Adversarial Robustification via Text-to-Image Diffusion Models

Daewon Choi[A*] Jongheon Jeong[B*] Huiwon Jang[A] Jinwoo Shin[A]

[A] Korea Advanced Institute of Science and Technology (KAIST)

[B] Korea University

ECCV 2024

* Equal contribution

Adversarial Robustness

Definition: Make consistent prediction under every possible perturbation

$$f(\mathbf{x}) = f(\mathbf{x} + \boldsymbol{\delta}), \quad \forall \boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq \varepsilon$$

Classifier

Perturbation

classified as

Stop Sign

+

Perturbation

=

classified as

Max Speed 100

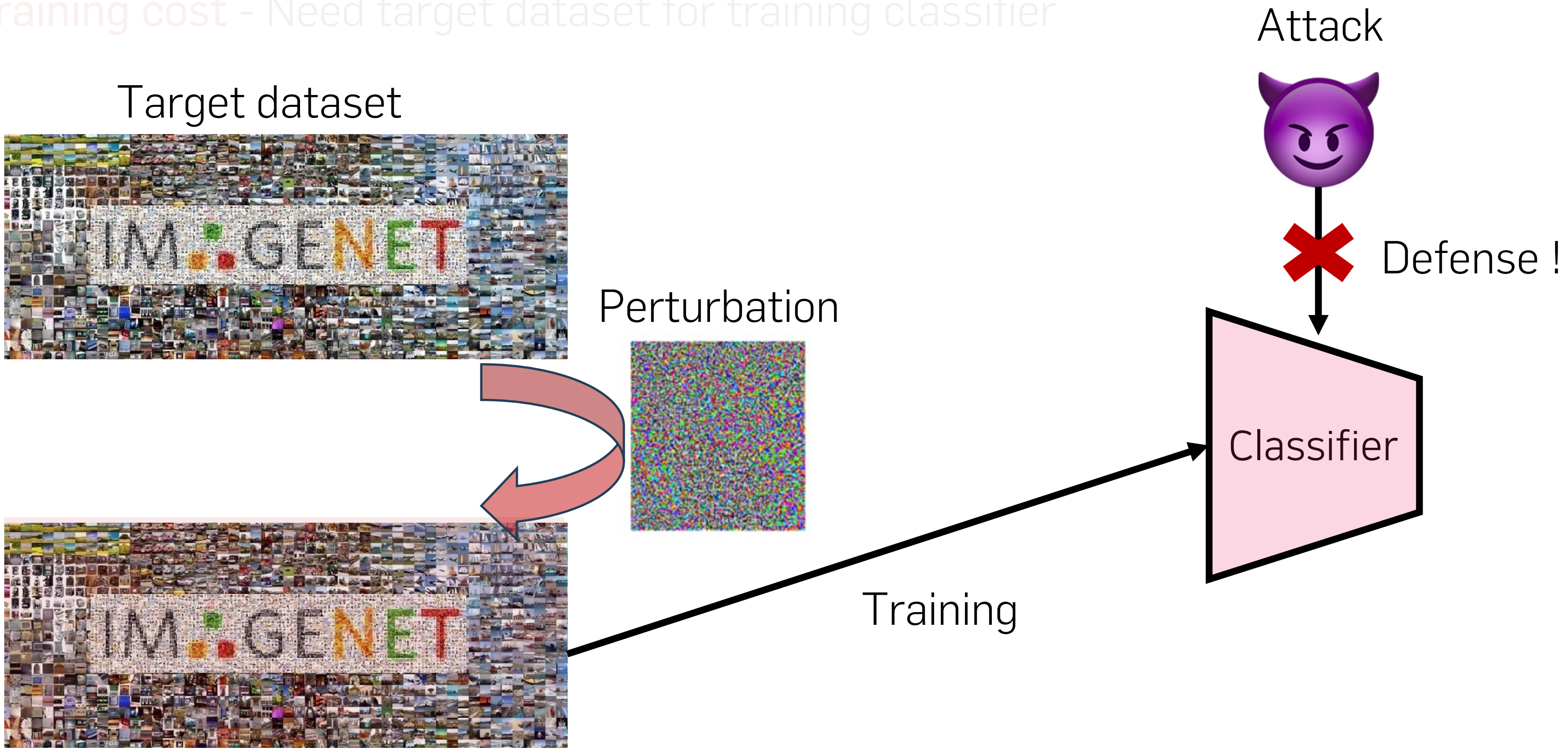
Source: https://www.altacognita.com/robust-attribution (left), GPT-4o (right)

Adversarial Training [Madry et al., 2018]

Adversarial Training : Training via perturbated examples (i.e., adversarial example)

(-) Empirical robustness – Only empirically ensure robustness to specific adversarial attack

(-) High training cost - Need target dataset for training classifier



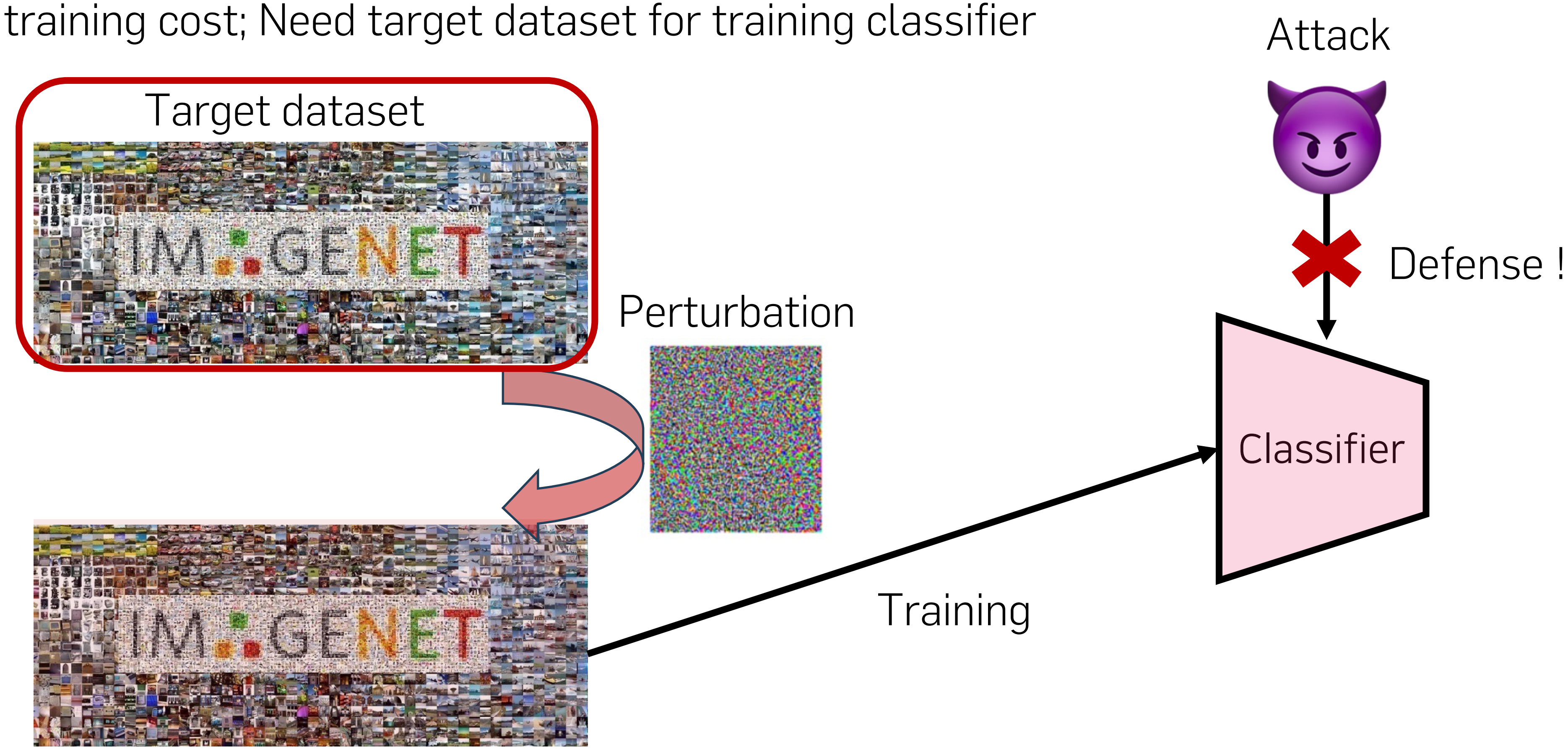
[Madry et al., 2018] Towards Deep Learning Models Resistant to Adversarial Attacks. ICML 2018.

Adversarial Training [Madry et al., 2018]

Adversarial Training : Training via perturbated examples (i.e., adversarial example)

(-) Empirical robustness; Only empirically ensure robustness, not provable

(-) High training cost; Need target dataset for training classifier



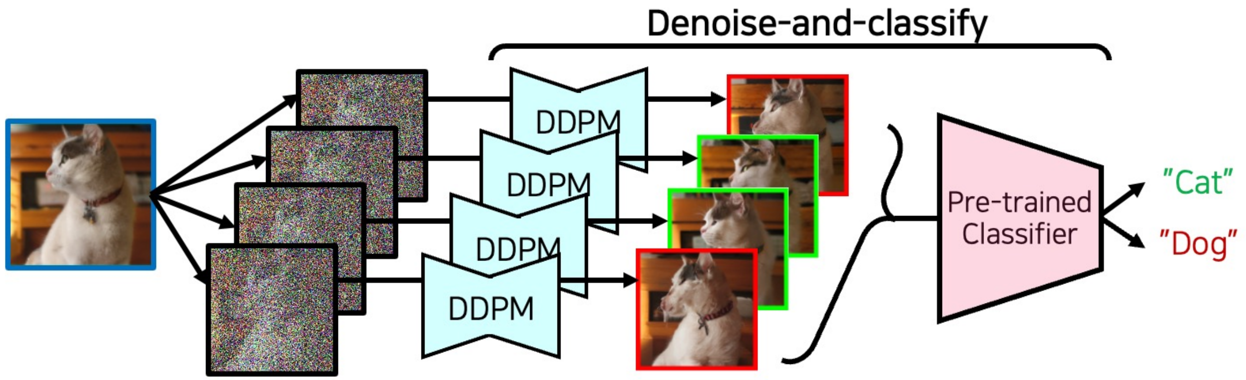
[Madry et al., 2018] Towards Deep Learning Models Resistant to Adversarial Attacks. ICML 2018.

Adversarial Defense for Pre-trained Models

Recent defense methods partially overcome crucial limitations of adversarial training.

Denoised Smoothing

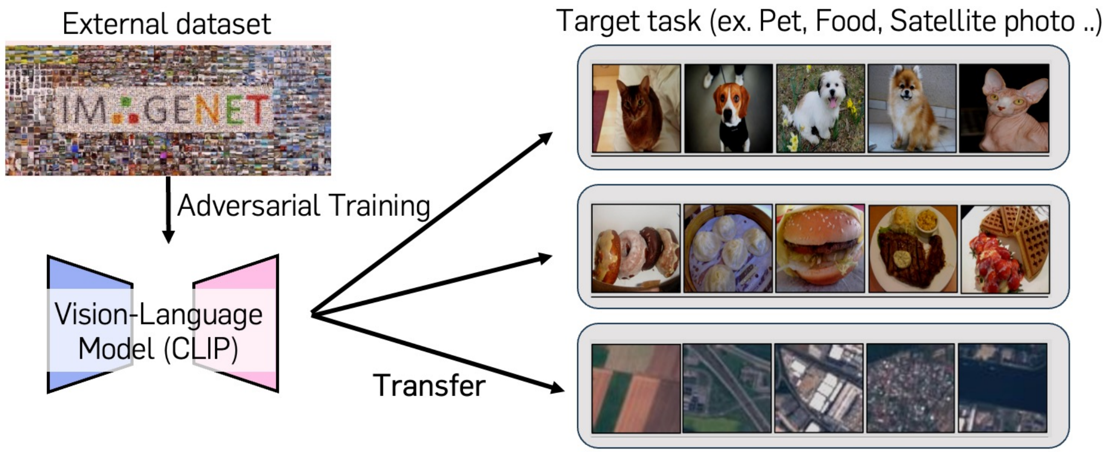
[Salman et al., 2020; Carlini et al., 2023]



(+) Provable robustness
(+) Not need to training classifier

Zero-shot Adversarial Robustness

[Mao et al., 2023]



(+) Not need target dataset

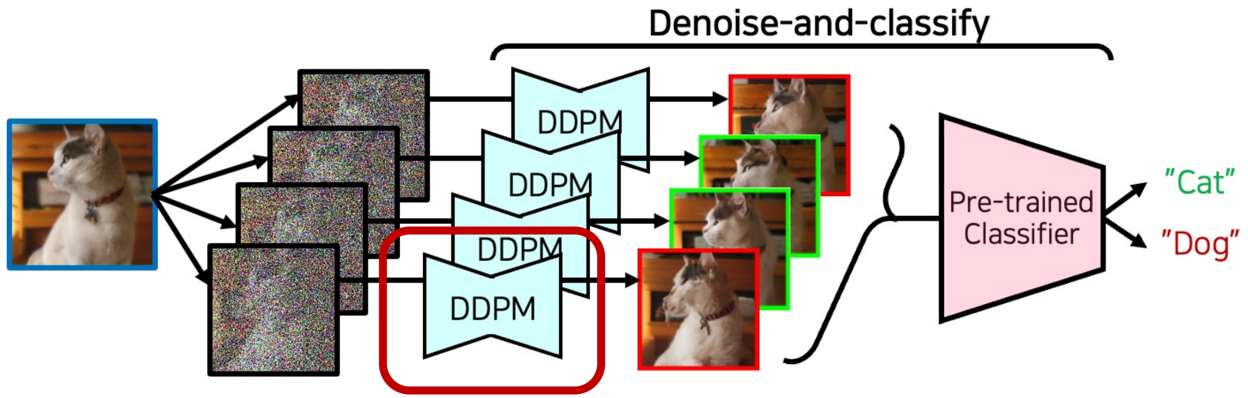
[Salman et al., 2020] Denoising Smoothing: A Provable Defense for Pretrained Classifiers, NeurIPS 2020.
[Carlini et al., 2023] (Certified!!) Adversarial Robustness for Free!, ICLR 2023.
[Mao et al., 2023] Understanding Zero-shot Adversarial Robustness for Large-scale Models, ICLR 2023.

Adversarial Defense for Pre-trained Models

However, they still require plenty of training data for robustification..😢

Denoised Smoothing

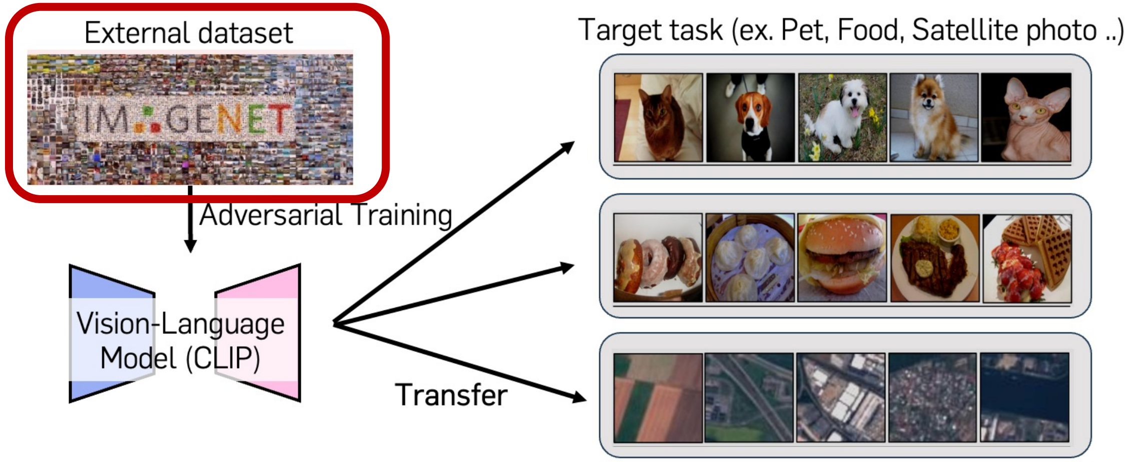
[Salman et al., 2020; Carlini et al., 2023]



(-) Need target dataset for separate training denoiser

Zero-shot Adversarial Robustness

[Mao et al., 2023]



(-) Still need external dataset for obtaining robustness

[Salman et al., 2020] Denoising Smoothing: A Provable Defense for Pretrained Classifiers, NeurIPS 2020.
[Carlini et al., 2023] (Certified!!) Adversarial Robustness for Free!, ICLR 2023.
[Mao et al., 2023] Understanding Zero-shot Adversarial Robustness for Large-scale Models, ICLR 2023.

Research Question

However, they require plenty of training data for robustification..

Existing approaches require plenty of training data for robustification.

Denoised Smoothing

(-) Need target dataset for separate training denoiser

Zero-shot Adversarial Robustness

(-) Still need external dataset for obtaining robustness

[Salman et al., 2020] Denoising Smoothing: A Provable Defense for Pretrained Classifiers, NeurIPS 2020.
[Carlini et al., 2023] (Certified!!) Adversarial Robustness for Free!, ICLR 2023.
[Mao et al., 2023] Understanding Zero-shot Adversarial Robustness for Large-scale Models, ICLR 2023.

Research Question

Existing approaches require plenty of training data for robustification.

🧐 Can we robustify a classifier without using external data?

Motivation

🧐 Can we robustify a classifier without using any external data?

Motivation: Text to image diffusion models (T2I) is versatile tool for promising solution

Generation

Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

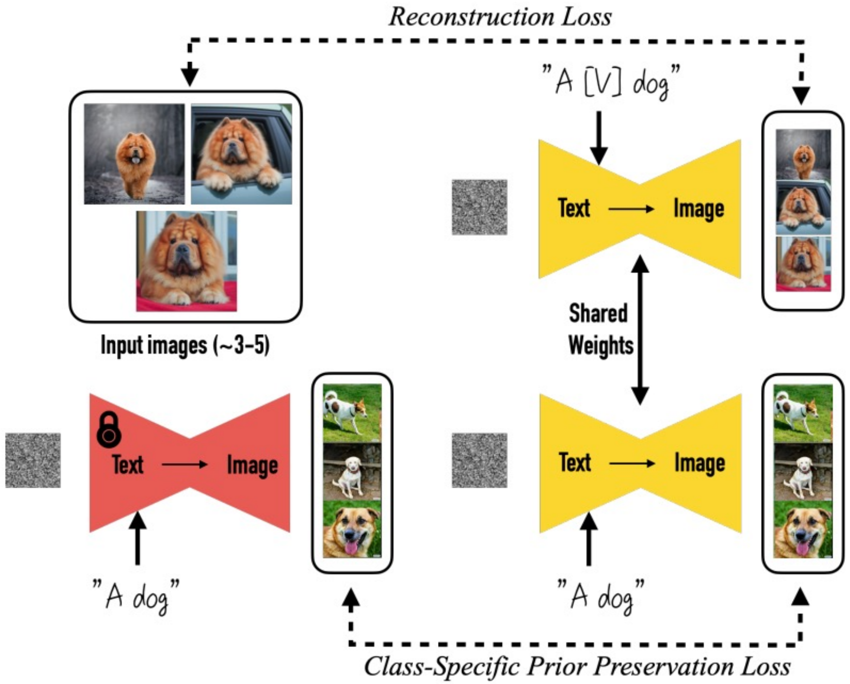
A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

Teddy bears swimming at the Olympics 400m Butter-fly event.

A cute corgi lives in a house made out of sushi.

A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

Personalization



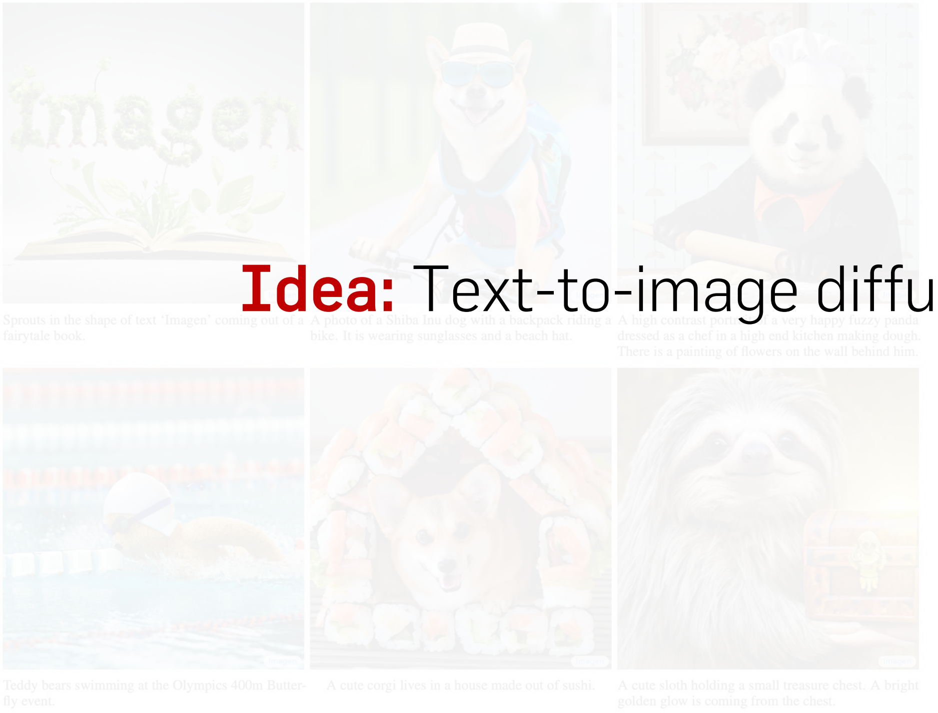
Source: https://imagen.research.google (left), https://dreambooth.github.io (right)

Motivation

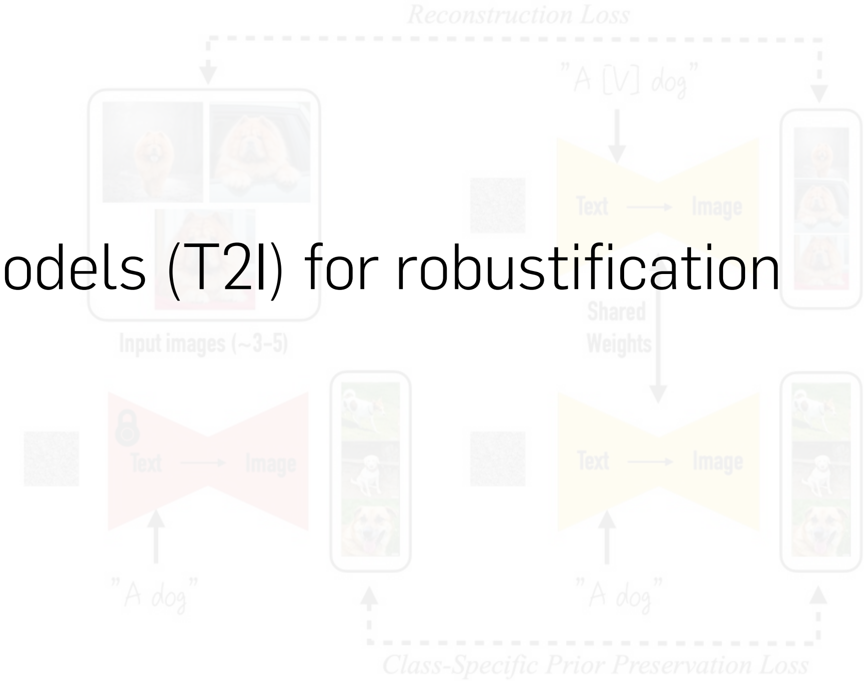
🧐 Can we robustify a classifier without using any external data?

Motivation: Text to image diffusion models (T2I) is versatile tool for promising solution

Generate high quality image



Personalization



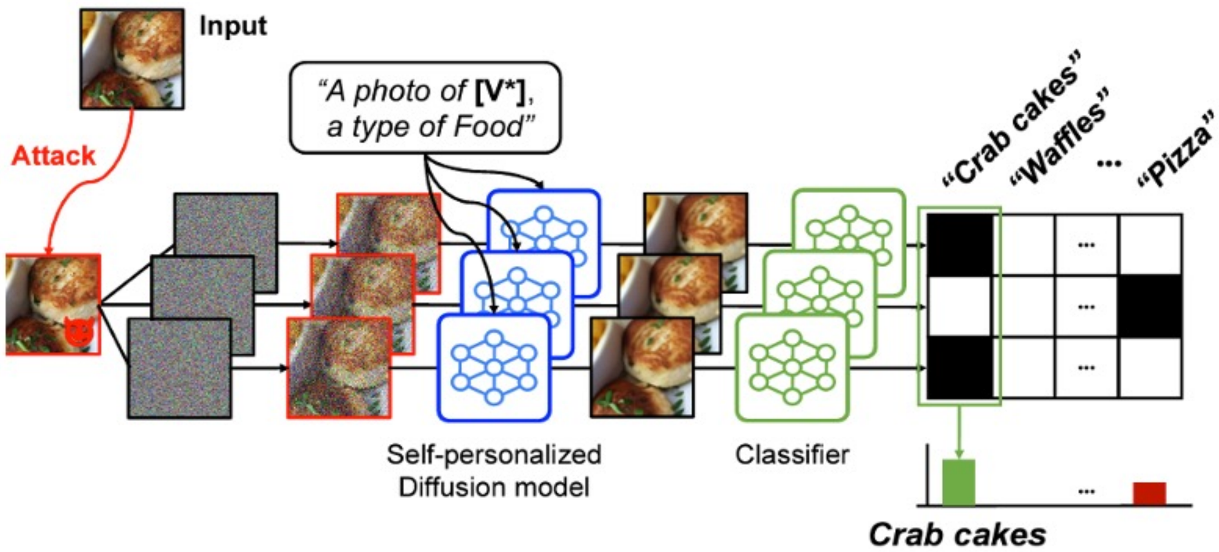
Idea: Text-to-image diffusion models (T2I) for robustification

Goal: Text-to-Image Diffusion Models for Robustification

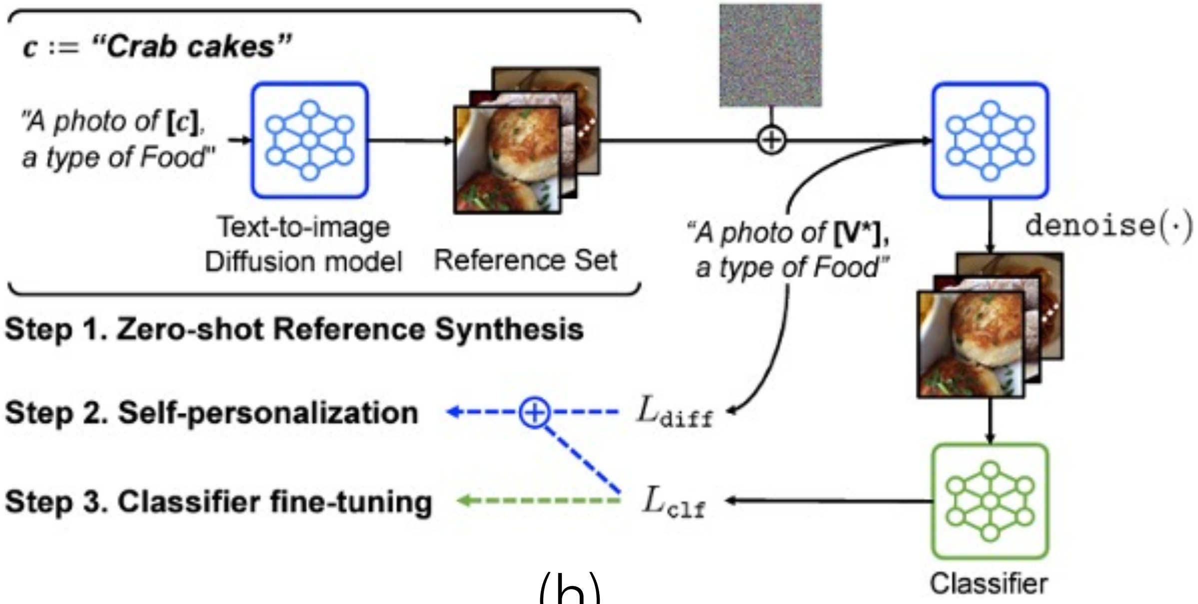
We utilize text-to-image diffusion models (T2I) mainly in two different ways !

(a) Incorporate T2I into the denoised smoothing pipeline

(b) Personalize T2I on target tasks using few-synthetic data



(a)

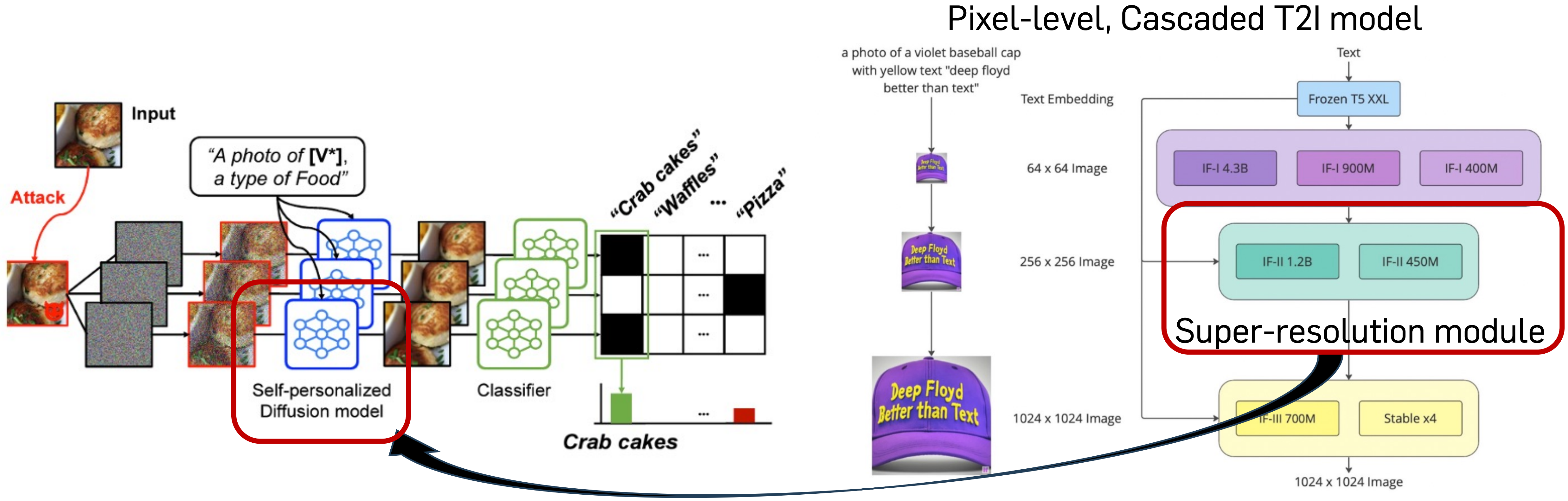


(b)

(a) Incorporate T2I into the Denoised Smoothing Pipeline

Use super-resolution module in pixel-level, cascaded T2I model as a denoiser

✓ The module is biased to reconstruct the original image contents

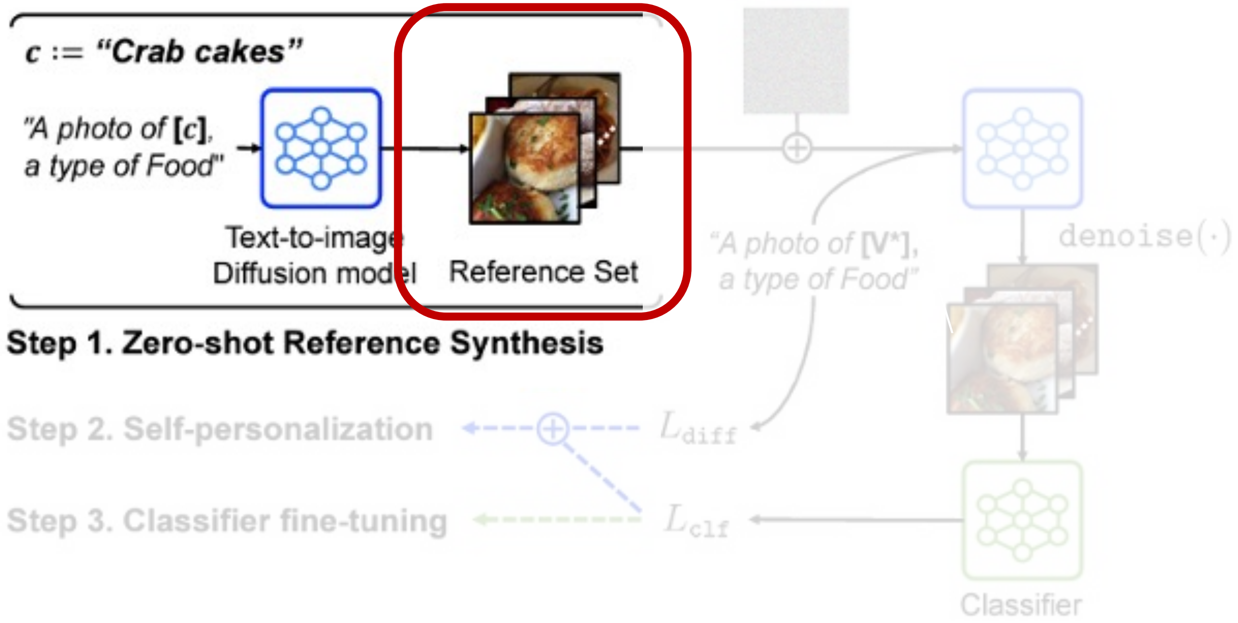


[Saharia et al., 2022] Photorealistic text-to-image diffusion models with deep language understanding, NeurIPS 2022.
DeepFloyd-IF: https://stability.ai/news/deepfloyd-if-text-to-image-model

(b) Personalize T2I on Target Tasks Using Synthetic Data

Step1. Synthesize a few reference samples via T2I, given textual label

✓ Given only textual labels for the target task, high-quality images are generated

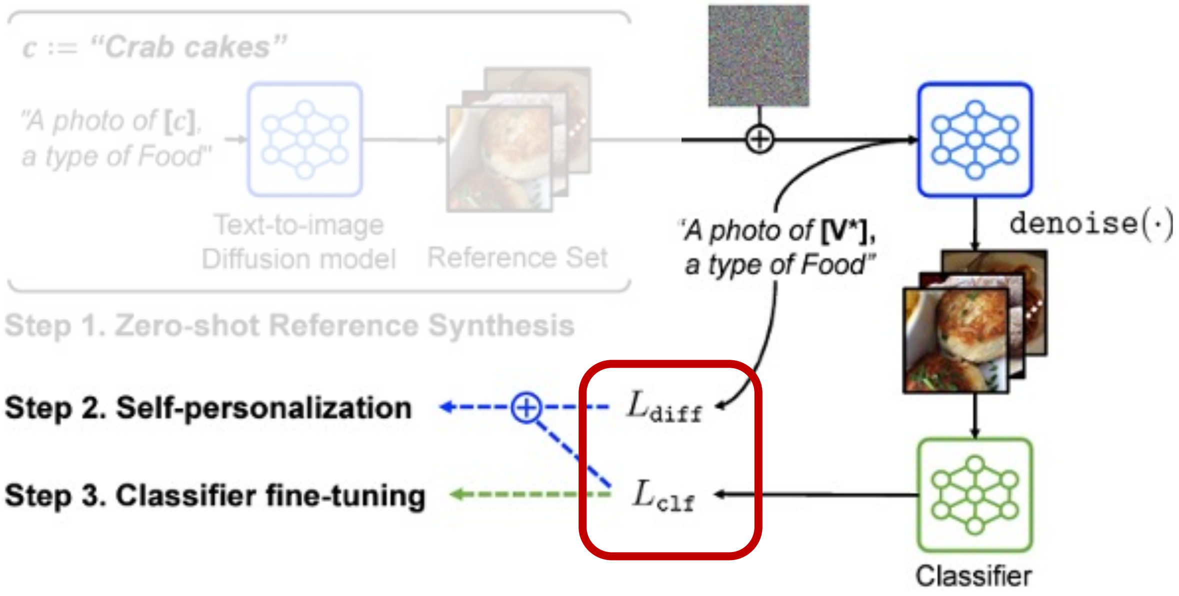


(b) Personalize T2I on Target Tasks Using Synthetic Data

Step 2&3. By leveraging synthetic samples, fine-tuning both the T2I as well as classifier

✓ DreamBooth enables to personalize T2I with few-reference images

✓ Classifier-guided regularization makes personalization suitable for denoised smoothing



- Dream-Booth Objective [Ruiz et al., 2023]

$$L_{\texttt{diff}}(\theta) := \mathbb{E}_{x^g,\varepsilon,t}\left[||\varepsilon - \varepsilon_\theta(x_t^g, t, \tau_\theta(\texttt{C}(\text{“}sks\text{”}))|x_t^g, kt)||_2^2\right]$$

- Classifier-Guided Regularization

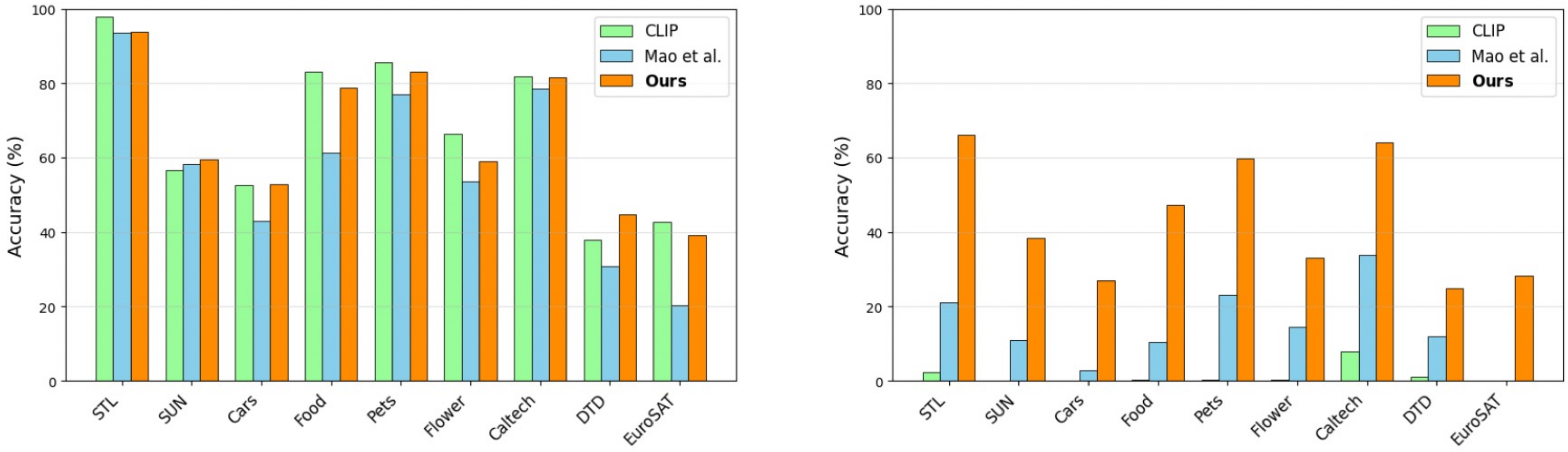
$$L_{\texttt{clf}}(\theta,\psi) := \mathbb{E}_{(x^g,c)\sim D^g,t}\left[\mathbb{CE}(f_\psi(\tilde{x}^g), c)\right]$$

[Ruiz et al., 2023] Dream-Booth: Fine tuning text-to-image diffusion models for subject-driven generation, CVPR 2023.

Experiment: Robustification of CLIP

Our framework significantly enhance CLIP with a new robustness-accuracy frontier

✓ Outperforming zero-shot robustness method [Mao et al., 2023] on 8 zero-shot benchmarks



(a) Clean accuracy

(b) Robust accuracy

[Mao et al., 2023] Understanding Zero-shot Adversarial Robustness for Large-scale Models, ICLR 2023.

Experiment: Robustification of CLIP

Our framework significantly enhance CLIP with a new robustness-accuracy frontier

✓ Competitive and even surpassing other approaches directly accessing training data

Method	Data-free?	Robust accuracy (%) $\varepsilon = 0.5$	Robust accuracy (%) $\varepsilon = 1.0$	Clean accuracy (%) $\varepsilon = 0.5$	Clean accuracy (%) $\varepsilon = 1.0$
CLIP	✓	1.4	0.2	**58.2**	**58.2**
CLIP-Smooth	✓	16.8 (9.8)	2.2 (1.2)	45.2 (25.0)	35.2 (3.8)
Ours (w/o adapt)	✓	<u>40.0</u> (29.6)	31.0 (17.6)	56.2 (50.8)	55.2 (42.0)
Ours	✓	**42.6** (**34.2**)	<u>31.4</u> (**20.6**)	<u>57.6</u> (**53.4**)	<u>56.2</u> (**46.0**)
Mao et al. [38]	✗	26.0	12.3	51.2	47.2
Carlini et al. [7]	✗	38.6 (30.2)	**32.4** (19.8)	54.4 (49.8)	53.6 (44.2)

-> Access training data corresponding test data

[Mao et al., 2023] Understanding Zero-shot Adversarial Robustness for Large-scale Models, ICLR 2023.
[Carlini et al., 2023] (Certified!!) Adversarial Robustness for Free!, ICLR 2023.

Experiment: Robustification of Other Classifier

Our framework can also be effective in robustifying other generic vision classifiers

✓ Combining with ResNet-50, surpassing standard approaches accessing training data

		Robust accuracy (%)		Clean accuracy (%)	
Method	Data-free?	$\varepsilon = 0.5$	$\varepsilon = 1.0$	$\varepsilon = 0.5$	$\varepsilon = 1.0$
Standard Training	✗	5.2	1.0	**74.4**	**74.4**
+ **Ours (w/o adapt)**	✓	56.2 (47.0)	44.2 (27.4)	73.0 (67.0)	68.8 (57.2)
+ **Ours**	✓	**57.0** (**50.4**)	**47.8** (34.0)	70.4 (**68.2**)	71.8 (**60.8**)
Adversarial Training [37]	✗	51.0	46.8	55.0	55.0
Randomized Smoothing [13]	✗	55.2 (48.6)	43.8 (**37.0**)	65.4 (66.8)	55.4 (57.0)
Carlini et al. [7]	✗	56.2 (49.2)	45.2 (33.2)	72.6 (67.4)	70.0 (57.8)

-> Access training data corresponding test data

[Madry et al., 2018] Towards Deep Learning Models Resistant to Adversarial Attacks. ICML 2018.
[Cohen et al., 2019] Certified adversarial robustness via randomized smoothing. ICML 2019.

Summary

Pursuing adversarial robustness in practice has been viewed as a costly design decision.

- Existing techniques for adversarial robustness require a plenty of training data.

We introduce a new formulation of robustifying vision classifiers without external data.

- Incorporating T2I into the inference of a classifiers in novel ways.
- Applicable for any pre-trained classifiers, even when training data is limited.

Please drop by our poster session for more information

- Contact: daeone0920@kaist.ac.kr

Paper

Code