NatADiff: Adversarial Boundary Guidance for Natural Adversarial Diffusion

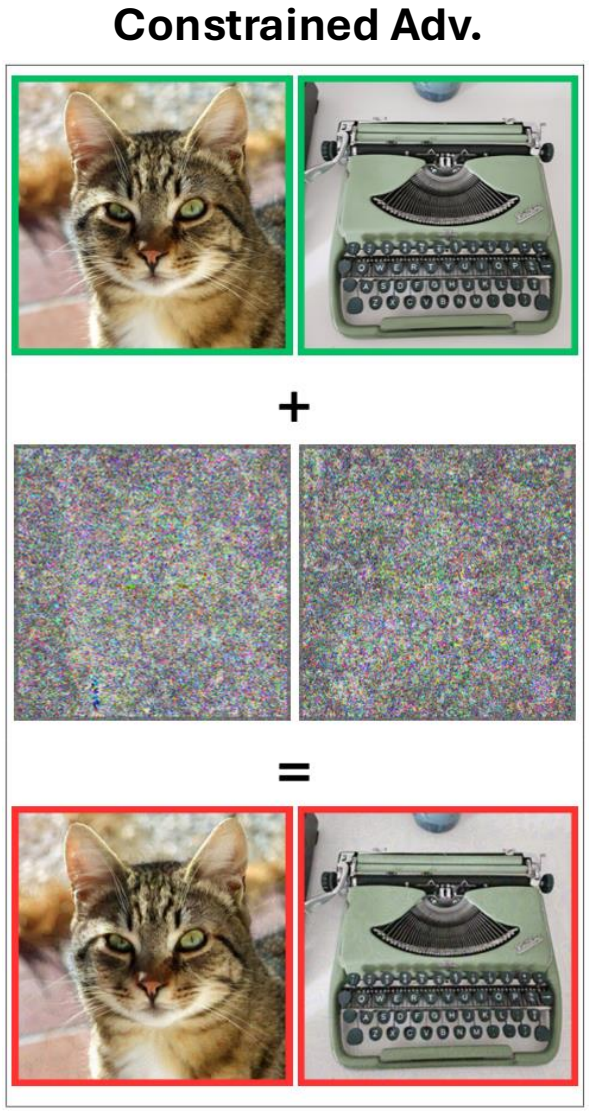
Max Collins, Jordan Vice, Tim French, and Ajmal Mian

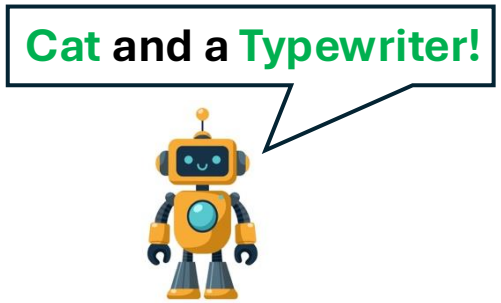
ICLR

THE UNIVERSITY OF WESTERN AUSTRALIA

Background: Adversarial Samples

- *Constrained adversarial attacks* "fool" classifiers by adding imperceptible image perturbations to a clean image.
  - These attacks are understood to leverage irregularities in the learned image manifold.
- Defences to these adversarial attacks have been proposed; however, they address the adversarial perturbation…
- Test-time errors can be seen as a form of *natural adversarial sample*.
  - *Current literature suggests that they occur when classifiers rely on "contextual cues" to shortcut classification.*
- **Perturbation-based defences do not work against natural adversarial samples!**





NatADiff

Can we generate natural adversarial samples?

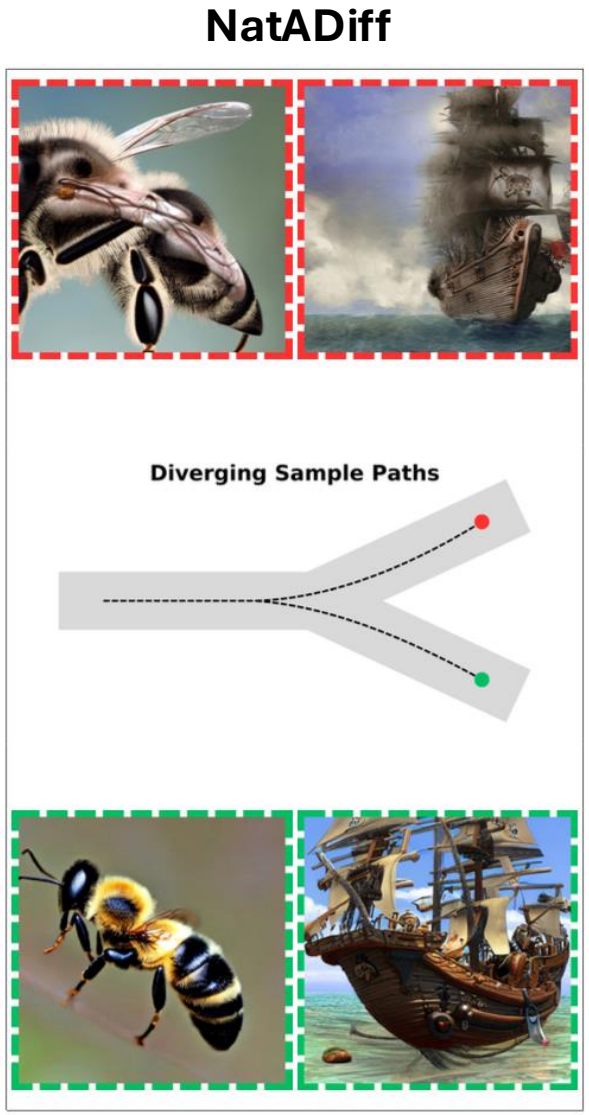
We propose NatADiff, a highly transferable, diffusion-based adversarial sample generation method.

Our main contributions are:

i. NatADiff: an adversarial sampling scheme that more closely mimics natural adversarial samples.

ii. Adversarial boundary guidance: a method of directing the diffusion sampling trajectory towards class intersections.

iii. We explore how convolution and transformer based classifiers perceive natural adversarial samples.



NatADiff's Guidance

- NatADiff uses a diffusion score that is conditioned on both the true (y) and target adversarial ($\tilde{y}$) classes:

Adversarial boundary guidance

Augmented adversarial classifier guidance

$$\nabla_{\boldsymbol{x}_t} \log(\bar{p}(\boldsymbol{x}_t|y,\tilde{y})) = -\frac{1}{\beta(t)}\left[\boldsymbol{\epsilon}_{\theta^\star}(\boldsymbol{x}_t,t) + (\omega - \mu\omega)\boldsymbol{v}_y + \mu\rho\boldsymbol{v}_{y\cap\tilde{y}}\right] + s\nabla_{\boldsymbol{x}_t}\log(p(\tilde{y}|\boldsymbol{x}_t))$$

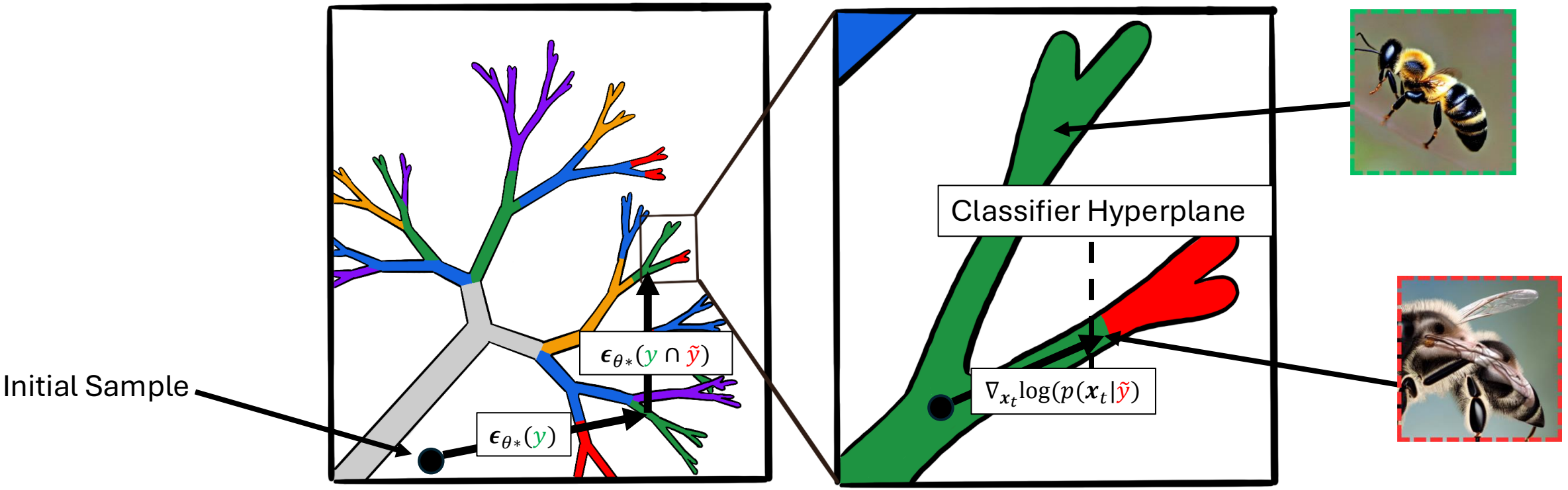
- Adversarial boundary guidance directs the sample towards the boundary of the true and adversarial classes.
  - This incorporates features from the target adversarial class ($\tilde{y}$) whilst maintaining the true class identity (y) .
  - We hypothesise that these will correlate with the erroneous contextual cues in natural adversarial samples.
- Augmented adversarial classifier guidance directs the sample towards misalignments on the classifier decision boundary.
  - Image transformations reduce the strength of constrained adversarial perturbations.

$$\nabla_{\boldsymbol{x}_t}\log(p(\tilde{y}|\boldsymbol{x}_t)) = \frac{\boldsymbol{g}(\boldsymbol{x}_t)}{\|\boldsymbol{g}(\boldsymbol{x}_t)\|_2}$$

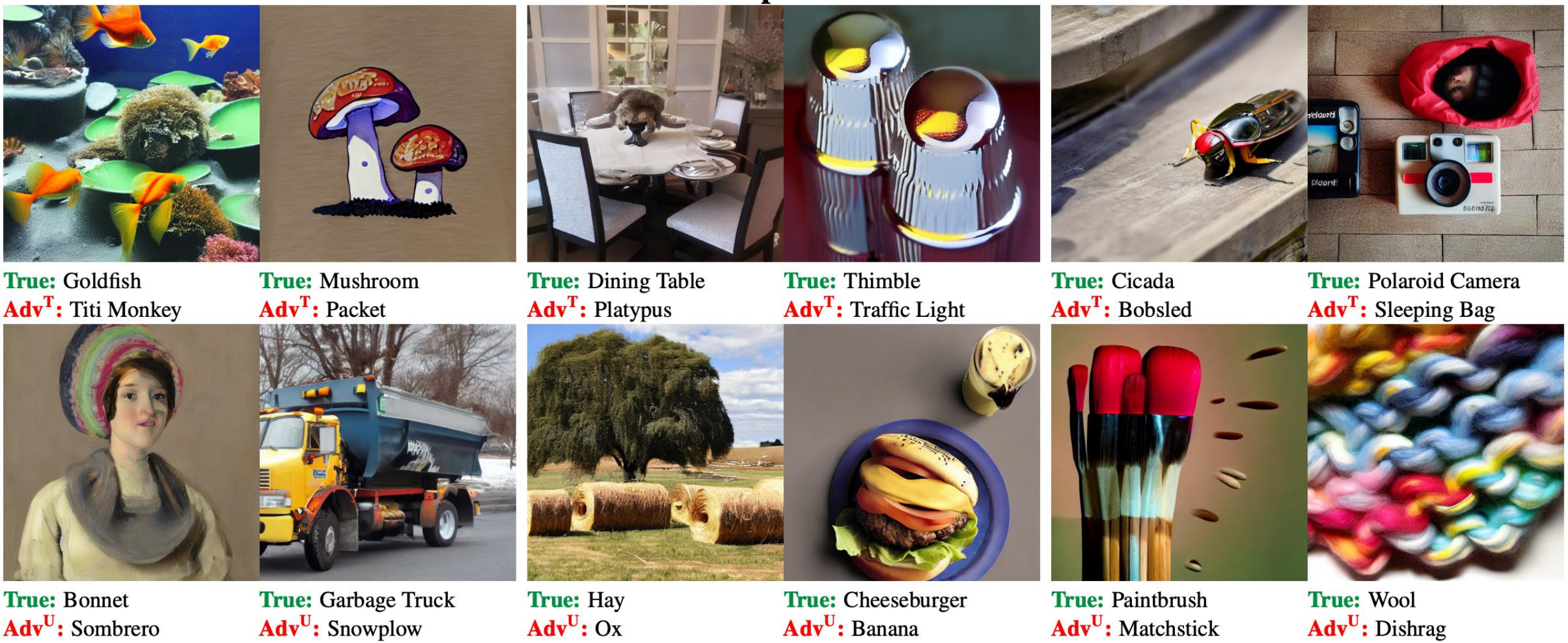
$$\boldsymbol{g}(\boldsymbol{x}_t) = \nabla_{\boldsymbol{x}_t}\log\left(\sigma_{\tilde{y}}\left(\frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|} h(T_i(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)))\right)\right)$$

NatADiff's Guidance – Visualization

- Classifier-free diffusion guidance directs the sample towards regions of the true class (bee), y.
- Adversarial boundary guidance directs the sample towards the true and adversarial class (ostrich) intersection, $y \cap \tilde{y}$.
- Augmented adversarial classifier guidance directs the sample towards the misaligned hyperplane.

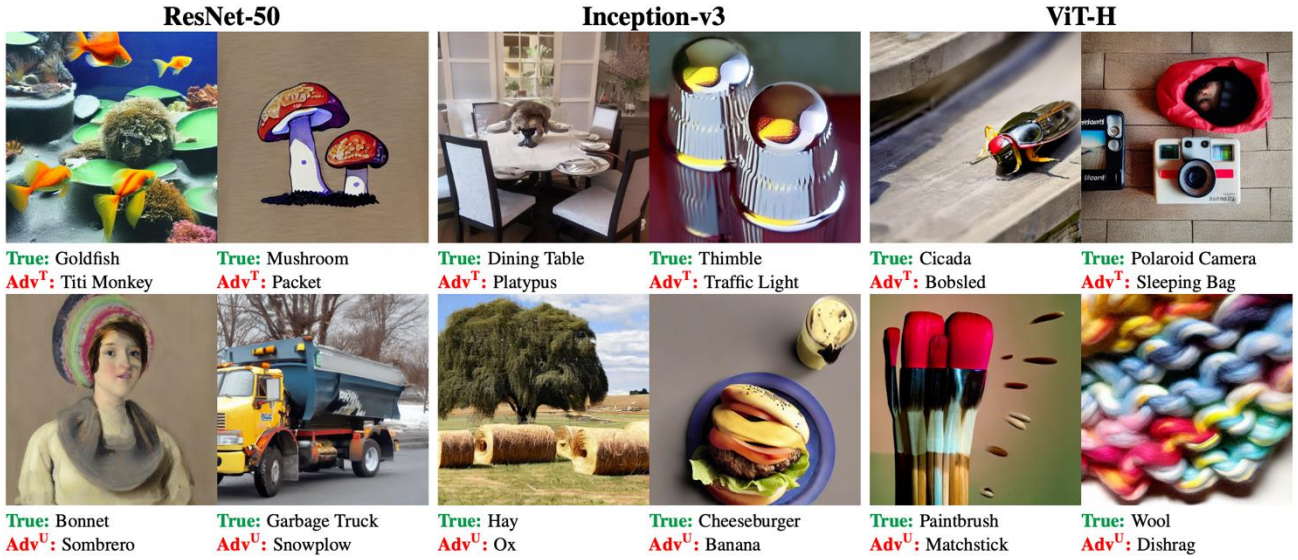


Results


ResNet-50
Inception-v3
ViT-H
True: Goldfish
Adv^T: Titi Monkey
True: Mushroom
Adv^T: Packet
True: Dining Table
Adv^T: Platypus
True: Thimble
Adv^T: Traffic Light
True: Cicada
Adv^T: Bobsled
True: Polaroid Camera
Adv^T: Sleeping Bag
True: Bonnet
Adv^U: Sombrero
True: Garbage Truck
Adv^U: Snowplow
True: Hay
Adv^U: Ox
True: Cheeseburger
Adv^U: Banana
True: Paintbrush
Adv^U: Matchstick
True: Wool
Adv^U: Dishrag

Results

- We compare NatADiff against state-of-the-art constrained and unconstrained adversarial attacks.
- NatADiff exhibits comparable white-box attack performance, and **vastly** superior attack transferability.
- NatADiff samples targeting convolution vs transformer architectures transferred better to the same architecture.
- There was marginal degradation in NatADiff's image quality as compared to base model.
- NatADiff samples had superior alignment with natural adversaries as compared to other generative attacks.



Surrogate Model	Attack	Victim Model ASR (%)									Average ASR	IS (↑)	FID-Val (↓)	FID-A (↓)
		CNNs					Transformers							
		RN-50	Inc-v3	RN-152	AdvRes	AdvInc	ViT-H	Max-ViT	Swin-B	DeiT				
	Clean	5.3	7.6	2.9	3.0	5.8	10.9	3.8	4.5	7.4	5.7	55.0	58.0	94.7
RN-50	PGD	99.4*	11.8	5.2	4.9	8.1	10.5	4.4	5.5	8.2	17.6	-	-	-
	AA	**100***	13.3	10.0	3.9	8.8	10.5	5.4	5.6	8.0	18.4	-	-	-
	NCF	74.8*	33.4	37.3	28.2	31.2	17.2	24.0	31.7	37.2	35.0	30.4	69.7	85.5
	DiffAttack	92.5*	47.1	52.5	35.3	43.3	28.4	44.6	42.4	38.9	47.2	26.8	64.1	**76.8**
	ACA	78.8*	53.3	52.7	49.8	53.1	41.8	46.4	49.3	50.6	52.9	23.9	65.0	77.9
	AdvClass^T	99.6*	35.0	32.1	31.4	33.5	25.8	30.0	30.8	32.8	39.0	38.3	**48.9**	92.4
	AdvClass^U	99.9*	42.5	44.3	38.7	41.1	29.7	37.6	38.4	39.1	45.7	38.5	50.2	92.7
	NatADiff^T	96.9*	60.1	56.5	55.3	58.9	36.8	45.3	49.0	52.3	56.8	26.0	66.5	77.3
	NatADiff^U	99.3*	**68.3**	**72.1**	**65.3**	**66.8**	**45.3**	**64.1**	**65.2**	**67.0**	**68.2**	**43.2**	51.4	95.9
Inc-v3	PGD	6.0	99.7*	4.0	5.1	10.4	10.2	4.1	5.6	7.4	16.9	-	-	-
	AA	7.3	**100***	4.9	4.8	12.8	10.6	5.7	6.1	8.0	17.8	-	-	-
	NCF	31.0	66.7*	23.1	29.0	36.3	15.8	18.3	20.4	30.5	30.1	31.7	69.1	83.0
	DiffAttack	29.0	74.6*	23.7	30.0	39.9	18.9	22.9	26.5	25.8	32.4	33.2	63.7	**78.2**
	ACA	50.9	67.8*	48.2	54.2	60.1	43.6	45.1	48.8	51.3	52.2	23.1	68.0	78.8
	AdvClass^T	35.1	99.6*	34.5	35.6	39.5	28.8	32.4	34.0	35.7	41.7	33.7	51.0	89.2
	AdvClass^U	38.0	99.9*	38.7	40.4	44.2	30.0	36.0	36.6	38.9	44.8	39.7	**49.4**	93.3
	NatADiff^T	53.4	97.9*	49.4	57.3	62.6	35.4	44.4	45.1	50.8	55.2	27.7	66.6	**78.2**
	NatADiff^U	**67.4**	99.4*	**65.7**	**70.1**	**75.7**	**44.4**	**60.3**	**60.2**	**63.1**	**67.4**	**47.0**	50.5	98.9
ViT-H	PGD	5.8	11.0	3.6	4.0	7.8	96.2*	4.5	5.4	9.2	16.4	-	-	-
	AA	6.5	9.8	3.9	4.3	8.6	**100***	4.5	5.9	9.9	17.0	-	-	-
	NCF	20.0	19.4	14.8	15.4	18.5	50.6*	11.9	15.6	21.2	20.8	**39.8**	63.1	86.4
	DiffAttack	20.5	25.0	17.2	18.9	22.4	73.2*	18.1	22.3	20.6	26.5	35.2	63.4	**80.0**
	ACA	50.5	54.5	48.1	49.1	52.8	75.8*	47.5	49.7	50.5	53.2	25.5	64.2	80.9
	AdvClass^T	33.9	35.9	33.4	34.4	34.4	92.6*	31.9	33.4	36.0	40.7	38.9	**48.5**	95.2
	AdvClass^U	35.2	37.5	35.8	35.2	36.0	98.7*	33.9	34.9	37.7	42.8	39.2	**48.5**	98.8
	NatADiff^T	**70.7**	**73.5**	**68.4**	**71.3**	**72.1**	98.5*	**65.7**	**66.9**	**71.7**	**73.2**	15.3	88.0	93.5
	NatADiff^U	66.8	67.0	65.3	64.9	65.8	99.6*	63.9	65.4	68.6	69.7	31.9	53.9	96.2