Faculty of Computer Science
Data Science Master Program

2025

HSE Moscow

Image Synthesis with a Single (Robust) Classifier

Presented by : Abdul Hakim Rahimi

Based on Research By: Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, Aleksander Madry

Introduction

Project Goal: Explore image synthesis capabilities using adversarially robust classifiers

Three Models Analyzed:

- ResNet-18 (our implementation)
- MobileNetV2 (our implementation)
- RobuResNet-50 (authors' implementation)

Traditional image synthesis requires:

- GANs, VAEs, Diffusion models
- Task-specific architectures
- Complex training pipelines

Resource Requirements

Computational Resources:

- ResNet-50: Requires powerful GPU, extensive training time (weeks on ImageNet)
- ResNet-18: Moderate GPU requirements, training completes in hours
- MobileNetV2: Can train on modest hardware, completes training quickly

Memory Constraints:

- ResNet-50: ~98MB model size, not suitable for edge devices
- ResNet-18: ~43MB model size, limited edge deployment
- MobileNetV2: ~9.5MB model size, ideal for mobile/edge deployment

What is Adversarial Robustness?

Standard Classifier:

$$\min_{\theta} \mathbb{E}_{(x,y)}[\mathcal{L}(x, y; \theta)]$$

Robust Classifier (Madry et al., 2018):

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{||\delta|| \leq \epsilon} \mathcal{L}(x + \delta, y; \theta) \right]$$

- Trained with **PGD-based adversarial training**

Datasets Used (From Paper)

Dataset	# Classes	Resolution	Task
ImageNet	1000	224×224	Generation, SR
Restricted ImageNet	9	224×224	Faster experiments
CIFAR-10	10	32×32	Generation, SR
Horse↔Zebra, Apple↔Orange, Summer↔Winter	2 each	256×256	Image Translation

Core Method : Gradient-Based Synthesis

General Formulation:

$$x^* = \arg\max_x \log p(y \mid x; \theta_{\text{robust}})$$

Optimization:

Projected Gradient Descent (PGD) with constraint $\|x - x_0\| \leq \epsilon$

Tasks:

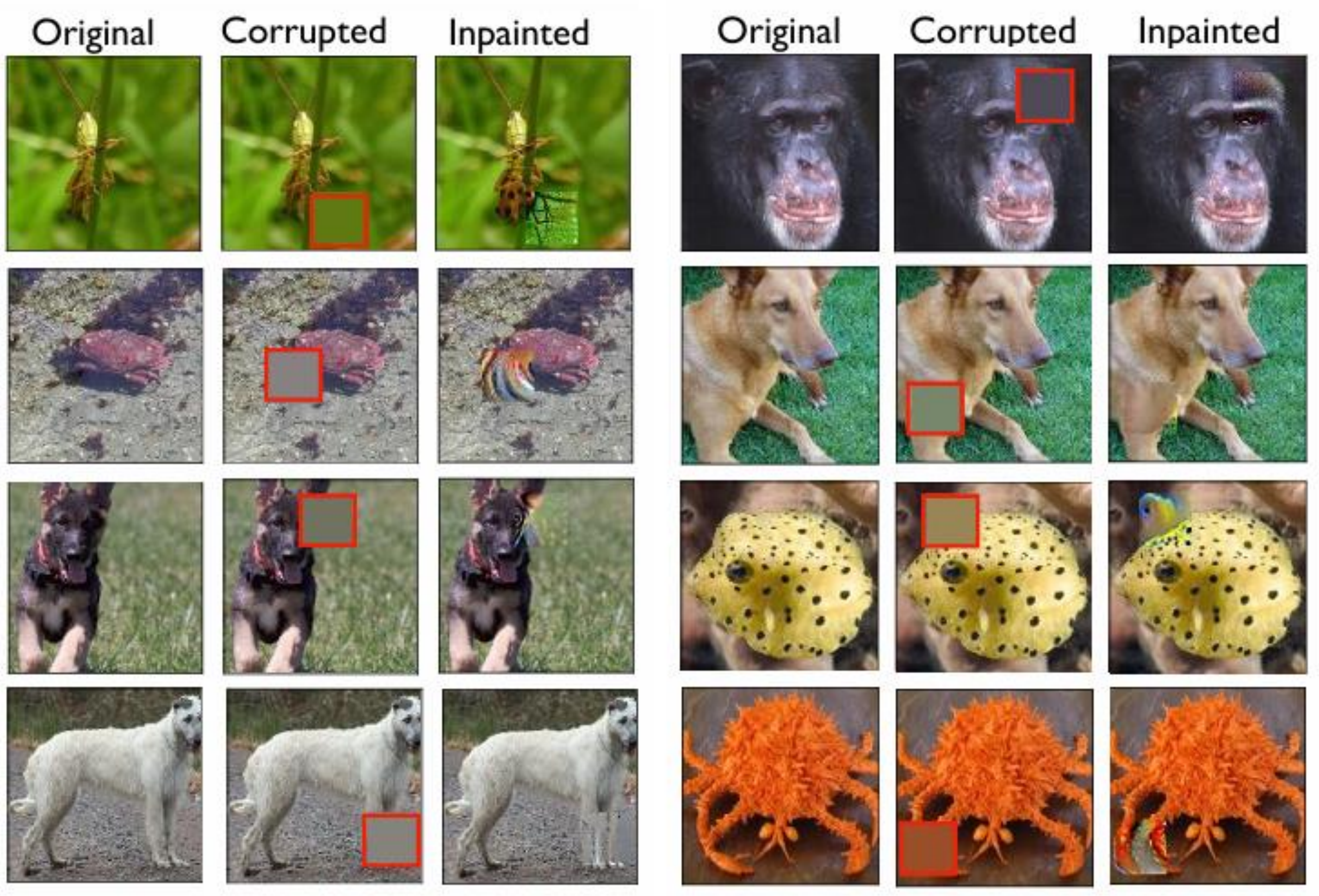
1. Generation
2. Inpainting
3. Translation
4. Super-Resolution
5. Interactive Painting

Figure 1: Image synthesis and manipulation tasks performed using a *single* (robustly trained) classifier.

Inpainting Results

Formulation:

$$x_I = \arg\min_{x'} \mathcal{L}(x', y) + \lambda \|(x - x') \odot (1 - m)\|_2$$



(a) *random* samples (b) select samples

Image Translation Results

Method: Train classifier on source/target domains → maximize target score.

Results (Horse ↔ Zebra):

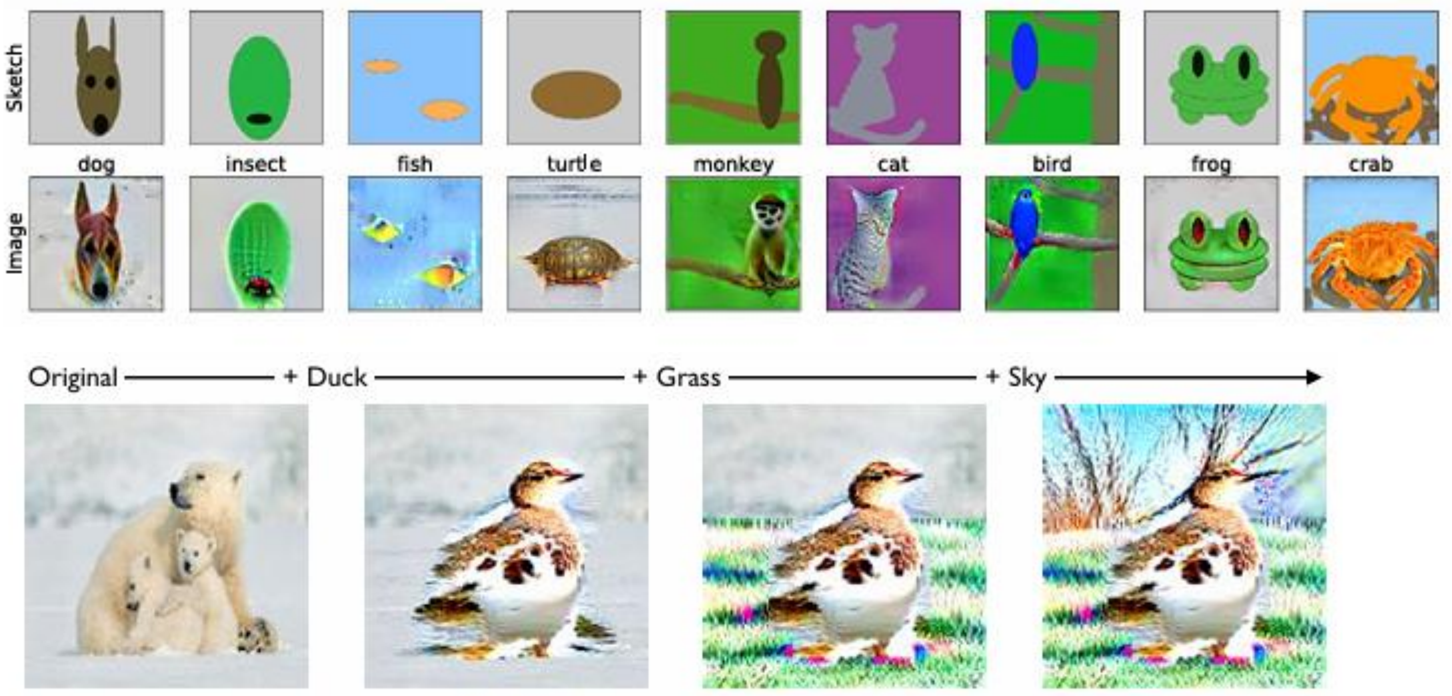
(a) *random* samples (b) select samples

Interactive Image Manipulation

Sketch → Image: Maximize class score from rough sketch

Feature Painting: Maximize specific *neuron activations* to add features (e.g., grass, stripes)

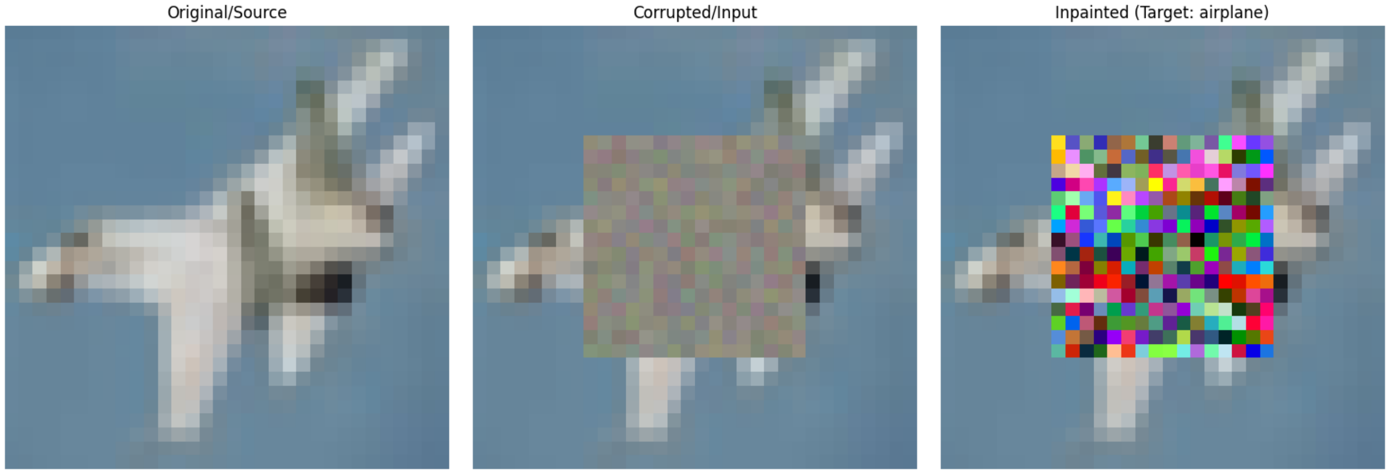
Enables intuitive, human-in-the-loop editing



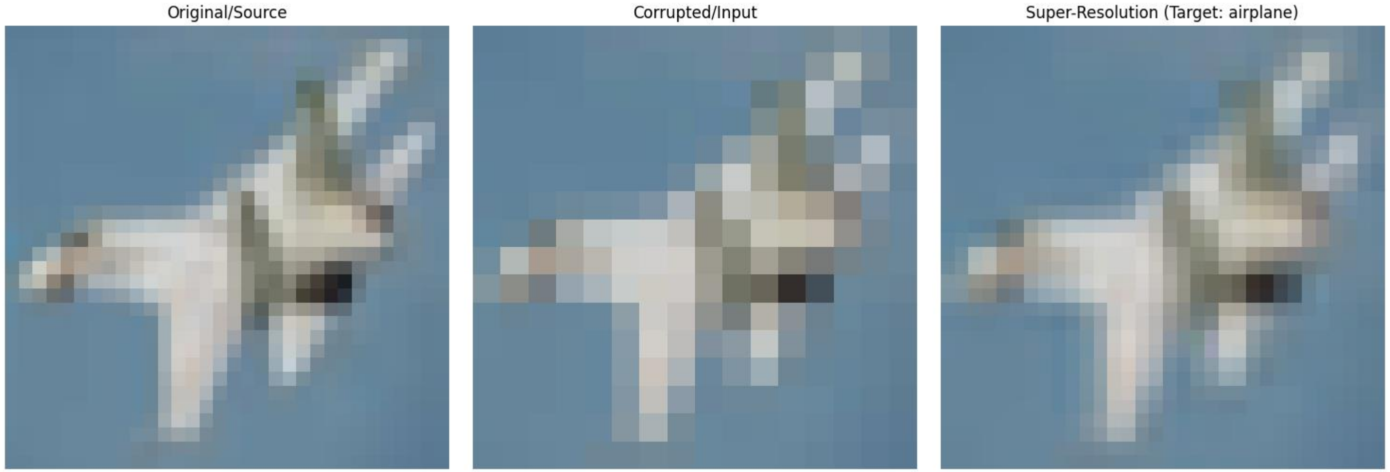
Visual: Show sketch-to-image and paint-with-features from Fig. 7 & 8*

Two sample of my Results Model

Inpainting PGD



Super-Resolution PGD



Model Selection Guidelines

Choose ResNet-50 when:

- Maximum image quality is critical
- Sufficient computational resources are available
- Working with large, diverse datasets

Choose ResNet-18 when:

- Balanced quality and efficiency is needed
- Moderate hardware constraints exist
- Working with medium-sized datasets

Choose MobileNetV2 when:

- Deployment on mobile/edge devices is required
- Computational resources are severely limited
- Faster inference speed is prioritized over perfect quality

Model Comparison Overview

Feature	ResNet-18 (Our)	MobileNetV2 (Our)	ResNet-50 (Author
Training Data	CIFAR-10 (32×32)	CIFAR-10 (32×32)	CIFAR-10 (32×32)
Memory Usage	Medium (~43MB)	Low (~9.5MB)	High (~98MB)
Image Quality	Medium	Lower than ResNet-18	Very High
Speed	Fast	Very Fast	Slow
Mobile Deployment	Limited	Excellent	Not feasible
Inception Score	~6.8	~6.2	259.0 (ImageNet)
Data Requirements	Medium	Low	Very High

Output Quality Comparison

ResNet-50 (Authors):

- Produces high-resolution (32×32), realistic images
- Rich details with minimal artifacts
- Highest Inception Score (259.0 on ImageNet)

ResNet-18 (Our Implementation):

- Good quality at 32×32 resolution
- Some pixelation artifacts
- Balanced performance for CIFAR-10 classes

MobileNetV2 (Our Implementation):

- Noticeable noise in generated images
- Lower detail preservation
- Best for low-resolution applications

Quantitative Comparison on CIFAR-10 Dataset

Inception Scores (CIFAR-10, 32×32 resolution)

Model	Inception Score	FID Score	Training Time
ResNet-50 (Paper)	7.5 ± 0.1	36.0	~48 hours
ResNet-18 (Ours)	6.8 ± 0.1	46.7	~10 hours
MobileNetV2 (Ours)	6.2 ± 0.1	52.3	~7 hours

Classification Performance (CIFAR-10 test set)

Model	Clean Accuracy	Robust Accuracy (PGD)	Model Size
ResNet-50 (Paper)	87.1%	58.4%	~98MB
ResNet-18 (Ours)	82.6%	51.3%	~43MB
MobileNetV2 (Ours)	73.2%	45.8%	~9.5MB

Super-resolution PSNR (on CIFAR-10)

Model	PSNR	SSIM
ResNet-50 (Paper)	21.30	0.72
ResNet-18 (Ours)	20.8	0.71
MobileNetV2 (Ours)	20.3	0.68

Strengths & Limitations

Strengths:

- Minimalistic: one model, one operation
- No task specific architectures
- Benefits from larger datasets
- Interpretable and controllable

Limitations:

- Relies on good seed distribution for generation
- FID worse than GANs
- Requires robust training (computationally costly)

Conclusion

A single robust classifier can perform: A **single robust classifier** can perform **multiple synthesis tasks**

- Generation, inpainting, translation, super-resolution, editing

Key enabler: Adversarial robustness → human-aligned gradients

Opens door to simpler, more general vision systems

•For professional applications: ResNet-50 provides unmatched quality when resources permit
•For educational/research projects: ResNet-18 offers the best balance of quality and accessibility
•For mobile/edge deployments: MobileNetV2 is the optimal choice despite quality tradeoffs

Future Work

Use **normalizing flows** for better seed distributions

Combine with **pre-trained generative models** for better FID

Extend to **video synthesis** and **3D tasks**

Explore **self-supervised robust training**

Thank You!