Contents

A On the Relationship between $\mathcal{P}_{x,f}$ and $\mathcal{P}_{x,f\circ g}$: Formal Proof

Experiments of paper Section **??** show that the domain of attacks is being shifted such that their range does not lie on the vulnerable space $\mathcal{P}_{x,f}$. We conduct a formal analysis of the relationship between perturbation spaces $\mathcal{P}_{x,f}$ and $\mathcal{P}_{x,f\circ g}$ and show that one is a subset of the other. There are a few simplifications required for the proof, which will be accounted for at the end of this section. First, we use the results of BIM for an $\mathcal{L}_2 = 0.05$ tested on ResNet 50 as a reference. Values for the baseline (no defense), White-Box$^+$ and, Gray-Box$^-$ are 0.0, 0.0 and 0.672041 respectively. Computing the relative drop in performance (*i.e.*, normalizing by 0.75004: the accuracy under no attack) yields 0.0, 0.0 and 0.8960. To simplify the handling of fuzzy sets, we assign the membership functions $\mu(\alpha(x))$ for each perturbation set to be either 0 or 1 if the fooling ratio is below random chance or above 0.8 respectively. With this in mind, we can define four axioms that come from the domain of the adversarial attack $\alpha : \nabla_f \cup \nabla_{f\circ g} \rightarrow \mathcal{P}$, and the aforementioned experiments with simplified membership functions. The proof is a simple proof by contradiction with case analysis on the first disjunction.

Proof.

1. $\forall_x(x \in \nabla_f \vee x \in \nabla_{f\circ g})$ (Def. domain of α)
2. $\forall_x(x \in \nabla_f \rightarrow \alpha(x) \in \mathcal{P}_{x,f})$ (Baseline Exp.)
3. $\forall_x(x \in \nabla_f \rightarrow \alpha(x) \in \mathcal{P}_{x,f\circ g})$ (White-Box$^+$)
4. $\forall_x(x \in \nabla_{f\circ g} \rightarrow \alpha(x) \notin \mathcal{P}_{x,f})$ (Gray-Box$^-$)

---

5. $\exists_x(\alpha(x) \in \mathcal{P}_{x,f} \wedge \alpha(x) \notin \mathcal{P}_{x,f\circ g})$ (Assumption, $\not\subseteq$)
6. $\alpha(a) \in \mathcal{P}_{a,f} \wedge \alpha(a) \notin \mathcal{P}_{a,f\circ g}$ (Skolemization $x \rightarrow a, 5$)
7. $a \in \nabla_f \vee a \in \nabla_{f\circ g}$ (U.I. $x \rightarrow a$, 1)
8. $a \in \nabla_f \rightarrow \alpha(a) \in \mathcal{P}_{a,f\circ g}$ (U.I. $x \rightarrow a$, 3)
9. $a \in \nabla_{f\circ g} \rightarrow \alpha(a) \notin \mathcal{P}_{a,f}$ (U.I. $x \rightarrow a$, 4)

---

10. $a \in \nabla_{f\circ g}$ (Assumption, 7)
11. $\alpha(a) \notin \mathcal{P}_{a,f}$ ($\rightarrow$ 10, 9)
12. $\alpha(a) \notin \mathcal{P}_{a,f} \wedge \alpha(a) \in \mathcal{P}_{a,f}$ ($\wedge$, 11, 6)
13. Contradiction! $\lightning$

---

14. $\neg(a \in \nabla_{f \circ g})$ (Q.E.A. 10)

15. $a \in \nabla_f$ (D.Syllogism, 14, 7)

16. $\alpha(a) \in \mathcal{P}_{a,f \circ g}$ ($\rightarrow$ 15, 8)

17. $\alpha(a) \notin \mathcal{P}_{a,f \circ g} \wedge \alpha(a) \in \mathcal{P}_{a,f \circ g}$ ($\wedge$, 6, 16)

18. Contradiction! $\lightning$

---

19. $\neg\exists_x(\alpha(x) \in \mathcal{P}_{x,f} \wedge \alpha(x) \notin \mathcal{P}_{x,f \circ g})$ (Q.E.A., 5)

20. $\forall_x\neg(\alpha(x) \in \mathcal{P}_{x,f} \wedge \alpha(x) \notin \mathcal{P}_{x,f \circ g})$ ($\neg\exists$, 19)

21. $\forall_x(\alpha(x) \notin \mathcal{P}_{x,f} \vee \alpha(x) \in \mathcal{P}_{x,f \circ g})$ (Distr. $\neg$, 20)

22. $\forall_x(\alpha(x) \in \mathcal{P}_{x,f} \rightarrow \alpha(x) \in \mathcal{P}_{x,f \circ g})$ (MI $\rightarrow$., 21)

23. $\mathcal{P}_{x,f} \subseteq \mathcal{P}_{x,f \circ g}$ (Def. $\subseteq$, 22)

□

Going back to the simplified membership function, it follows that different reference experiments (network, attack and $\mathcal{L}_2$) will naturally yield different results. This is especially true if we take the raw accuracy as the membership function, instead of the simplified one. However, one can argue that the subset relationship is, in general terms, valid since overall, most experiments under the simplified membership function yield the same axioms). This is why we denote such a relationship by the approximate subset relationship $\subseteq$ to refer to this result.

A similar analysis can be done graphically as shown in Figure 1 (middle). Case `A` is covered by Gray-Box$^-$ experiments which show that $\mathcal{P}_{x,f}$ can only be reached 10.4% of the time, which we know now, actually corresponds to case `B` due to the inclusion $\mathcal{P}_{x,f} \subsetneq \mathcal{P}_{x,f \circ g}$. This leaves cases `C`, `D` which are covered by White-Box experiments. Here, using the BIM on ResNet 50 and $\mathcal{L}_2$ = 0.05 as reference, yields that $\alpha(x) \in \mathcal{P}_{x,f \circ g}$ with a membership $\mu(\alpha(x)) = 1 - (0.5/0.75) = 0.333$. That leaves case `D` with perturbations in the remaining 0.667 in $\mathcal{P} - \{\mathcal{P}_{x,f} \cup \mathcal{P}_{x,f \circ g}\}$.

Likewise, for cases where the domain is ∇_f, we see that they all fall into $\mathcal{P}_{x,f}$ which we know lies in $\mathcal{P}_{x,f \circ g}$ hence, they all fall into case `F`, eliminating samples falling into the remaining ones `E`, `G`, `H`.

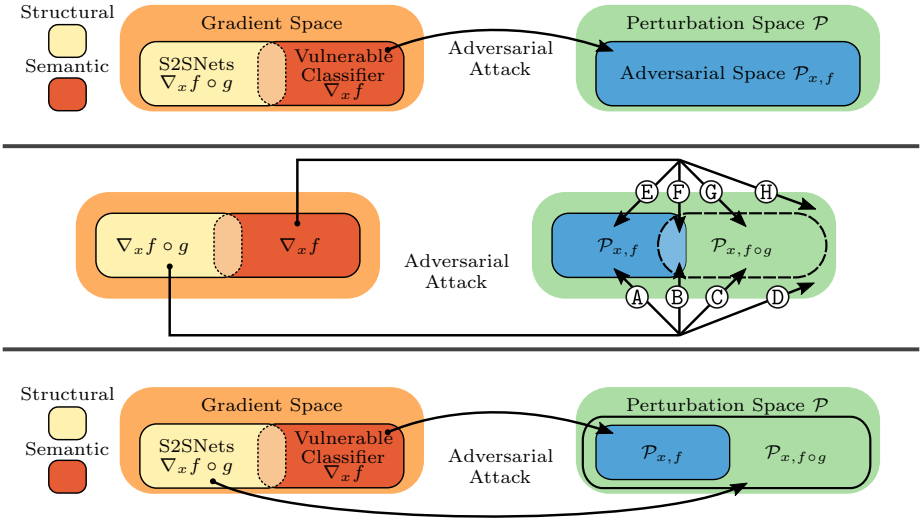

Figure 1: Analysis of the relationship between gradient space and perturbation space. By doing case analysis (middle) and a proof by contradiction, we infer that the perturbation space approximates $\mathcal{P}_{x,f} \subsetneq \mathcal{P}_{x,f\circ g}$

B Perturbed Images and their Reconstructions

The following figures (2-8) show attack attempts on ResNet 50 (R) with correctly classified images from our random shuffle of the ImageNet validation set. x (column 1) denotes the clean image and $\hat{x}$ its perturbed variants. Columns 2-7 show a combination of attack method (FGSM, BIM, CW) and gradient sources ($\nabla_x R(x)$, $\nabla_x R \circ \mathbb{A}_A(x)$). Below are perturbations δ (row 2), reconstructions $\hat{x}' = \mathbb{A}_R(\hat{x})$ (row 3), and remaining perturbation in reconstructions $\delta' = \mathbb{A}_R(\hat{x}) - \mathbb{A}_R(x)$ (row 4). Perturbation images are subject to histogram equalization for increases visibility. These examples further illustrate the structural nature of attacks through S2SNets.

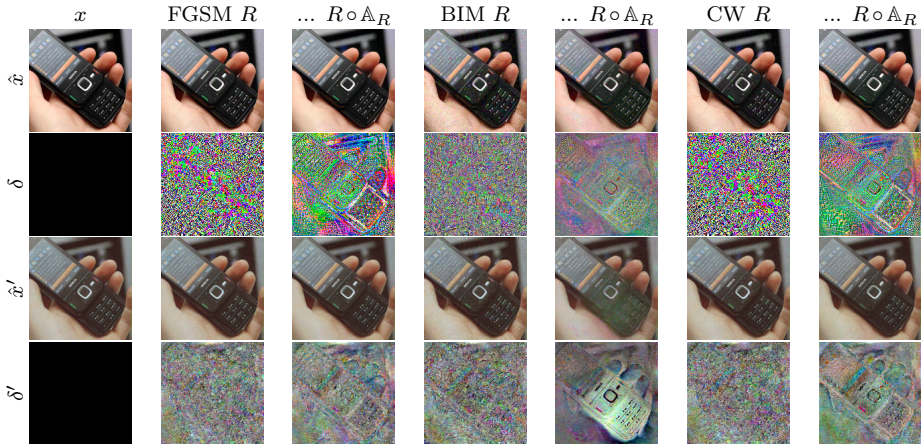

Figure 2: Example of different attacks on ResNet 50 (R) with and without $\mathbb{A}_R$, along with perturbations and reconstructions. $\mathcal{L}_2$ (left to right): 0.032, 0.032, 0.098, 0.115, 0.052, 0.064.

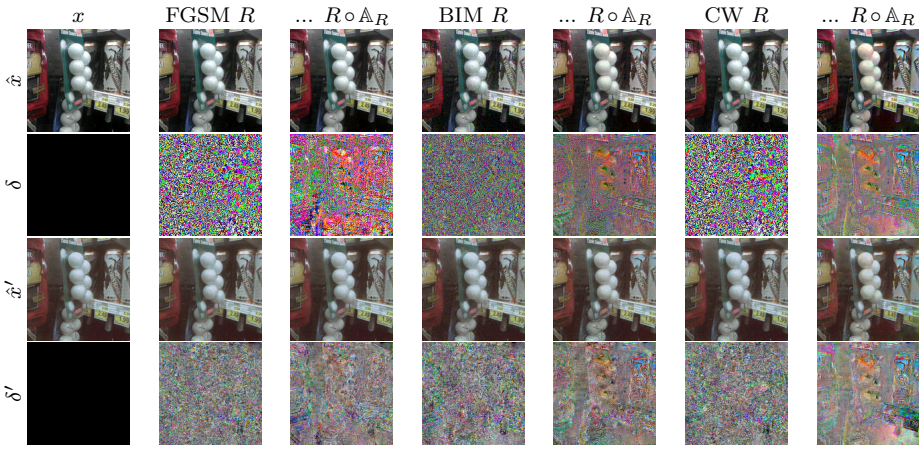

Figure 3: Example of different attacks on ResNet 50 (R) with and without $\mathbb{A}_R$, along with perturbations and reconstructions. $\mathcal{L}_2$ (left to right): 0.036, 0.036, 0.111, 0.104, 0.059, 0.172.

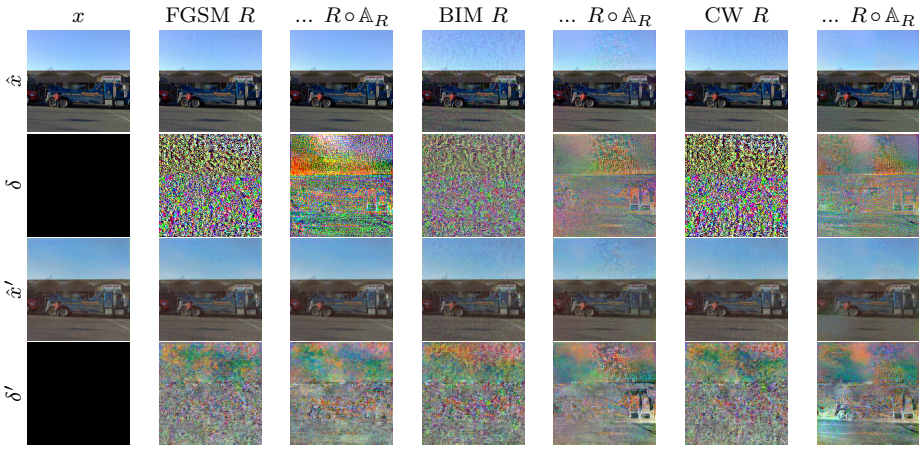

Figure 4: Example of different attacks on ResNet 50 (R) with and without $\mathbb{A}_R$, along with perturbations and reconstructions. $\mathcal{L}_2$ (left to right): 0.028, 0.028, 0.084, 0.105, 0.045, 0.095.

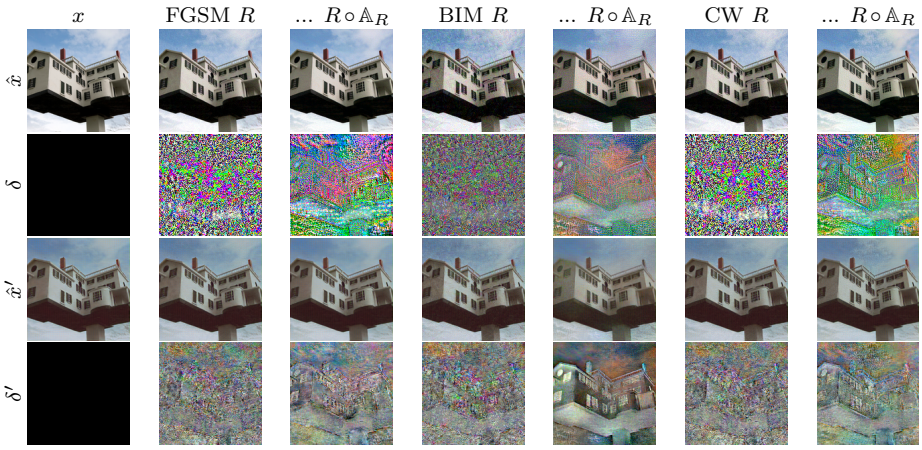

Figure 5: Example of different attacks on ResNet 50 (R) with and without $\mathbb{A}_R$, along with perturbations and reconstructions. $\mathcal{L}_2$ (left to right): 0.028, 0.028, 0.088, 0.119, 0.046, 0.054.

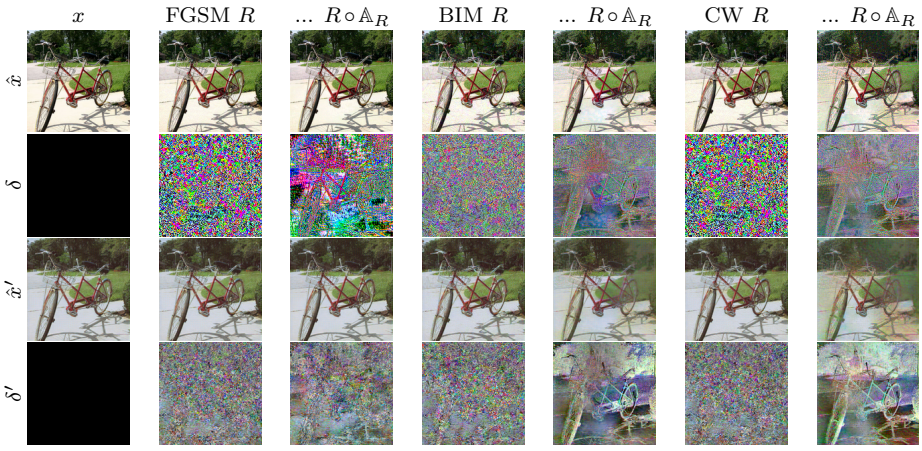

Figure 6: Example of different attacks on ResNet 50 (R) with and without $\mathbb{A}_R$, along with perturbations and reconstructions. $\mathcal{L}_2$ (left to right): 0.025, 0.025, 0.077, 0.124, 0.041, 0.212.

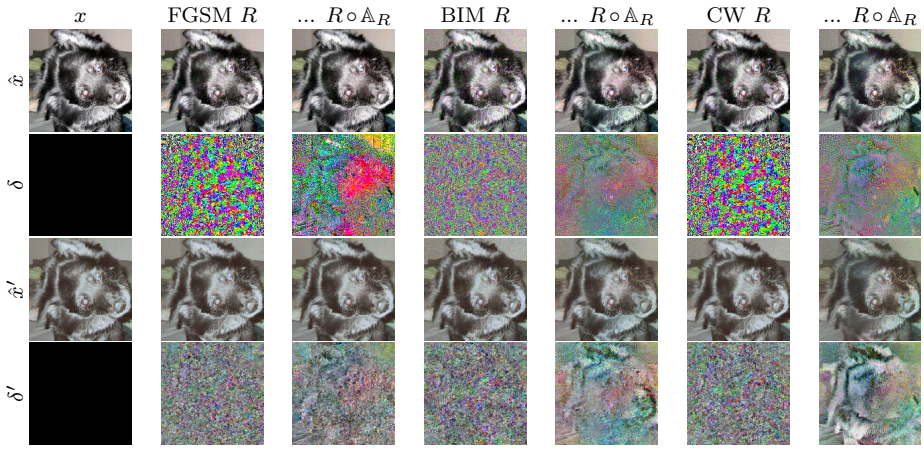

Figure 7: Example of different attacks on ResNet 50 (R) with and without $\mathbb{A}_R$, along with perturbations and reconstructions. $\mathcal{L}_2$ (left to right): 0.030, 0.030, 0.095, 0.107, 0.050, 0.150.

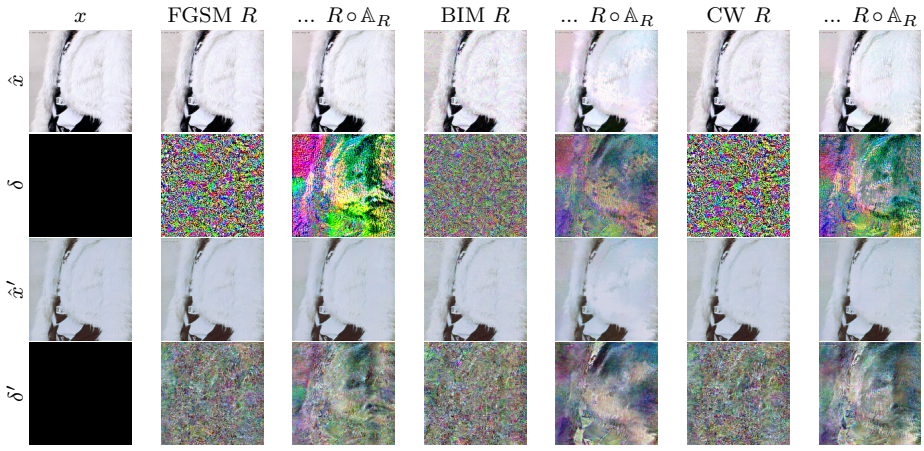

Figure 8: Example of different attacks on ResNet 50 (R) with and without $\mathbb{A}_R$, along with perturbations and reconstructions. $\mathcal{L}_2$ (left to right): 0.018, 0.018, 0.053, 0.067, 0.029, 0.045.

C Raw Experiment Data

This section contains raw values for all experiments we conducted. A subset was used to create the plots seen in the paper, but we also provide results we ended up not using for our argumentation. In the following tables, $\mathcal{L}_2$ denotes the normalized $L2$ dissimilarity (Equation 1) wrt. reference image x_1 and another image x_2. We also provide L_∞ values, the usual infinity norm divided by 255 (Equation 2). Settings for attacks besides ϵ are defined in paper Section **??**.

$$\mathcal{L}_2(x_1, x_2) = \frac{||x_1 - x_2||_2}{||x_1||_2} \tag{1}$$

$$L_\infty(x_1, x_2) = \frac{max(|x_1 - x_2|)}{255} \tag{2}$$

Tables 1, 2, and 3 contain White-Box results for the respective attack method. They cover experiments described in paper Section **??** Tables 4, 5, and 6 contain results attacks where the network that supplied the gradients (source network) that the attack is based on is different from the network that is attacked (target network). They cover White-Box$^+$ and Gray-Box$^-$ experiments described in paper Section **??** and **??**. Finally, Table 7 contains results on BIM attacks mitigated by different types and strengths (η) of noise.

Table 1: Accuracy and distance statistics for FGSM attacks.

Network	ϵ	top-1 acc. (%)	...(adversarial)	$\mathcal{L}_2$	L_∞
A	0.5	56.48	20.30	0.00413	0.00196
A	1.0	56.48	8.60	0.00825	0.00392
A	2.0	56.48	2.80	0.01647	0.00784
A	4.0	56.48	1.14	0.03283	0.01569
A	8.0	56.48	0.83	0.06533	0.03137
A	16.0	56.48	0.69	0.12945	0.06275
I	0.5	76.86	32.58	0.00413	0.00196
I	1.0	76.86	25.32	0.00825	0.00392
I	2.0	76.86	21.32	0.01647	0.00784
I	4.0	76.86	20.04	0.03284	0.01569
I	4.0	76.86	20.04	0.03284	0.01569
I	8.0	76.86	21.37	0.06535	0.03137
I	16.0	76.86	25.09	0.12948	0.06275
R	0.5	76.03	23.49	0.00413	0.00196
R	1.0	76.03	12.58	0.00826	0.00392
R	2.0	76.03	8.00	0.01648	0.00784
R	4.0	76.03	6.90	0.03284	0.01569
R	8.0	76.03	8.02	0.06535	0.03137
R	16.0	76.03	10.37	0.12947	0.06275
V	0.5	73.28	8.82	0.00413	0.00196
V	1.0	73.28	3.75	0.00826	0.00392

Table 1: Accuracy and distance statistics for FGSM attacks.

Network	ϵ	top-1 acc. (%)	...(adversarial)	$\mathcal{L}_2$	L_∞
V	2.0	73.28	2.74	0.01648	0.00784
V	4.0	73.28	2.85	0.03285	0.01569
V	8.0	73.28	3.79	0.06537	0.03137
V	16.0	73.28	6.64	0.12951	0.06275
$A \circ \mathbb{A}_A$	0.5	56.38	54.90	0.00413	0.00196
$A \circ \mathbb{A}_A$	1.0	56.38	53.79	0.00827	0.00392
$A \circ \mathbb{A}_A$	2.0	56.38	52.37	0.01650	0.00784
$A \circ \mathbb{A}_A$	4.0	56.38	50.68	0.03289	0.01569
$A \circ \mathbb{A}_A$	8.0	56.38	48.56	0.06545	0.03137
$A \circ \mathbb{A}_A$	16.0	56.38	42.86	0.12970	0.06275
$I \circ \mathbb{A}_I$	0.5	76.79	75.46	0.00413	0.00196
$I \circ \mathbb{A}_I$	1.0	76.79	74.51	0.00825	0.00392
$I \circ \mathbb{A}_I$	2.0	76.79	73.41	0.01647	0.00784
$I \circ \mathbb{A}_I$	4.0	76.79	71.91	0.03283	0.01569
$I \circ \mathbb{A}_I$	8.0	76.79	69.89	0.06532	0.03137
$I \circ \mathbb{A}_I$	16.0	76.79	65.40	0.12942	0.06275
$R \circ \mathbb{A}_R$	0.5	75.00	73.35	0.00413	0.00196
$R \circ \mathbb{A}_R$	1.0	75.00	72.33	0.00826	0.00392
$R \circ \mathbb{A}_R$	2.0	75.00	70.99	0.01648	0.00784
$R \circ \mathbb{A}_R$	4.0	75.00	69.33	0.03285	0.01569
$R \circ \mathbb{A}_R$	8.0	75.00	66.65	0.06535	0.03137
$R \circ \mathbb{A}_R$	16.0	75.00	60.47	0.12948	0.06275
$V \circ \mathbb{A}_V$	0.5	68.14	65.20	0.00413	0.00196
$V \circ \mathbb{A}_V$	1.0	68.14	63.38	0.00827	0.00392
$V \circ \mathbb{A}_V$	2.0	68.14	61.01	0.01650	0.00784
$V \circ \mathbb{A}_V$	4.0	68.14	58.16	0.03289	0.01569
$V \circ \mathbb{A}_V$	8.0	68.14	54.85	0.06544	0.03137
$V \circ \mathbb{A}_V$	16.0	68.14	47.38	0.12965	0.06275
$I \circ \mathbb{A}_S$	0.5	73.80	71.37	0.00413	0.00196
$I \circ \mathbb{A}_S$	1.0	73.80	69.71	0.00825	0.00392
$I \circ \mathbb{A}_S$	2.0	73.80	67.63	0.01647	0.00784
$I \circ \mathbb{A}_S$	4.0	73.80	65.41	0.03283	0.01569
$I \circ \mathbb{A}_S$	8.0	73.80	63.10	0.06532	0.03137
$I \circ \mathbb{A}_S$	16.0	73.80	59.38	0.12943	0.06275
$R \circ \mathbb{A}_S$	0.5	72.07	68.75	0.00413	0.00196
$R \circ \mathbb{A}_S$	1.0	72.07	66.33	0.00825	0.00392
$R \circ \mathbb{A}_S$	2.0	72.07	63.47	0.01647	0.00784
$R \circ \mathbb{A}_S$	4.0	72.07	59.89	0.03283	0.01569
$R \circ \mathbb{A}_S$	8.0	72.07	56.03	0.06533	0.03137
$R \circ \mathbb{A}_S$	16.0	72.07	48.87	0.12943	0.06275

Table 2: Accuracy and distance statistics for BIM attacks.

Network	ϵ	top-1 acc. (%)	...(adversarial)	$\mathcal{L}_2$	L_∞
A	0.5	56.48	0.11	0.02278	0.01961
A	1.0	56.48	0.05	0.03842	0.03922
A	2.0	56.48	0.03	0.06572	0.07843
A	4.0	56.48	0.02	0.11770	0.15686
A	8.0	56.48	0.01	0.22130	0.31373
I	0.5	76.86	0.50	0.01508	0.01961
I	1.0	76.86	0.20	0.02698	0.03921
I	2.0	76.86	0.15	0.05147	0.07841
I	4.0	76.86	0.10	0.10090	0.15682
I	8.0	76.86	0.08	0.19724	0.31366
R	0.5	76.03	0.10	0.01669	0.01961
R	1.0	76.03	0.05	0.02847	0.03922
R	2.0	76.03	0.02	0.05232	0.07843
R	4.0	76.03	0.01	0.10067	0.15686
R	8.0	76.03	0.00	0.19510	0.31373
V	0.5	73.28	0.34	0.01724	0.01961
V	1.0	73.28	0.27	0.02920	0.03922
V	2.0	73.28	0.23	0.05305	0.07843
V	4.0	73.28	0.18	0.10117	0.15686
V	8.0	73.28	0.12	0.19571	0.31372
$A \circ \mathbb{A}_A$	0.5	56.38	44.11	0.02637	0.01961
$A \circ \mathbb{A}_A$	1.0	56.38	35.57	0.04537	0.03922
$A \circ \mathbb{A}_A$	2.0	56.38	24.47	0.07559	0.07843
$A \circ \mathbb{A}_A$	4.0	56.38	12.69	0.12549	0.15686
$A \circ \mathbb{A}_A$	8.0	56.38	4.00	0.22330	0.31373
$I \circ \mathbb{A}_I$	0.5	76.79	66.48	0.02235	0.01961
$I \circ \mathbb{A}_I$	1.0	76.79	60.47	0.03852	0.03922
$I \circ \mathbb{A}_I$	2.0	76.79	51.12	0.06634	0.07843
$I \circ \mathbb{A}_I$	4.0	76.79	37.65	0.11604	0.15686
$I \circ \mathbb{A}_I$	8.0	76.79	23.03	0.21053	0.31373
$R \circ \mathbb{A}_R$	0.5	75.00	62.05	0.02445	0.01961
$R \circ \mathbb{A}_R$	1.0	75.00	53.64	0.04197	0.03922
$R \circ \mathbb{A}_R$	2.0	75.00	41.21	0.07040	0.07843
$R \circ \mathbb{A}_R$	4.0	75.00	25.93	0.11964	0.15686
$R \circ \mathbb{A}_R$	8.0	75.00	11.78	0.21596	0.31373
$V \circ \mathbb{A}_V$	0.5	68.14	46.16	0.02467	0.01961
$V \circ \mathbb{A}_V$	1.0	68.14	34.45	0.04230	0.03922
$V \circ \mathbb{A}_V$	2.0	68.14	22.14	0.07086	0.07843
$V \circ \mathbb{A}_V$	4.0	68.14	12.65	0.11932	0.15686
$V \circ \mathbb{A}_V$	8.0	68.14	6.33	0.21275	0.31373
$R \circ \mathbb{A}_S$	0.5	72.07	45.11	0.02507	0.01961
$R \circ \mathbb{A}_S$	1.0	72.07	31.29	0.04294	0.03922

Table 2: Accuracy and distance statistics for BIM attacks.

Network	ϵ	top-1 acc. (%)	... (adversarial)	$\mathcal{L}_2$	L_∞
$R \circ \mathbb{A}_S$	2.0	72.07	18.32	0.07170	0.07843
$R \circ \mathbb{A}_S$	4.0	72.07	9.86	0.12014	0.15686
$R \circ \mathbb{A}_S$	8.0	72.07	5.42	0.21249	0.31373
$R \circ \mathbb{A}_S$	16.0	72.07	1.09	0.40976	0.62744

Table 3: Accuracy and distance statistics for CW attacks.

Network	ϵ	top-1 acc. (%)	... (adversarial)	$\mathcal{L}_2$	L_∞
I	0.5	76.86	0.16	0.00805	0.00587
I	1.0	76.86	0.29	0.01475	0.00966
I	2.0	76.86	0.99	0.03084	0.01848
I	4.0	76.86	2.13	0.05731	0.03350
R	0.5	76.03	0.10	0.00752	0.00511
R	1.0	76.03	0.02	0.01387	0.00850
R	2.0	76.03	0.42	0.02964	0.01644
R	4.0	76.03	1.32	0.05459	0.02917
$I \circ \mathbb{A}_I$	0.5	76.80	44.49	0.07357	0.11869
$I \circ \mathbb{A}_I$	1.0	76.80	33.58	0.10195	0.17756
$I \circ \mathbb{A}_I$	2.0	76.80	25.19	0.13851	0.24872
$I \circ \mathbb{A}_I$	4.0	76.80	19.72	0.18622	0.31932
$R \circ \mathbb{A}_R$	0.5	75.01	36.66	0.07310	0.11166
$R \circ \mathbb{A}_R$	1.0	75.01	26.59	0.09461	0.15410
$R \circ \mathbb{A}_R$	2.0	75.01	20.63	0.11958	0.19562
$R \circ \mathbb{A}_R$	4.0	75.01	16.68	0.15458	0.23814
$R \circ \mathbb{A}_S$	0.5	72.07	20.26	0.04971	0.06407
$R \circ \mathbb{A}_S$	1.0	72.07	14.08	0.06044	0.07838
$R \circ \mathbb{A}_S$	2.0	72.07	11.27	0.07758	0.09839
$R \circ \mathbb{A}_S$	4.0	72.07	10.45	0.10965	0.13377

Table 4: Accuracy for FGSM attacks when source and target network differ.

Source Network	Target Network	ϵ	top-1 acc. (%)	... (adversarial)
A	I	0.5	56.48	76.31
A	R	0.5	56.48	75.29
A	V	0.5	56.48	72.30
A	$A \circ \mathbb{A}_A$	0.5	56.48	41.39
A	I	1.0	56.48	75.70
A	R	1.0	56.48	74.39

Table 4: Accuracy for FGSM attacks when source and target network differ.

Source Network	Target Network	ϵ	top-1 acc. (%)	... (adversarial)
A	V	1.0	56.48	71.19
A	$A \circ \mathbb{A}_A$	1.0	56.48	29.07
A	I	2.0	56.48	74.27
A	R	2.0	56.48	72.36
A	V	2.0	56.48	68.84
A	$A \circ \mathbb{A}_A$	2.0	56.48	14.42
A	I	4.0	56.48	71.20
A	R	4.0	56.48	68.07
A	V	4.0	56.48	63.51
A	$A \circ \mathbb{A}_A$	4.0	56.48	5.32
A	I	8.0	56.48	64.46
A	R	8.0	56.48	58.65
A	V	8.0	56.48	52.28
A	$A \circ \mathbb{A}_A$	8.0	56.48	2.23
A	$A \circ \mathbb{A}_A$	16.0	56.48	1.34
I	A	0.5	76.86	56.11
I	R	0.5	76.86	74.01
I	V	0.5	76.86	70.97
I	$I \circ \mathbb{A}_I$	0.5	76.86	62.88
I	A	1.0	76.86	55.86
I	R	1.0	76.86	71.87
I	V	1.0	76.86	68.69
I	$I \circ \mathbb{A}_I$	1.0	76.86	51.84
I	A	2.0	76.86	55.02
I	R	2.0	76.86	68.23
I	V	2.0	76.86	64.78
I	$I \circ \mathbb{A}_I$	2.0	76.86	40.63
I	A	4.0	76.86	52.48
I	R	4.0	76.86	63.06
I	V	4.0	76.86	59.33
I	$I \circ \mathbb{A}_I$	4.0	76.86	33.02
I	A	8.0	76.86	44.76
I	R	8.0	76.86	56.61
I	V	8.0	76.86	51.99
I	$I \circ \mathbb{A}_I$	8.0	76.86	29.92
I	$I \circ \mathbb{A}_I$	16.0	76.86	29.42
R	A	0.5	76.03	55.93
R	I	0.5	76.03	74.63
R	V	0.5	76.03	69.25
R	$R \circ \mathbb{A}_R$	0.5	76.03	54.46
R	$R \circ \mathbb{A}_S$	0.5	76.03	47.90

Table 4: Accuracy for FGSM attacks when source and target network differ.

Source Network	Target Network	ϵ	top-1 acc. (%)	... (adversarial)
R	A	1.0	76.03	55.43
R	I	1.0	76.03	72.26
R	V	1.0	76.03	65.32
R	$R \circ \mathbb{A}_R$	1.0	76.03	39.20
R	$R \circ \mathbb{A}_S$	1.0	76.03	31.81
R	A	2.0	76.03	54.16
R	I	2.0	76.03	68.06
R	V	2.0	76.03	59.18
R	$R \circ \mathbb{A}_R$	2.0	76.03	23.53
R	$R \circ \mathbb{A}_S$	2.0	76.03	18.28
R	A	4.0	76.03	50.82
R	I	4.0	76.03	61.82
R	V	4.0	76.03	51.14
R	$R \circ \mathbb{A}_R$	4.0	76.03	14.43
R	$R \circ \mathbb{A}_S$	4.0	76.03	11.96
R	A	8.0	76.03	41.18
R	I	8.0	76.03	54.70
R	V	8.0	76.03	42.35
R	$R \circ \mathbb{A}_R$	8.0	76.03	11.67
R	$R \circ \mathbb{A}_S$	8.0	76.03	10.40
R	$R \circ \mathbb{A}_R$	16.0	76.03	11.25
V	A	0.5	73.28	56.01
V	I	0.5	73.28	75.11
V	R	0.5	73.28	73.22
V	$V \circ \mathbb{A}_V$	0.5	73.28	35.55
V	A	1.0	73.28	55.58
V	I	1.0	73.28	73.29
V	R	1.0	73.28	70.38
V	$V \circ \mathbb{A}_V$	1.0	73.28	18.50
V	A	2.0	73.28	54.38
V	I	2.0	73.28	69.89
V	R	2.0	73.28	65.24
V	$V \circ \mathbb{A}_V$	2.0	73.28	8.27
V	A	4.0	73.28	51.57
V	I	4.0	73.28	64.71
V	R	4.0	73.28	58.30
V	$V \circ \mathbb{A}_V$	4.0	73.28	5.41
V	A	8.0	73.28	43.48
V	I	8.0	73.28	58.73
V	R	8.0	73.28	51.20
V	$V \circ \mathbb{A}_V$	8.0	73.28	5.39

Table 4: Accuracy for FGSM attacks when source and target network differ.

Source Network	Target Network	ϵ	top-1 acc. (%)	... (adversarial)
V	$V \circ \mathbb{A}_V$	16.0	73.28	7.72
$A \circ \mathbb{A}_A$	A	0.5	56.38	56.26
$A \circ \mathbb{A}_A$	A	1.0	56.38	56.19
$A \circ \mathbb{A}_A$	A	2.0	56.38	55.89
$A \circ \mathbb{A}_A$	A	4.0	56.38	55.25
$A \circ \mathbb{A}_A$	A	8.0	56.38	52.33
$A \circ \mathbb{A}_A$	A	16.0	56.38	41.49
$I \circ \mathbb{A}_I$	I	0.5	76.79	76.71
$I \circ \mathbb{A}_I$	I	1.0	76.79	76.45
$I \circ \mathbb{A}_I$	I	2.0	76.79	75.90
$I \circ \mathbb{A}_I$	I	4.0	76.79	74.65
$I \circ \mathbb{A}_I$	I	8.0	76.79	72.45
$I \circ \mathbb{A}_I$	I	16.0	76.79	68.18
$R \circ \mathbb{A}_R$	R	0.5	75.00	75.71
$R \circ \mathbb{A}_R$	R	1.0	75.00	75.42
$R \circ \mathbb{A}_R$	R	2.0	75.00	74.68
$R \circ \mathbb{A}_R$	R	4.0	75.00	73.01
$R \circ \mathbb{A}_R$	R	8.0	75.00	69.59
$R \circ \mathbb{A}_R$	R	16.0	75.00	62.54
$V \circ \mathbb{A}_V$	V	0.5	68.14	72.52
$V \circ \mathbb{A}_V$	V	1.0	68.14	71.79
$V \circ \mathbb{A}_V$	V	2.0	68.14	70.12
$V \circ \mathbb{A}_V$	V	4.0	68.14	67.02
$V \circ \mathbb{A}_V$	V	8.0	68.14	62.11
$V \circ \mathbb{A}_V$	V	16.0	68.14	53.49
$R \circ \mathbb{A}_S$	R	0.5	72.07	75.03
$R \circ \mathbb{A}_S$	R	1.0	72.07	73.89
$R \circ \mathbb{A}_S$	R	2.0	72.07	71.70
$R \circ \mathbb{A}_S$	R	4.0	72.07	67.98
$R \circ \mathbb{A}_S$	R	8.0	72.07	62.50

Table 5: Accuracy for BIM attacks when source and target network differ.

Source Network	Target Network	ϵ	top-1 acc. (%)	... (adversarial)
A	I	0.5	56.48	71.70
A	R	0.5	56.48	68.98
A	V	0.5	56.48	63.81
A	$A \circ \mathbb{A}_A$	0.5	56.48	0.95
A	I	1.0	56.48	67.01

Table 5: Accuracy for BIM attacks when source and target network differ.

Source Network	Target Network	ϵ	top-1 acc. (%)	... (adversarial)
A	R	1.0	56.48	62.48
A	V	1.0	56.48	55.92
A	$A \circ \mathbb{A}_A$	1.0	56.48	0.13
A	I	2.0	56.48	58.45
A	R	2.0	56.48	50.59
A	V	2.0	56.48	42.54
A	$A \circ \mathbb{A}_A$	2.0	56.48	0.04
A	I	4.0	56.48	41.90
A	R	4.0	56.48	30.60
A	V	4.0	56.48	23.27
A	$A \circ \mathbb{A}_A$	4.0	56.48	0.02
A	I	8.0	56.48	18.61
A	R	8.0	56.48	10.25
A	V	8.0	56.48	6.32
A	$A \circ \mathbb{A}_A$	8.0	56.48	0.01
I	A	0.5	76.86	55.04
I	R	0.5	76.86	68.41
I	V	0.5	76.86	64.25
I	$I \circ \mathbb{A}_I$	0.5	76.86	23.14
I	A	1.0	76.86	53.51
I	R	1.0	76.86	62.54
I	V	1.0	76.86	57.50
I	$I \circ \mathbb{A}_I$	1.0	76.86	10.72
I	A	2.0	76.86	48.11
I	R	2.0	76.86	52.16
I	V	2.0	76.86	45.69
I	$I \circ \mathbb{A}_I$	2.0	76.86	3.87
I	A	4.0	76.86	33.50
I	R	4.0	76.86	37.46
I	V	4.0	76.86	30.33
I	$I \circ \mathbb{A}_I$	4.0	76.86	1.23
I	A	8.0	76.86	11.08
I	R	8.0	76.86	21.29
I	V	8.0	76.86	15.10
I	$I \circ \mathbb{A}_I$	8.0	76.86	0.40
R	A	0.5	76.03	53.95
R	I	0.5	76.03	67.32
R	V	0.5	76.03	53.11
R	$R \circ \mathbb{A}_R$	0.5	76.03	2.65
R	$R \circ \mathbb{A}_S$	0.5	76.03	0.73
R	A	1.0	76.03	51.69

Table 5: Accuracy for BIM attacks when source and target network differ.

Source Network	Target Network	ϵ	top-1 acc. (%)	...(adversarial)
R	I	1.0	76.03	60.49
R	V	1.0	76.03	41.59
R	$R \circ \mathbb{A}_R$	1.0	76.03	0.44
R	$R \circ \mathbb{A}_S$	1.0	76.03	0.11
R	A	2.0	76.03	45.05
R	I	2.0	76.03	49.84
R	V	2.0	76.03	26.86
R	$R \circ \mathbb{A}_R$	2.0	76.03	0.08
R	$R \circ \mathbb{A}_S$	2.0	76.03	0.03
R	A	4.0	76.03	29.14
R	I	4.0	76.03	34.22
R	V	4.0	76.03	13.02
R	$R \circ \mathbb{A}_R$	4.0	76.03	0.01
R	$R \circ \mathbb{A}_S$	4.0	76.03	0.01
R	A	8.0	76.03	8.70
R	I	8.0	76.03	17.36
R	V	8.0	76.03	4.42
R	$R \circ \mathbb{A}_R$	8.0	76.03	0.00
R	$R \circ \mathbb{A}_S$	8.0	76.03	0.00
V	A	0.5	73.28	54.42
V	I	0.5	73.28	70.39
V	R	0.5	73.28	64.93
V	$V \circ \mathbb{A}_V$	0.5	73.28	0.59
V	A	1.0	73.28	52.67
V	I	1.0	73.28	65.72
V	R	1.0	73.28	57.52
V	$V \circ \mathbb{A}_V$	1.0	73.28	0.33
V	A	2.0	73.28	47.17
V	I	2.0	73.28	57.29
V	R	2.0	73.28	46.09
V	$V \circ \mathbb{A}_V$	2.0	73.28	0.24
V	A	4.0	73.28	34.14
V	I	4.0	73.28	44.44
V	R	4.0	73.28	32.11
V	$V \circ \mathbb{A}_V$	4.0	73.28	0.18
V	A	8.0	73.28	12.73
V	I	8.0	73.28	28.68
V	R	8.0	73.28	18.94
V	$V \circ \mathbb{A}_V$	8.0	73.28	0.12
$A \circ \mathbb{A}_A$	A	0.5	56.38	54.61
$A \circ \mathbb{A}_A$	A	1.0	56.38	51.96

Table 5: Accuracy for BIM attacks when source and target network differ.

Source Network	Target Network	ϵ	top-1 acc. (%)	... (adversarial)
$A \circ \mathbb{A}_A$	A	2.0	56.38	46.29
$A \circ \mathbb{A}_A$	A	4.0	56.38	34.28
$A \circ \mathbb{A}_A$	A	8.0	56.38	14.31
$I \circ \mathbb{A}_I$	I	0.5	76.79	74.27
$I \circ \mathbb{A}_I$	I	1.0	76.79	71.67
$I \circ \mathbb{A}_I$	I	2.0	76.79	67.21
$I \circ \mathbb{A}_I$	I	4.0	76.79	58.85
$I \circ \mathbb{A}_I$	I	8.0	76.79	46.60
$R \circ \mathbb{A}_R$	R	0.5	75.00	72.31
$R \circ \mathbb{A}_R$	R	1.0	75.00	68.83
$R \circ \mathbb{A}_R$	R	2.0	75.00	63.07
$R \circ \mathbb{A}_R$	R	4.0	75.00	53.10
$R \circ \mathbb{A}_R$	R	8.0	75.00	36.36
$V \circ \mathbb{A}_V$	V	0.5	68.14	64.68
$V \circ \mathbb{A}_V$	V	1.0	68.14	56.26
$V \circ \mathbb{A}_V$	V	2.0	68.14	43.38
$V \circ \mathbb{A}_V$	V	4.0	68.14	27.12
$V \circ \mathbb{A}_V$	V	8.0	68.14	13.25
$R \circ \mathbb{A}_S$	R	0.5	72.07	63.43
$R \circ \mathbb{A}_S$	R	1.0	72.07	51.71
$R \circ \mathbb{A}_S$	R	2.0	72.07	35.85
$R \circ \mathbb{A}_S$	R	4.0	72.07	20.81
$R \circ \mathbb{A}_S$	R	8.0	72.07	11.40

Table 6: Accuracy for CW attacks when source and target network differ.

Source Network	Target Network	ϵ	top-1 acc. (%)	... (adversarial)
I	$I \circ \mathbb{A}_I$	0.5	76.86	64.36
I	$I \circ \mathbb{A}_I$	1.0	76.86	43.53
I	$I \circ \mathbb{A}_I$	2.0	76.86	28.18
I	$I \circ \mathbb{A}_I$	4.0	76.86	20.69
R	$R \circ \mathbb{A}_R$	0.5	76.03	53.83
R	$R \circ \mathbb{A}_R$	1.0	76.03	28.23
R	$R \circ \mathbb{A}_R$	2.0	76.03	12.76
R	$R \circ \mathbb{A}_R$	4.0	76.03	8.50
$I \circ \mathbb{A}_I$	I	0.5	76.80	72.90
$I \circ \mathbb{A}_I$	I	1.0	76.80	70.08
$I \circ \mathbb{A}_I$	I	2.0	76.80	65.54
$I \circ \mathbb{A}_I$	I	4.0	76.80	59.46

Table 6: Accuracy for CW attacks when source and target network differ.

Source Network	Target Network	ϵ	top-1 acc. (%)	...(adversarial)
$R \circ \mathbb{A}_R$	R	0.5	75.01	71.15
$R \circ \mathbb{A}_R$	R	1.0	75.01	67.98
$R \circ \mathbb{A}_R$	R	2.0	75.01	63.42
$R \circ \mathbb{A}_R$	R	4.0	75.01	57.05

Table 7: BIM attacks on ResNet 50, mitigated with different types and strength (η) of noise.

Noise Type	ϵ	η	top-1 acc. (%)	...(adversarial)
Gaussian	0.5	0.5	76.03	2.77
Gaussian	0.5	1.0	76.03	3.15
Gaussian	0.5	2.0	76.03	4.73
Gaussian	0.5	4.0	76.03	11.15
Gaussian	0.5	8.0	76.03	29.17
Gaussian	0.5	12.0	76.03	41.43
Gaussian	0.5	15.0	76.03	46.46
Gaussian	0.5	20.0	76.03	50.48
Gaussian	0.5	30.0	76.03	48.21
Gaussian	1.0	0.5	76.03	0.45
Gaussian	1.0	1.0	76.03	0.50
Gaussian	1.0	2.0	76.03	0.65
Gaussian	1.0	4.0	76.03	1.88
Gaussian	1.0	8.0	76.03	11.37
Gaussian	1.0	12.0	76.03	24.19
Gaussian	1.0	15.0	76.03	31.81
Gaussian	1.0	20.0	76.03	40.08
Gaussian	1.0	30.0	76.03	43.35
Gaussian	2.0	0.5	76.03	0.08
Gaussian	2.0	1.0	76.03	0.08
Gaussian	2.0	2.0	76.03	0.10
Gaussian	2.0	4.0	76.03	0.16
Gaussian	2.0	8.0	76.03	1.05
Gaussian	2.0	12.0	76.03	5.00
Gaussian	2.0	15.0	76.03	10.38
Gaussian	2.0	20.0	76.03	20.76
Gaussian	2.0	30.0	76.03	32.58
Gaussian	4.0	0.5	76.03	0.01
Gaussian	4.0	1.0	76.03	0.01
Gaussian	4.0	2.0	76.03	0.01
Gaussian	4.0	4.0	76.03	0.02

Table 7: BIM attacks on ResNet 50, mitigated with different types and strength (η) of noise.

Noise Type	ϵ	η	top-1 acc. (%)	... (adversarial)
Gaussian	4.0	8.0	76.03	0.04
Gaussian	4.0	12.0	76.03	0.10
Gaussian	4.0	15.0	76.03	0.30
Gaussian	4.0	20.0	76.03	1.85
Gaussian	4.0	30.0	76.03	11.53
Gaussian	8.0	0.5	76.03	0.00
Gaussian	8.0	1.0	76.03	0.00
Gaussian	8.0	2.0	76.03	0.00
Gaussian	8.0	4.0	76.03	0.00
Gaussian	8.0	8.0	76.03	0.00
Gaussian	8.0	12.0	76.03	0.00
Gaussian	8.0	15.0	76.03	0.01
Gaussian	8.0	20.0	76.03	0.01
Gaussian	8.0	30.0	76.03	0.26
Sign	0.5	0.5	76.03	2.77
Sign	0.5	1.0	76.03	3.14
Sign	0.5	2.0	76.03	4.81
Sign	0.5	4.0	76.03	11.25
Sign	0.5	8.0	76.03	29.59
Sign	0.5	12.0	76.03	41.88
Sign	0.5	15.0	76.03	47.01
Sign	0.5	20.0	76.03	50.73
Sign	0.5	30.0	76.03	47.74
Sign	1.0	0.5	76.03	0.45
Sign	1.0	1.0	76.03	0.47
Sign	1.0	2.0	76.03	0.65
Sign	1.0	4.0	76.03	1.94
Sign	1.0	8.0	76.03	11.61
Sign	1.0	12.0	76.03	24.76
Sign	1.0	15.0	76.03	32.74
Sign	1.0	20.0	76.03	40.71
Sign	1.0	30.0	76.03	43.30
Sign	2.0	0.5	76.03	0.08
Sign	2.0	1.0	76.03	0.08
Sign	2.0	2.0	76.03	0.09
Sign	2.0	4.0	76.03	0.16
Sign	2.0	8.0	76.03	1.13
Sign	2.0	12.0	76.03	5.38
Sign	2.0	15.0	76.03	11.19
Sign	2.0	20.0	76.03	21.74
Sign	2.0	30.0	76.03	33.49

Table 7: BIM attacks on ResNet 50, mitigated with different types and strength (η) of noise.

Noise Type	ϵ	η	top-1 acc. (%)	... (adversarial)
Sign	4.0	0.5	76.03	0.01
Sign	4.0	1.0	76.03	0.01
Sign	4.0	2.0	76.03	0.01
Sign	4.0	4.0	76.03	0.02
Sign	4.0	8.0	76.03	0.04
Sign	4.0	12.0	76.03	0.12
Sign	4.0	15.0	76.03	0.38
Sign	4.0	20.0	76.03	2.25
Sign	4.0	30.0	76.03	13.05
Sign	8.0	0.5	76.03	0.00
Sign	8.0	1.0	76.03	0.00
Sign	8.0	2.0	76.03	0.00
Sign	8.0	4.0	76.03	0.00
Sign	8.0	8.0	76.03	0.00
Sign	8.0	12.0	76.03	0.00
Sign	8.0	15.0	76.03	0.01
Sign	8.0	20.0	76.03	0.01
Sign	8.0	30.0	76.03	0.44
Uniform	0.5	0.5	76.03	2.72
Uniform	0.5	1.0	76.03	2.81
Uniform	0.5	2.0	76.03	3.37
Uniform	0.5	4.0	76.03	5.57
Uniform	0.5	8.0	76.03	13.82
Uniform	0.5	12.0	76.03	24.69
Uniform	0.5	15.0	76.03	31.86
Uniform	0.5	20.0	76.03	40.42
Uniform	0.5	30.0	76.03	49.01
Uniform	1.0	0.5	76.03	0.44
Uniform	1.0	1.0	76.03	0.44
Uniform	1.0	2.0	76.03	0.49
Uniform	1.0	4.0	76.03	0.76
Uniform	1.0	8.0	76.03	2.69
Uniform	1.0	12.0	76.03	8.04
Uniform	1.0	15.0	76.03	13.55
Uniform	1.0	20.0	76.03	23.04
Uniform	1.0	30.0	76.03	37.01
Uniform	2.0	0.5	76.03	0.08
Uniform	2.0	1.0	76.03	0.08
Uniform	2.0	2.0	76.03	0.08
Uniform	2.0	4.0	76.03	0.10
Uniform	2.0	8.0	76.03	0.20

Table 7: BIM attacks on ResNet 50, mitigated with different types and strength (η) of noise.

Noise Type	ϵ	η	top-1 acc. (%)	... (adversarial)
Uniform	2.0	12.0	76.03	0.63
Uniform	2.0	15.0	76.03	1.43
Uniform	2.0	20.0	76.03	4.51
Uniform	2.0	30.0	76.03	15.73
Uniform	4.0	0.5	76.03	0.01
Uniform	4.0	1.0	76.03	0.01
Uniform	4.0	2.0	76.03	0.01
Uniform	4.0	4.0	76.03	0.01
Uniform	4.0	8.0	76.03	0.02
Uniform	4.0	12.0	76.03	0.03
Uniform	4.0	15.0	76.03	0.04
Uniform	4.0	20.0	76.03	0.09
Uniform	4.0	30.0	76.03	0.83
Uniform	8.0	0.5	76.03	0.00
Uniform	8.0	1.0	76.03	0.00
Uniform	8.0	2.0	76.03	0.00
Uniform	8.0	4.0	76.03	0.00
Uniform	8.0	8.0	76.03	0.00
Uniform	8.0	12.0	76.03	0.00
Uniform	8.0	15.0	76.03	0.00
Uniform	8.0	20.0	76.03	0.00
Uniform	8.0	30.0	76.03	0.01