Sample-specific Noise Injection for Diffusion-based Adversarial Purification

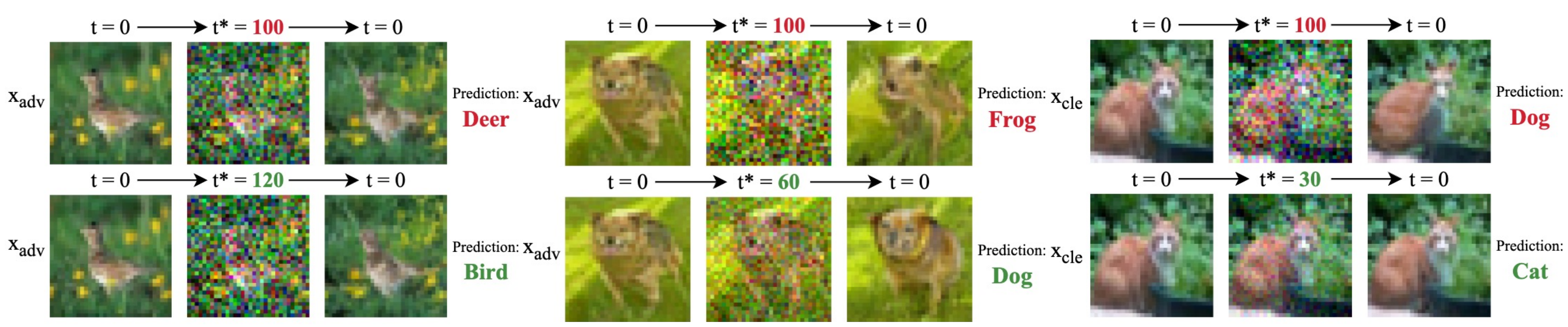
Yuhao Sun^, Jiacheng Zhang^, Zesheng Ye^, Chaowei Xiao, Feng Liu*

The University of Melbourne and The University of Wisconsin Email: fengliu.ml@gmail.com

Key Challenges in DBP Methods

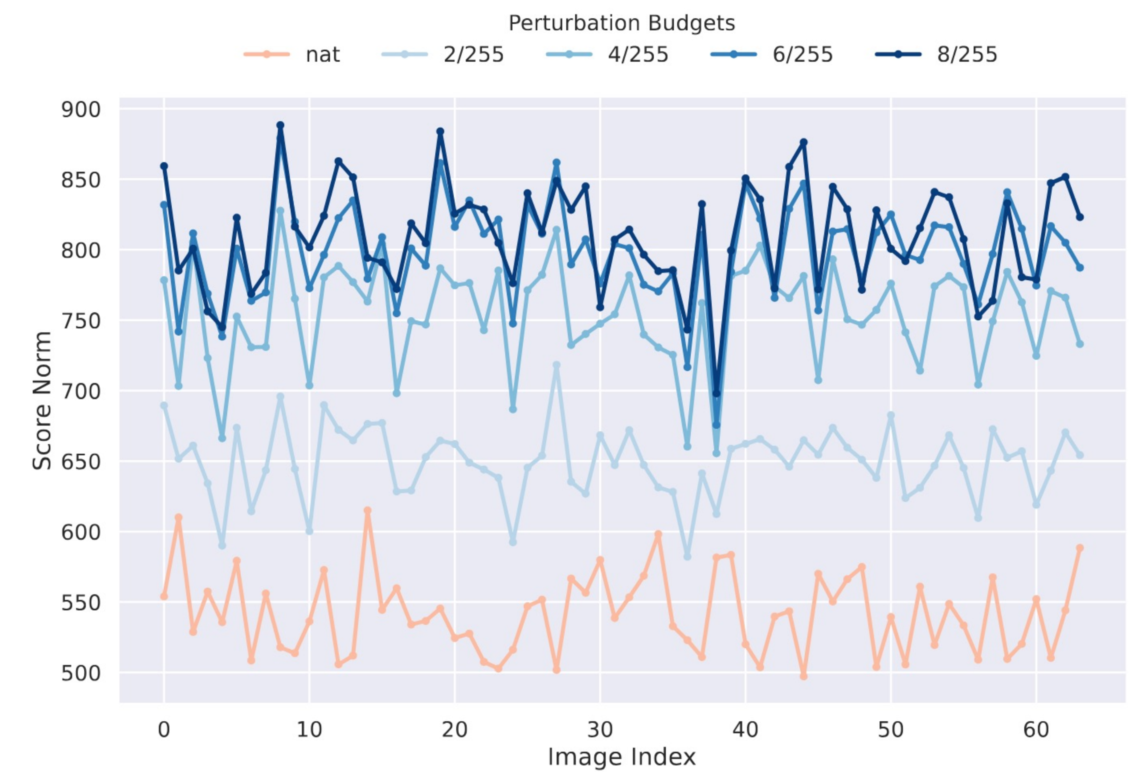
- If the noise level t is ***too small***, then adversarial noise cannot be fully removed.
- If the noise level t is ***too large***, then the purified image may have a different semantic meaning.
- Existing methods empirically select a ***fixed*** noise level t^* for all images, which is ***counterintuitive***.

Proof-of-concept Experiments



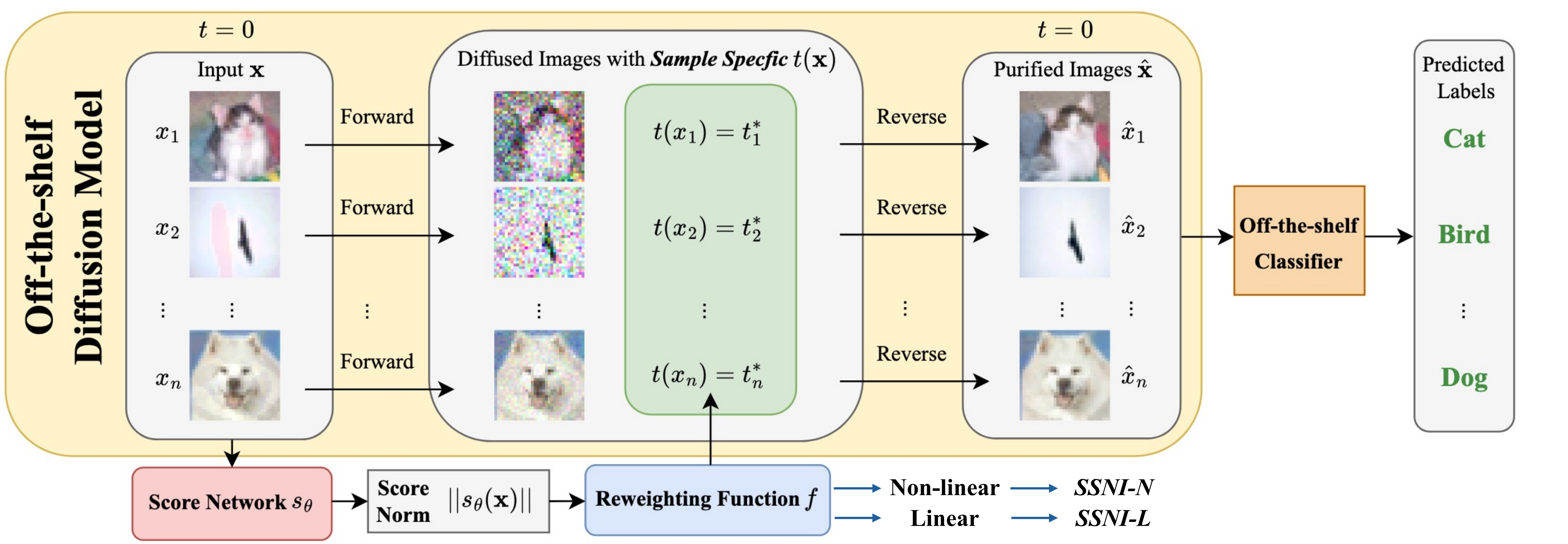
- Globally shared noise level $t^* = 100$ results in ***suboptimal*** prediction performance.
- $t^* = 100$ is ***insufficient*** for some images (e.g., some adversarial images), but ***excessive*** for others (e.g., clean images). For instance, the image is classified as "frog" (incorrect) with $t^* = 100$ but as "dog" (correct) with $t^* = 60$.
- These highlight the need for a ***sample-wise noise level adjustment***.

How to Reweight t?



- We find that score norms ***scale directly*** with perturbation budgets. A lower score norm means closer to clean data distribution.
- Score norms can act as ***proxies*** for estimating the sample-specific noise level.

A New Framework: Sample-specific Score-aware Noise Injection (SSNI)



Experiment Results

PGD+EOT ℓ_∞ ($\epsilon = 8/255$)

	DBP Method	Standard	Robust
WRN-28-10	Nie et al. (2022)	89.71±0.72	47.98±0.64
	+ *SSNI-N*	**93.29±0.37** (+3.58)	**48.63±0.56** (+0.65)
	Wang et al. (2022)	92.45±0.64	36.72±1.05
	+ *SSNI-N*	**94.08±0.33** (+1.63)	**40.95±0.65** (+4.23)
	Lee & Kim (2023)	90.10±0.18	56.05±1.11
	+ *SSNI-N*	**93.55±0.55** (+3.45)	**56.45±0.28** (+0.40)
WRN-70-16	Nie et al. (2022)	90.89±1.13	52.15±0.30
	+ *SSNI-N*	**94.47±0.51** (+3.58)	**52.47±0.66** (+0.32)
	Wang et al. (2022)	93.10±0.51	43.55±0.58
	+ *SSNI-N*	**95.57±0.24** (+2.47)	**46.03±1.33** (+2.48)
	Lee & Kim (2023)	89.39±1.12	56.97±0.33
	+ *SSNI-N*	**93.82±0.24** (+4.43)	**57.03±0.28** (+0.06)

PGD+EOT ℓ_2 ($\epsilon = 0.5$)

	DBP Method	Standard	Robust
WRN-28-10	Nie et al. (2022)	91.80±0.84	**82.81±0.97**
	+ *SSNI-N*	**93.95±0.70** (+2.15)	82.75±1.01 (-0.06)
	Wang et al. (2022)	92.45±0.64	82.29±0.82
	+ *SSNI-N*	**94.08±0.33** (+1.63)	**82.49±0.75** (+0.20)
	Lee & Kim (2023)	90.10±0.18	83.66±0.46
	+ *SSNI-N*	**93.55±0.55** (+3.45)	**84.05±0.33** (+0.39)
WRN-70-16	Nie et al. (2022)	92.90±0.40	82.94±1.13
	+ *SSNI-N*	**95.12±0.58** (+2.22)	**84.38±0.58** (+1.44)
	Wang et al. (2022)	93.10±0.51	**85.03±0.49**
	+ *SSNI-N*	**95.57±0.24** (+2.47)	84.64±0.51 (-0.39)
	Lee & Kim (2023)	89.39±1.12	84.51±0.37
	+ *SSNI-N*	**93.82±0.24** (+4.43)	**84.83±0.33** (+0.32)

PGD+EOT ℓ_∞ ($\epsilon = 4/255$)

	DBP Method	Standard	Robust
RN-50	Nie et al. (2022)	68.23±0.92	30.34±0.72
	+ *SSNI-N*	**70.25±0.56** (+2.02)	**33.66±1.04** (+3.32)
	Wang et al. (2022)	74.22±0.12	0.39±0.03
	+ *SSNI-N*	**75.07±0.18** (+0.85)	**5.21±0.24** (+4.82)
	Lee & Kim (2023)	70.18±0.60	42.45±0.92
	+ *SSNI-N*	**72.69±0.80** (+2.51)	**43.48±0.25** (+1.03)

BPDA+EOT ℓ_∞ ($\epsilon = 8/255$)

	DBP Method	Standard	Robust
WRN-28-10	Nie et al. (2022)	89.71±0.72	81.90±0.49
	+ *SSNI-N*	**93.29±0.37** (+3.58)	**82.10±1.15** (+0.20)
	Wang et al. (2022)	92.45±0.64	79.88±0.89
	+ *SSNI-N*	**94.08±0.33** (+1.63)	**80.99±1.09** (+1.11)
	Lee & Kim (2023)	90.10±0.18	**88.40±0.88**
	+ *SSNI-N*	**93.55±0.55** (+3.45)	87.30±0.42 (-1.10)

ℓ_∞ ($\epsilon = 8/255$)

	DBP Method	Standard	AutoAttack	DiffAttack	Diff-PGD
WRN-28-10	Nie et al. (2022)	89.71±0.72	66.73±0.21	47.16±0.48	54.95±0.77
	+ *SSNI-N*	**93.29±0.37** (+3.58)	**66.94±0.44** (+0.21)	**48.15±0.22** (+0.99)	**56.10±0.35** (+1.15)
	Wang et al. (2022)	92.45±0.64	64.48±0.62	54.27±0.72	41.45±0.60
	+ *SSNI-N*	**94.08±0.33** (+1.63)	**66.53±0.46** (+2.05)	**55.81±0.33** (+1.54)	**42.91±0.56** (+1.46)
	Lee & Kim (2023)	90.10±0.18	69.92±0.30	56.04±0.58	59.02±0.28
	+ *SSNI-N*	**93.55±0.55** (+3.45)	**72.27±0.19** (+2.35)	**56.80±0.41** (+0.76)	**61.43±0.58** (+2.41)

DBP Method	Noise Injection Method	Time (s)
Nie et al. (2022)	-	3.934
	SSNI-L	4.473
	SSNI-N	4.474
Wang et al. (2022)	-	5.174
	SSNI-L	5.793
	SSNI-N	5.829
Lee & Kim (2023)	-	14.902
	SSNI-L	15.624
	SSNI-N	15.534

DBP Method	Noise Injection Method	Time (s)
Nie et al. (2022)	-	8.980
	SSNI-L	14.515
	SSNI-N	14.437
Wang et al. (2022)	-	11.271
	SSNI-L	16.657
	SSNI-N	16.747
Lee & Kim (2023)	-	35.091
	SSNI-L	40.526
	SSNI-N	40.633

Paper **Code** **Contact**