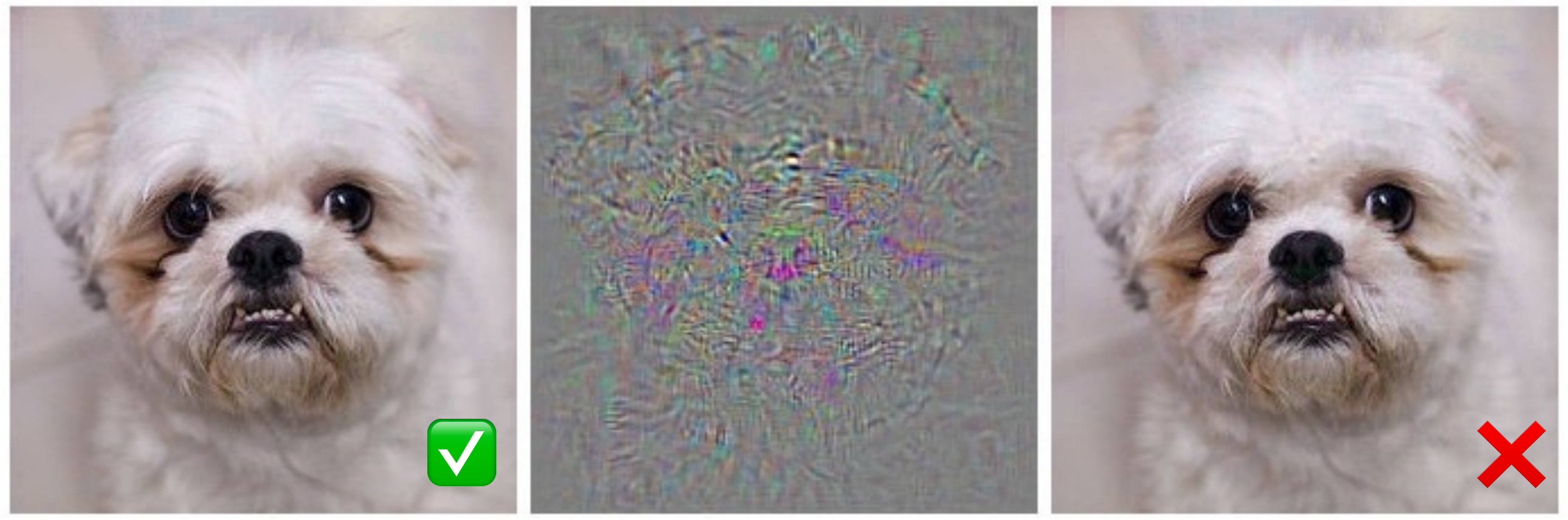
Diffusion-Based Adversarial Sample Generation for Improved Stealthiness and Controllability

Haotian Xue[1], Alexandre Araujo[2], Bin Hu[3], Yongxin Chen[1]

[1] GaTech, [2] NYU, [3] UIUC

Background & Motivation

It is easy to fool a DNN by crafting adversarial perturbations:

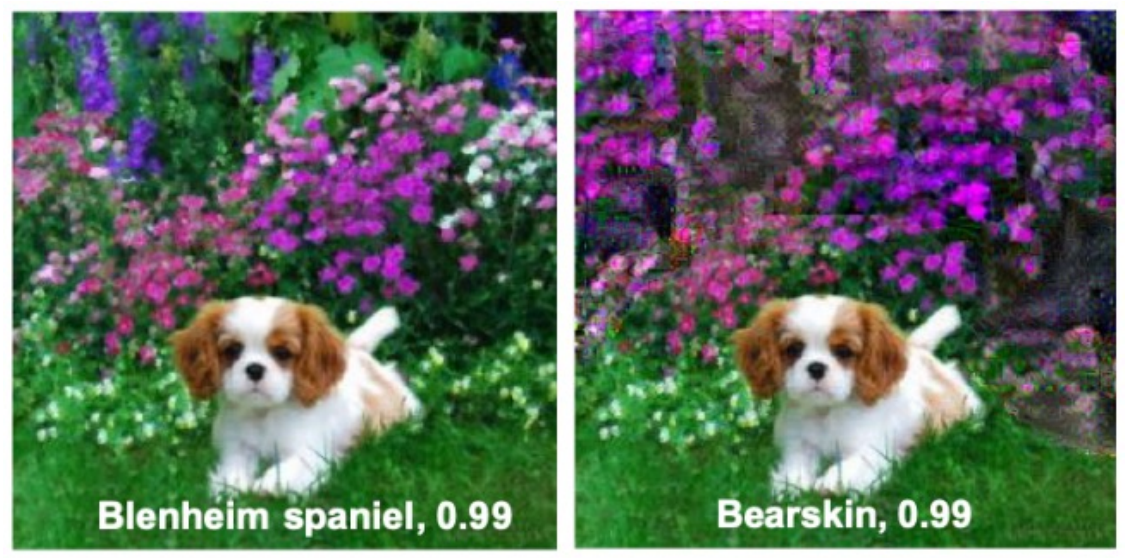


Prediction: Dog   Prediction: Cat

But **Scaling-Up** the Perturbation brings **Unrealism:**

(a) PGD Attack with larger Budget: $\epsilon_\infty = 32/255$

(b) Adversarial Patch

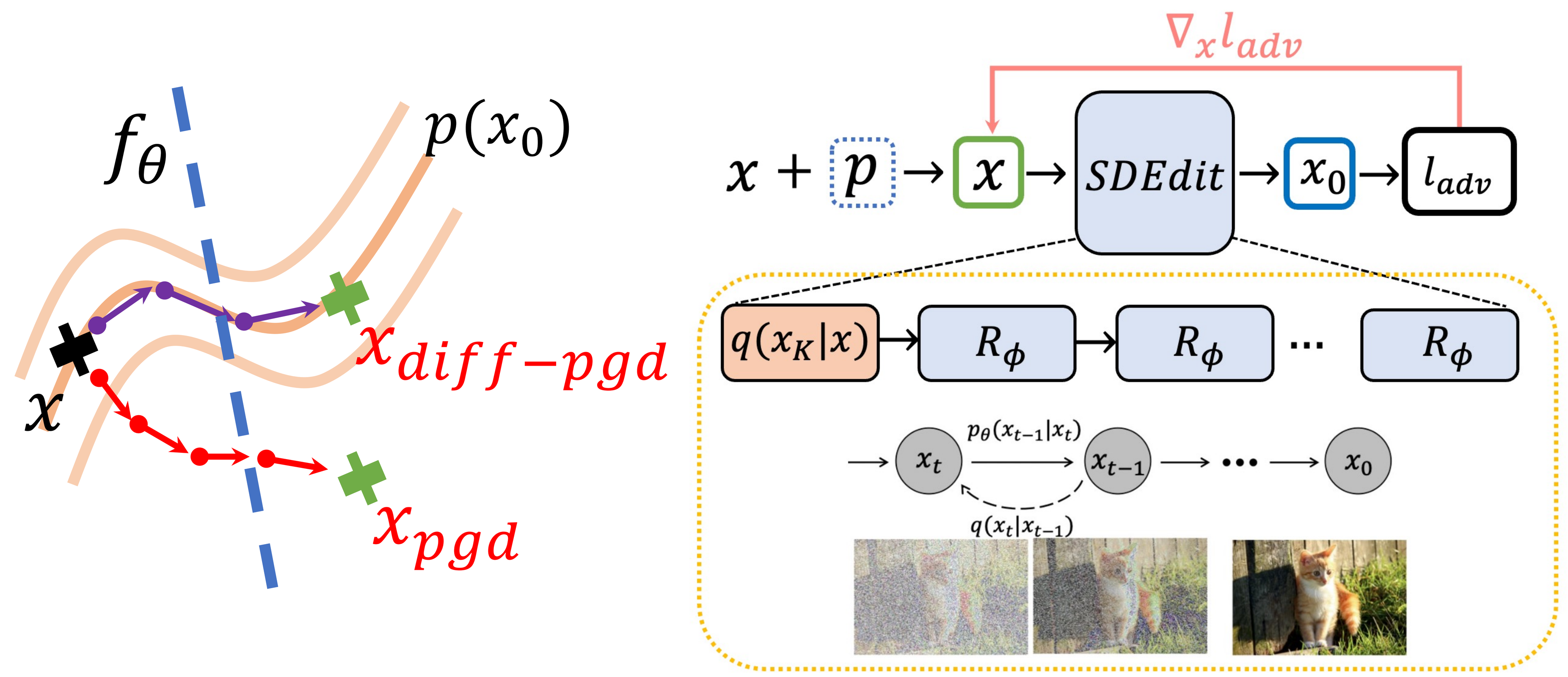


(c) Style-based Attacks

How to make the adversarial attacks **scalable**, and preserving the **stealthiness**?

Our Approach: Diff-PGD

Key Idea: using Diffusion Model off-the-shelf to preserve the stealthiness of generated adversarial samples



$$x_0^t = \mathrm{SDEdit}(x^t, K) \quad \text{and} \quad x^{t+1} = \mathcal{P}_{B_\infty(x,\epsilon)}\left[x^t + \eta\, \mathrm{sign} \nabla_{x^t} l(f_\theta(x_0^t), y)\right].$$

Key Strengths:

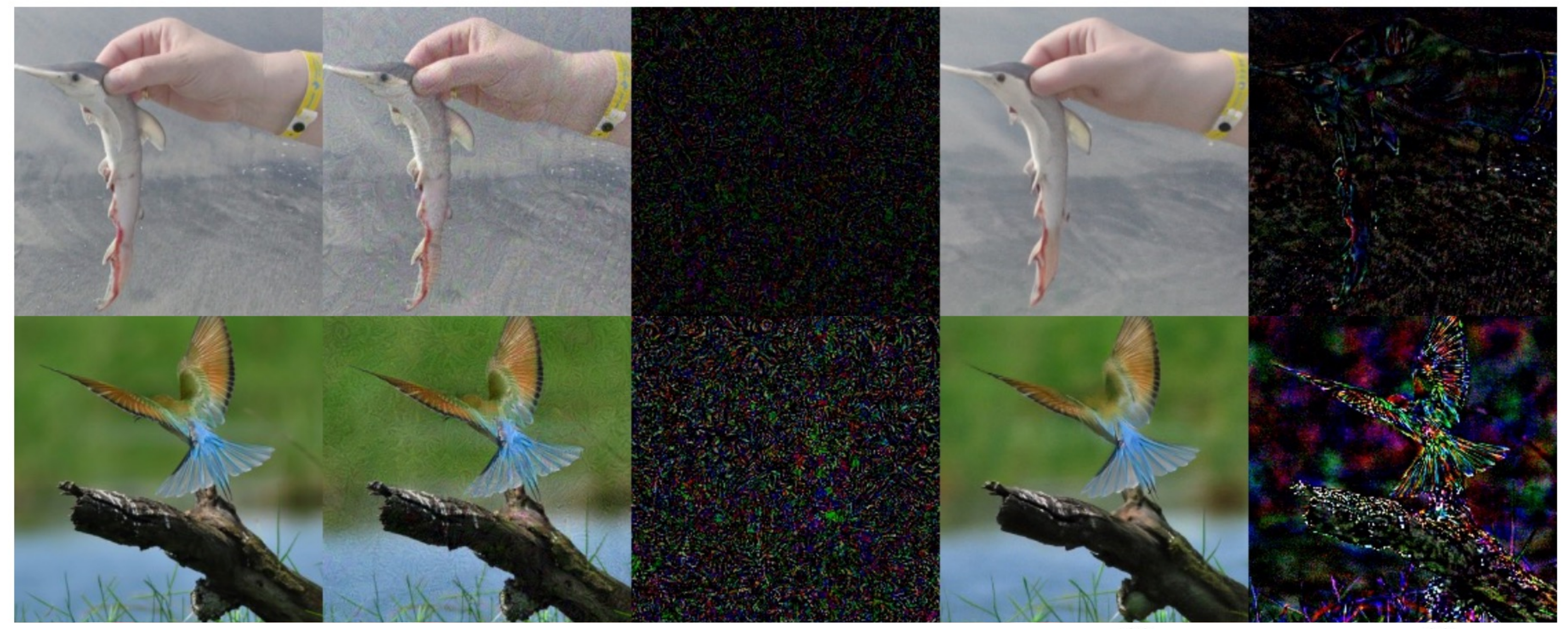
- Plug-and-play, fully off-the-shelf
- Can generate samples with higher stealthiness
- Can be easily applied to many attacks (e.g. digital attacks, style-based attacks, physical-world attacks)

Effectiveness analysis:

- Can still achieve 100% success rate with 5 PGD steps
- Cost more, but can be largely optimized using approximated gradient
- Higher stealthiness, proved using human evaluations
- Stronger anti-purification and transferability

Main Results

(a) Digital Attacks



x x_{PGD} δ_{PGD} x_0^n δ_0^n

(b) Style-based Attacks

x mask   style x_{adv}

(c) Physical-world Attack

Takeaway: Diff-PGD makes it possible to flexibly scale-up adversarial perturbations to preserve the stealthiness !

- More results can be found in our paper and GitHub repo 👉
- Feel free to contact me if you have further questions 👉