓	FID↓	ASR(%)↑
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**	<u>0.991</u>	**0.0035**	**3.413**	98.3
	AdvINN-UAP	<u>0.27</u>	**0.03**	**0.993**	<u>0.0037</u>	3.982	**99.6**
	AdvINN-CGT	**0.23**	**0.03**	**0.993**	<u>0.0037</u>	<u>3.921</u>	<u>99.5</u>
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**	<u>0.993</u>	0.0180	3.654	<u>99.9</u>
	AvdINN-HCT	<u>0.18</u>	**0.03**	**0.995**	0.0033	2.627	<u>99.9</u>
	AdvINN-UAP	0.19	**0.03**	**0.995**	<u>0.0031</u>	2.791	<u>99.9</u>
	AdvINN-CGT	**0.17**	**0.03**	**0.995**	**0.0030**	<u>2.480</u>	<u>99.9</u>

3. Experimental Results

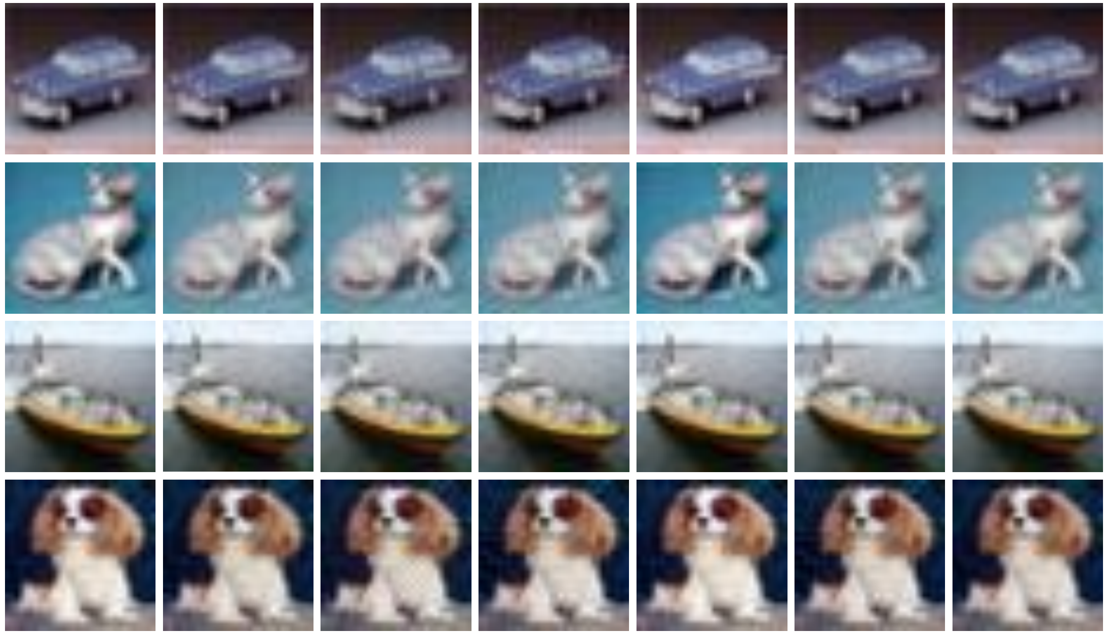
Visual Comparison on CIFAR-100 and CIFAR-10

CIFAR-100



Original StepLL C&W PGD SSAH PerC-AL Ours

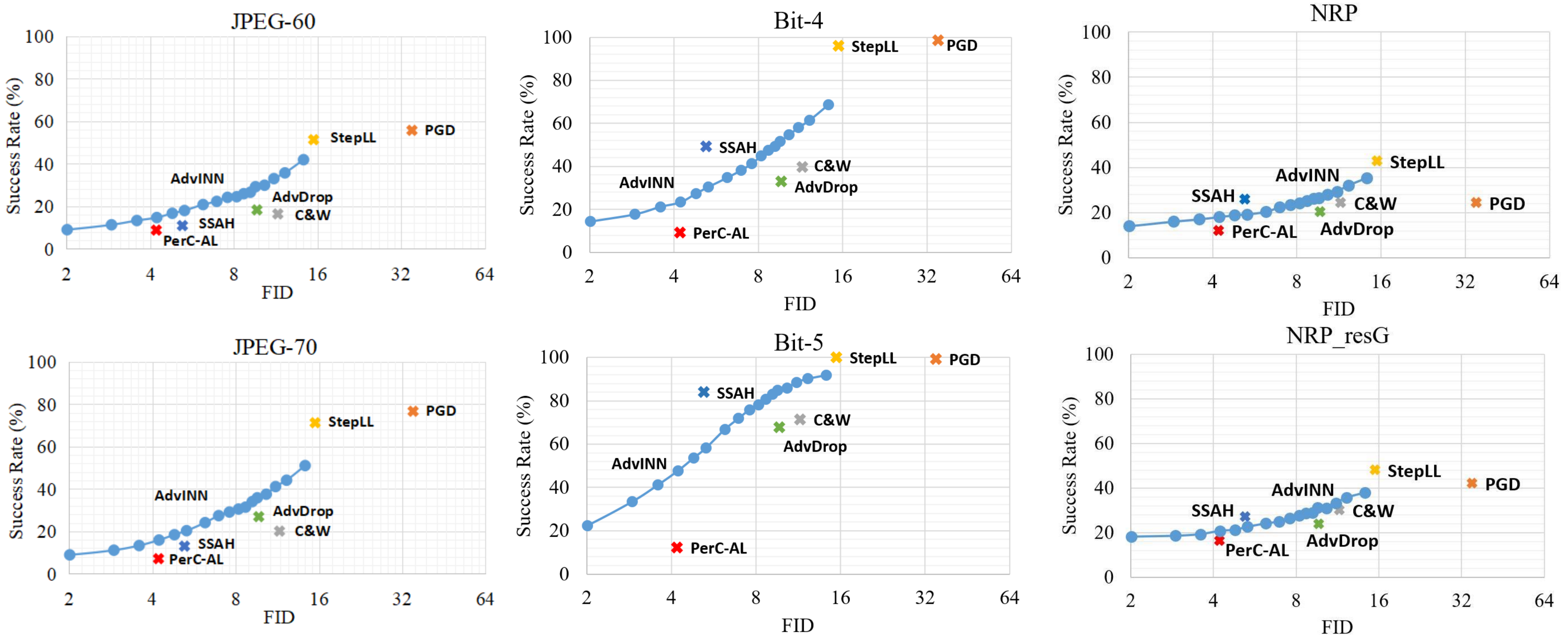
CIFAR-10



Original StepLL C&W PGD SSAH PerC-AL Ours

3. Experimental Results

Robustness Towards Adversarial Defense

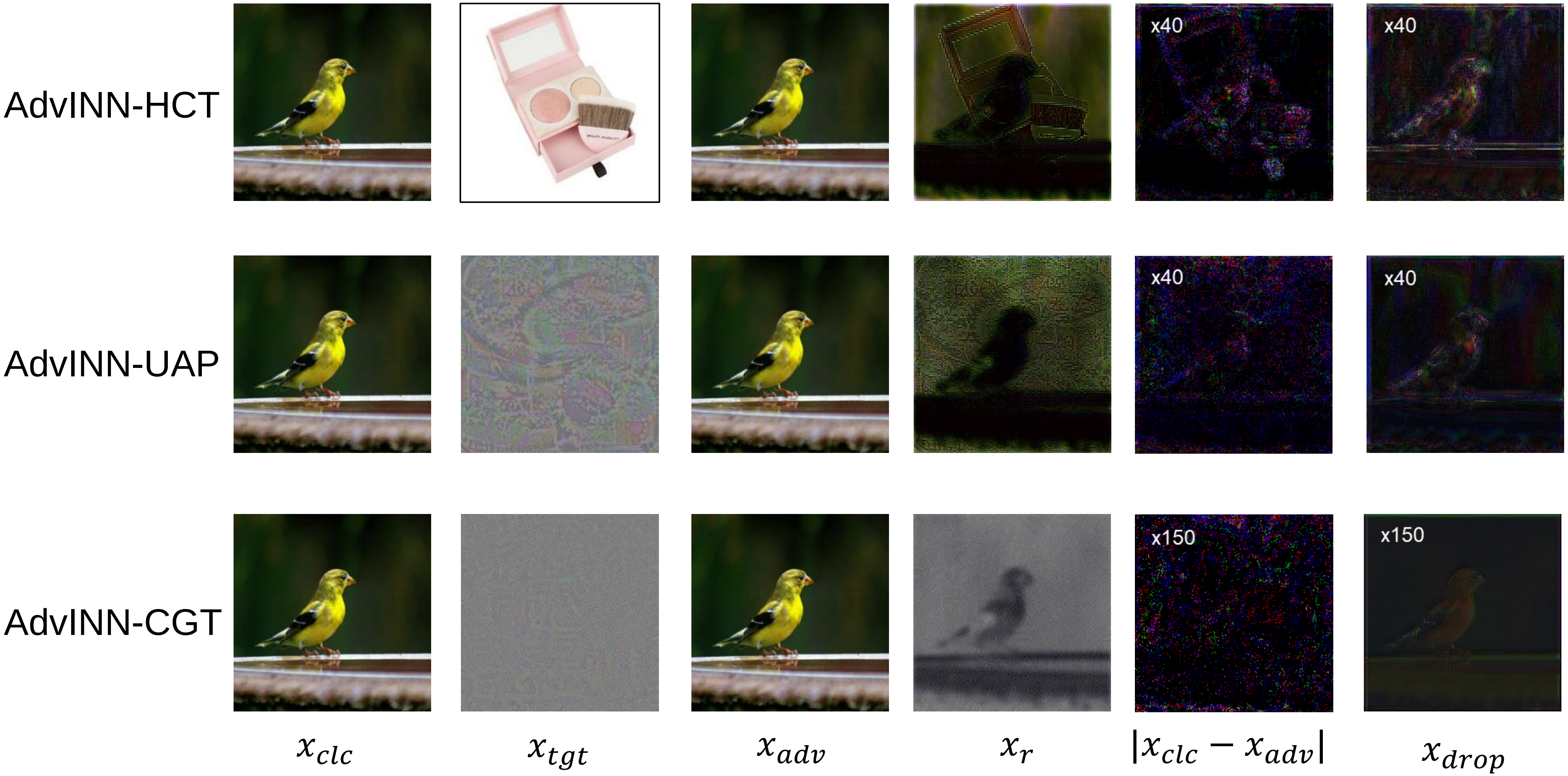


(a) JPEG Compression (b) Bit-Depth Reduction (c) NRP

Naseer, Muzammal, et al. "A self-supervised approach for adversarial robustness." CVPR 2020.

3. Experimental Results

Visualizaiton and Interpretation



3. Experimental Results

Ablation Studies

Adversarial Budget Constraints

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

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

Different Classifiers

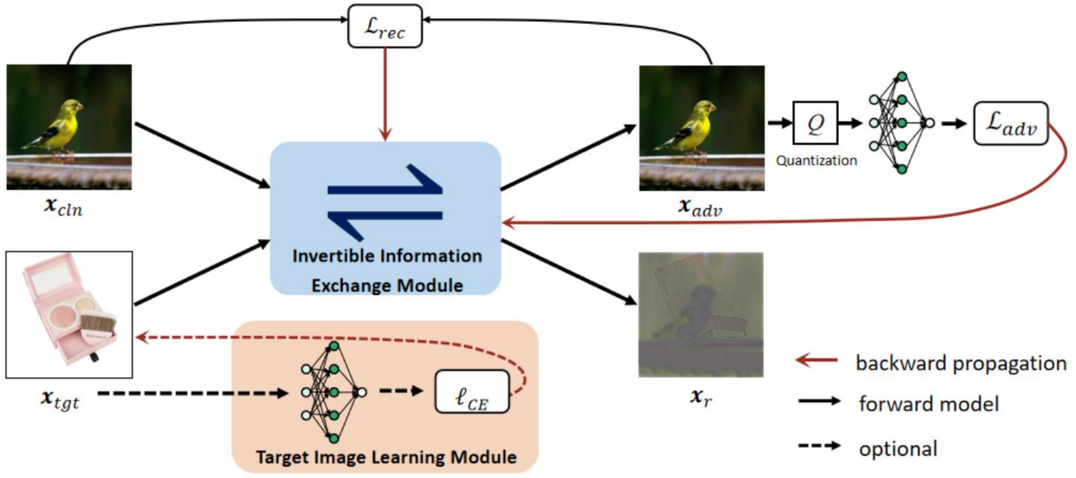
Table 4: The performance of AdvINN on different classifiers. The adversarial weights λ_{adv} are set to 10 and 3 on Inception_v3 and Densenet121, respectively.

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

4. Conclusions

AdvINN: Adversarial Attack via **I**nvertible **N**eural **N**etworks

- Generate imperceptible and robust adversarial examples by simultaneously adding and dropping information in an unified framework
- Fully utilize the information preservation property of INNs
- Improve the interpretability of adversarial examples



国防科技大学
National University of Defense Technology

Thanks for watching!

Website

Paper

Code

Email: chenzihan21@nudt.edu.cn