DataFreeShield: Defending Adversarial Attacks without Training Data

Hyeyoon Lee[1], Kanghyun Choi[1], Dain Kwon[1], Sunjong Park[1], Mayoore Selvarasa Jaiswal[2], Noseong Park[3], Jonghyun Choi[1], Jinho Lee[1]

Seoul National University[1] NVIDIA[2] KAIST[3]

Adversarial Attacks

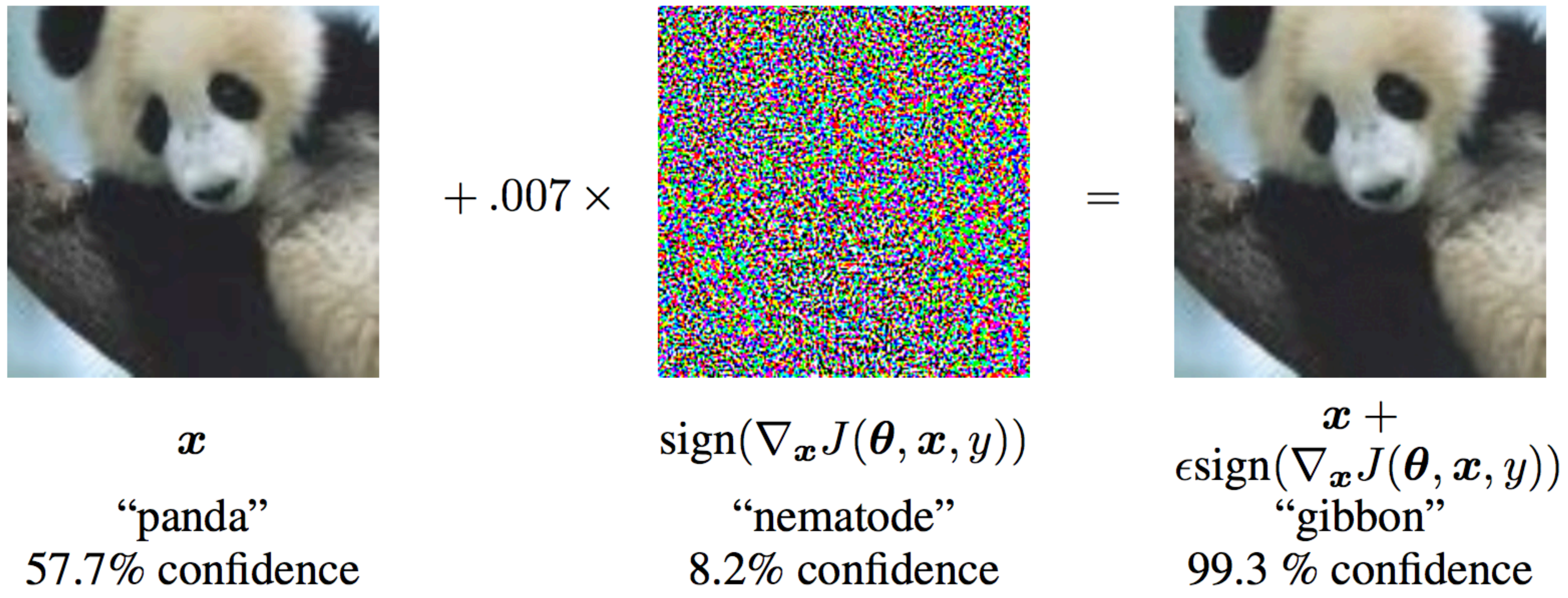
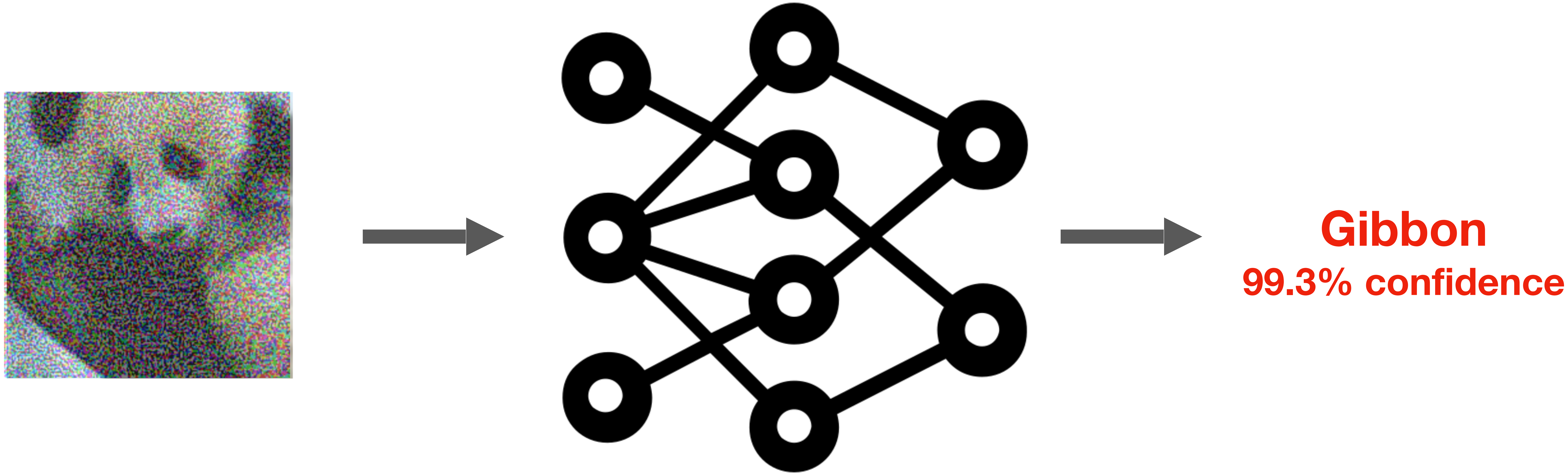


Figure: Picture from Goodfellow et al. (2015)

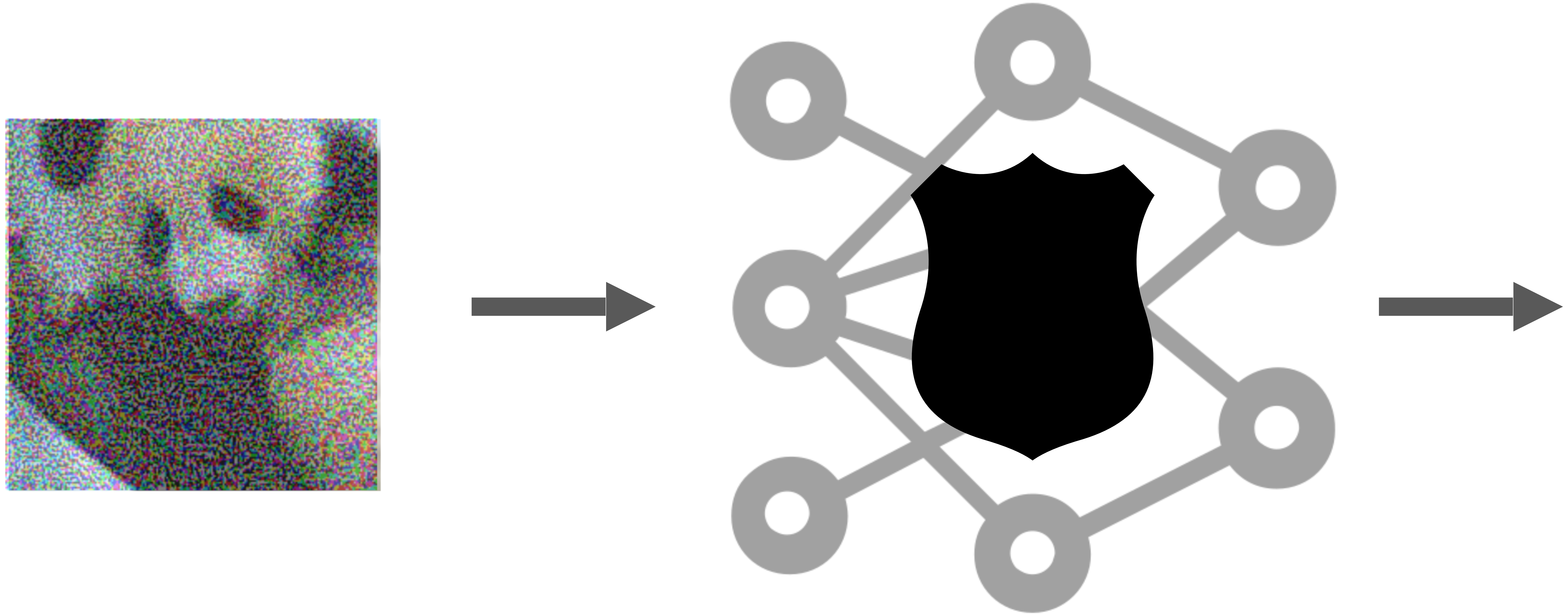
- A small perturbation to the input can cause misclassification to a well-trained neural network.

Adversarial Attacks



- A small perturbation to the input can cause misclassification to a well-trained neural network.

Adversarial Attacks

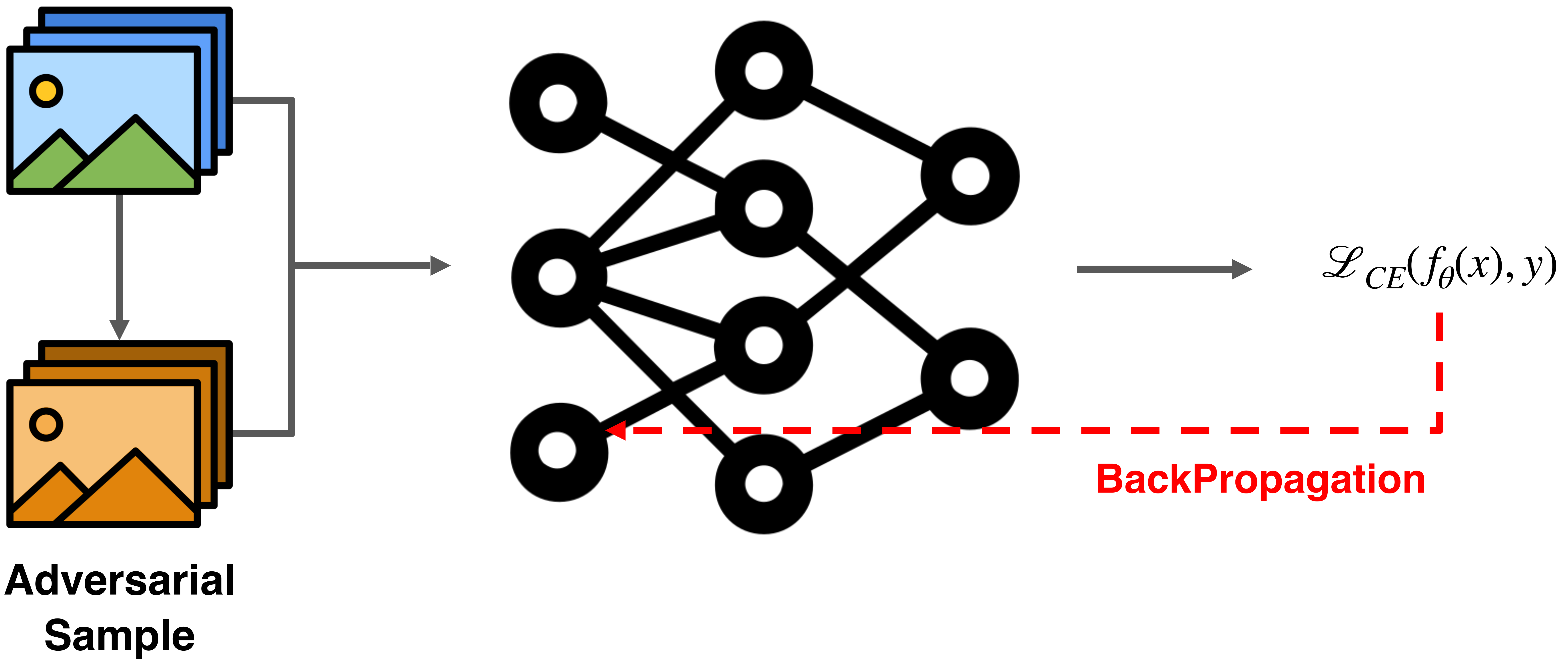


• A sm… …work.

How to defend against these attacks?

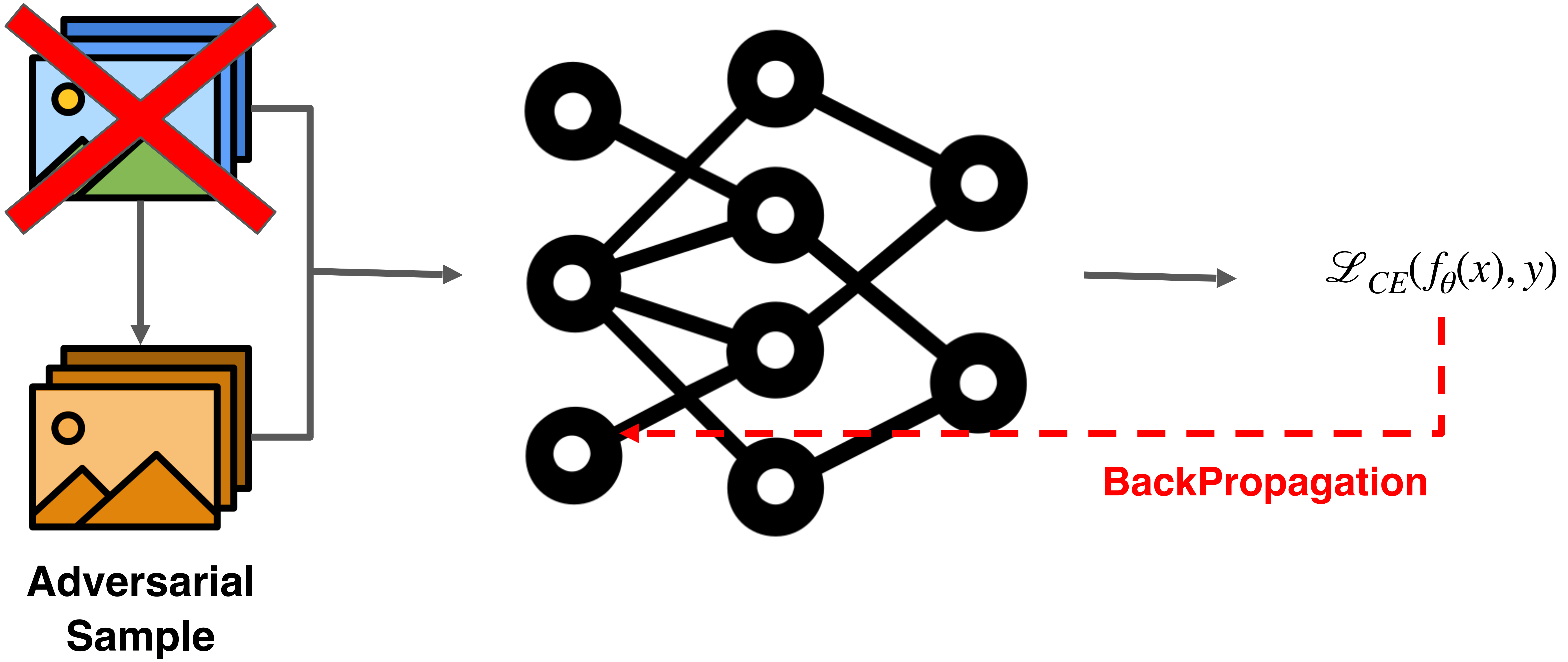
Adversarial Training

Given a pretrained model, how can we transform it to a robust one?



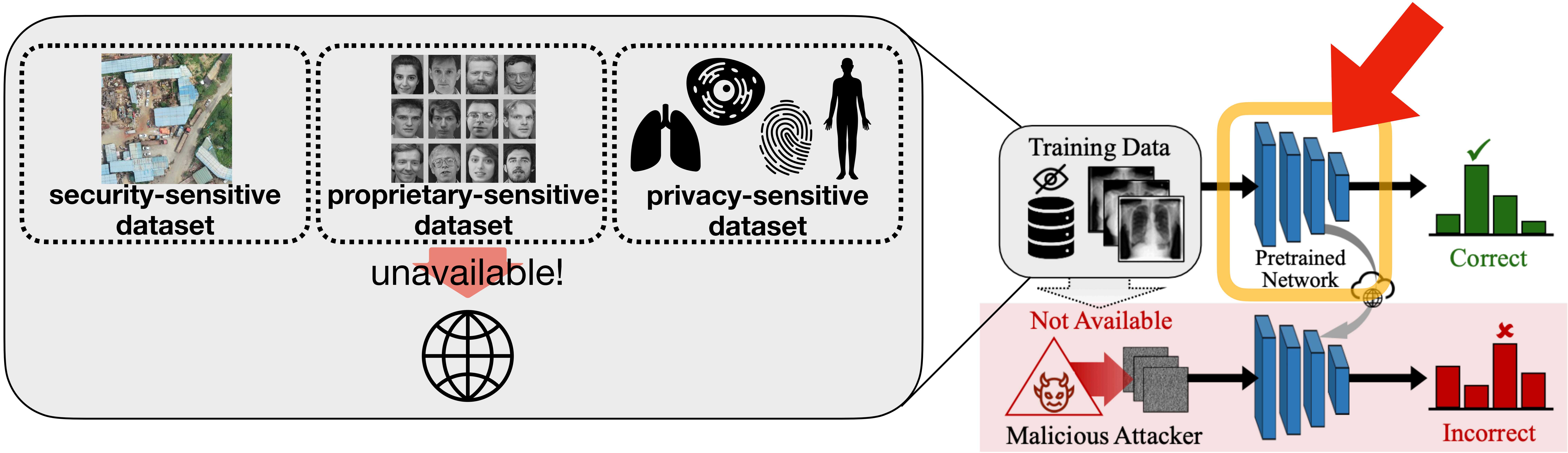
Adversarial Training

Given a pretrained model, how can we transform it to a robust one
with no access to train data?



Problem Scenario

Why the need to achieve robustness “data-free”?



- Training data is kept private for privacy / security / proprietary reasons.
- Attack vulnerability exists in most vanilla-trained DNNs.
- However, existing methods for robustness naturally assumes train dataset is always available. (Unrealistic)

Problem Scenario

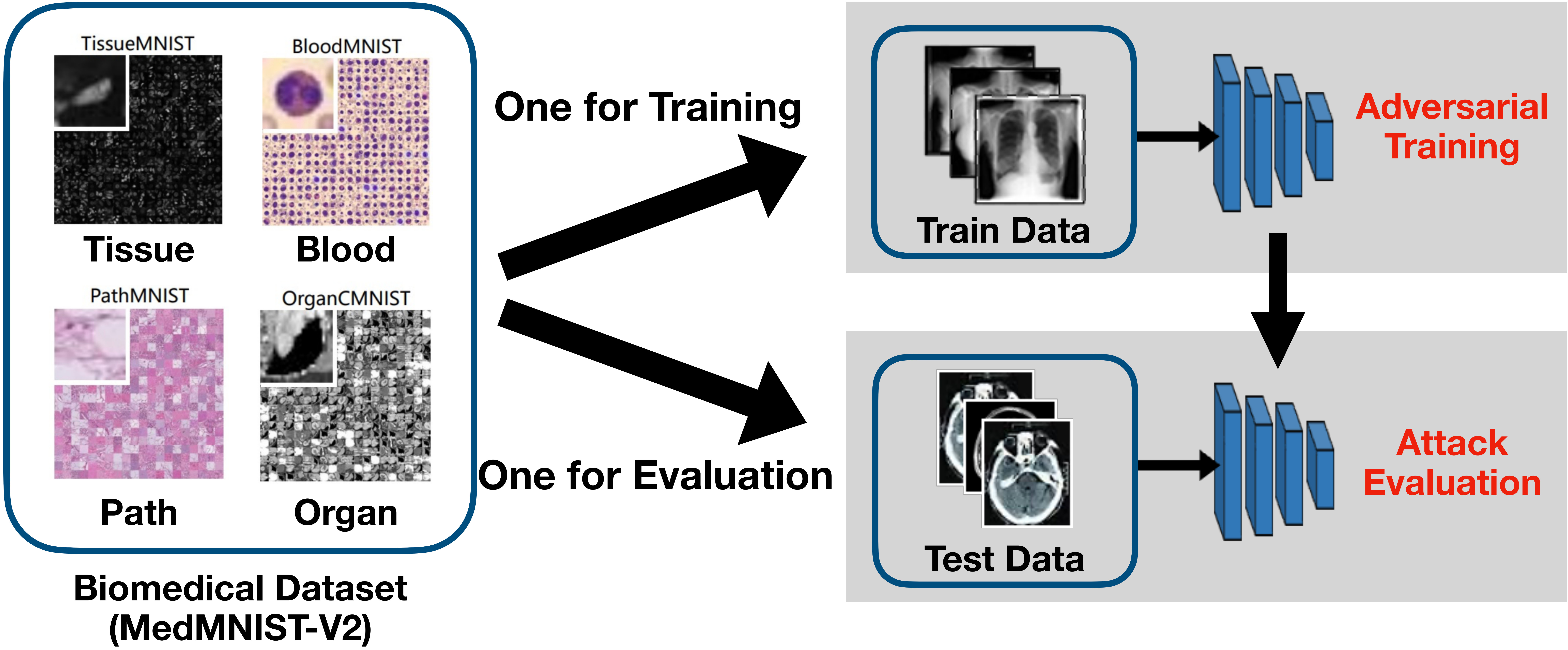
Why the need to achieve robustness “data-free”?

“Why can’t we just use another dataset?”

- Training data is kept private for privacy / security / proprietary reasons.
- Attack vulnerability exists in all vanilla-trained DNNs.
- However, existing methods for robustness naturally assumes train dataset is always available. (Unrealistic)

Motivational Experiment

Using an alternative dataset



Motivational Experiment

Using an alternative dataset

Used for adversarial training

Used for attack

	Tissue	Blood	Path	Organ	CIFAR10
Tissue	37.53	0.00	23.69	8.69	0.02
Blood	9.09	71.94	18.18	0.35	9.09
Path	0.44	0.44	52.53	12.16	0.00
Organ	10.27	23.23	25.82	81.06	40.10

Train Data == Attack Data

PGD-10 ($\epsilon = 8/255$)

- Conventional AT becomes ineffective without the original dataset.

Motivational Experiment

Using an alternative dataset

Used for adversarial training

Used for attack

	Tissue	Blood	Path	Organ	CIFAR10
Tissue	37.53	0.00	23.69	8.69	0.02
Blood	9.09	71.94	18.18	0.35	9.09
Path	0.44	0.44	52.53	12.16	0.00
Organ	10.27	23.23	25.82	81.06	40.10

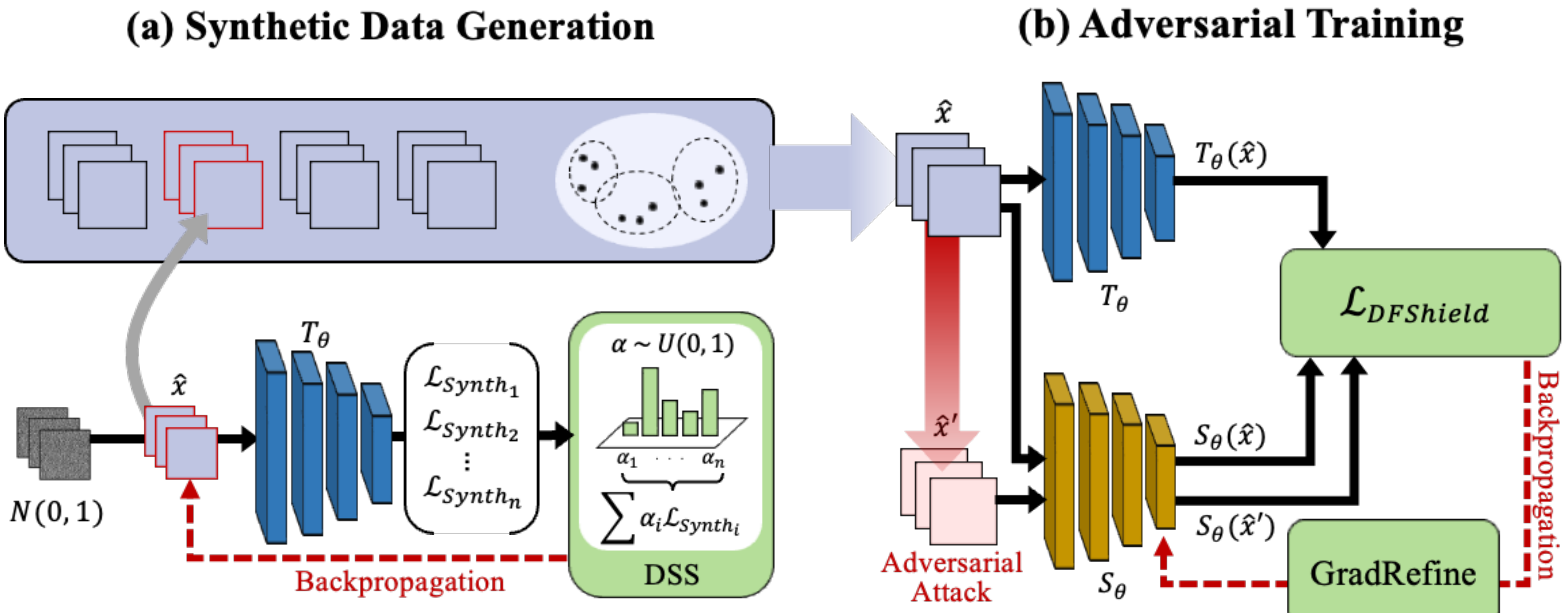
PGD-10 ($\epsilon = 8/255$)

Little to no robustness!

- Conventional AT becomes ineffective without the original dataset.

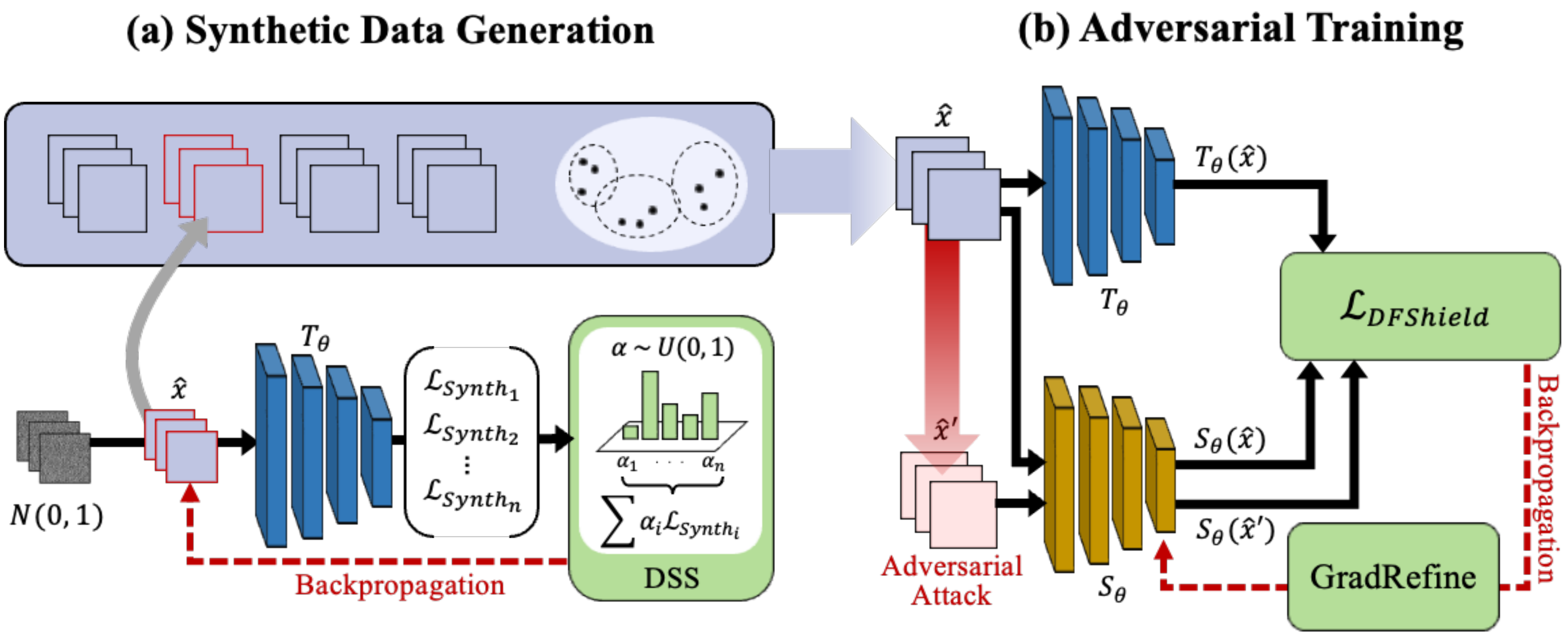
Proposed Method

Overall Procedure



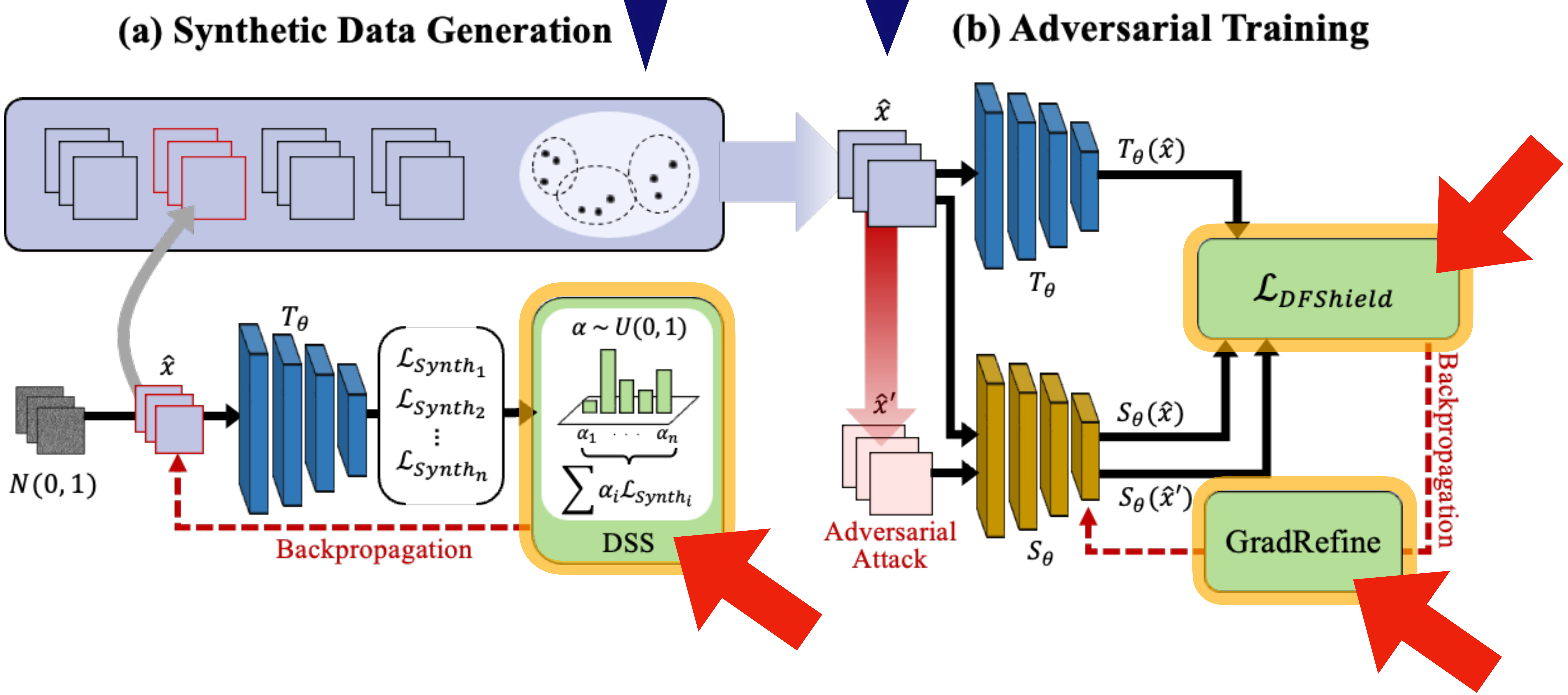
Proposed Method

Overall Procedure

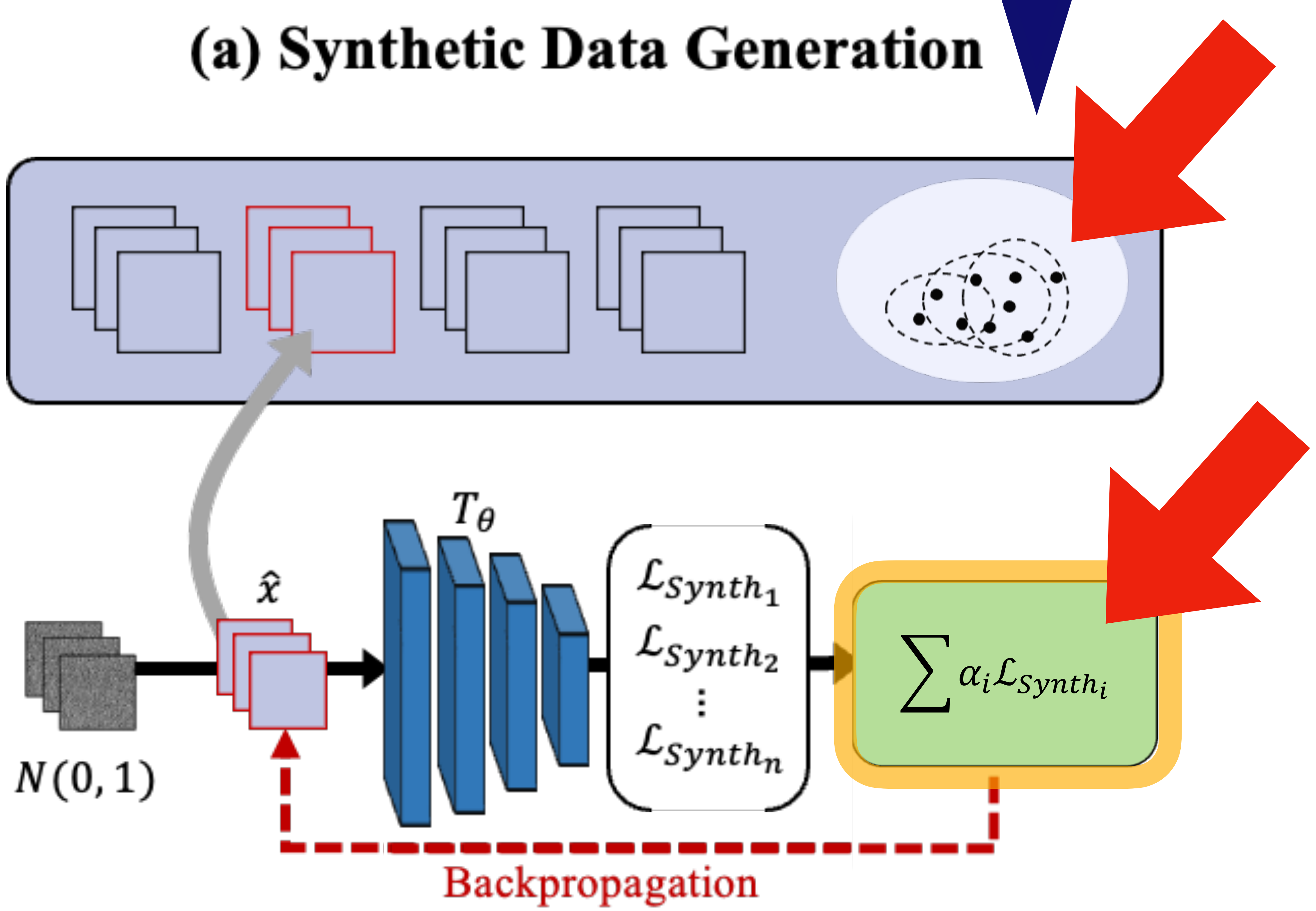


Key Challenge 1: Limited Diversity

Key Challenge 2: Poor Generalization to Real Adversarial Samples



Key Challenge 1: Limited Diversity



Inter-batch diversity

Given a set of synthesis loss functions,

$$\mathbb{S} = \{\mathcal{L}_{Synth_1}, \mathcal{L}_{Synth_2}, \ldots, \mathcal{L}_{Synth_n}\}$$

Conventional approach:

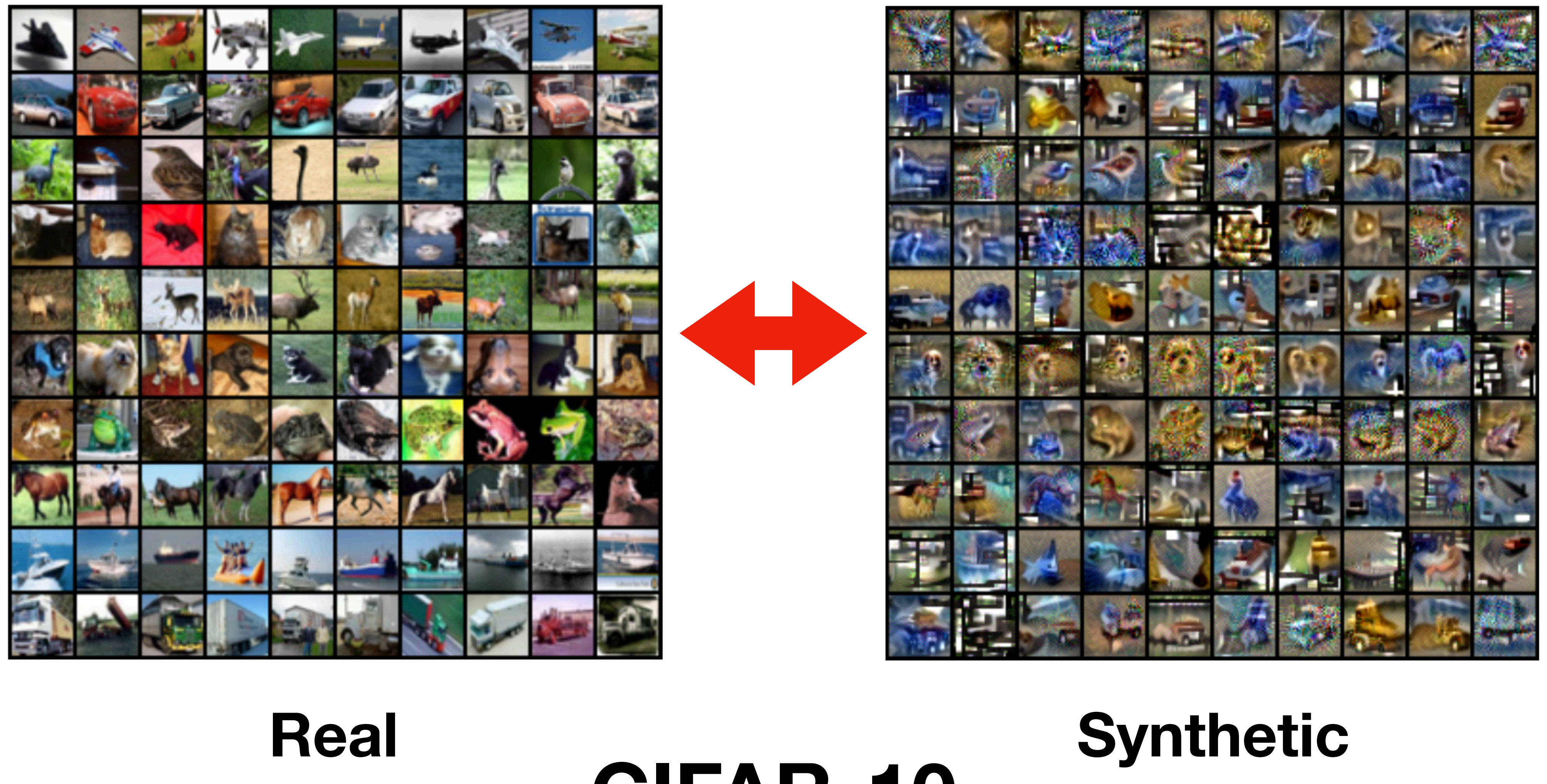
$$\mathcal{L}_{Synth} = \alpha_1 \mathcal{L}_{synth_1} + \alpha_2 \mathcal{L}_{synth_2} + \alpha_3 \mathcal{L}_{synth_3}$$

Diversified Sample Synthesis (DSS)

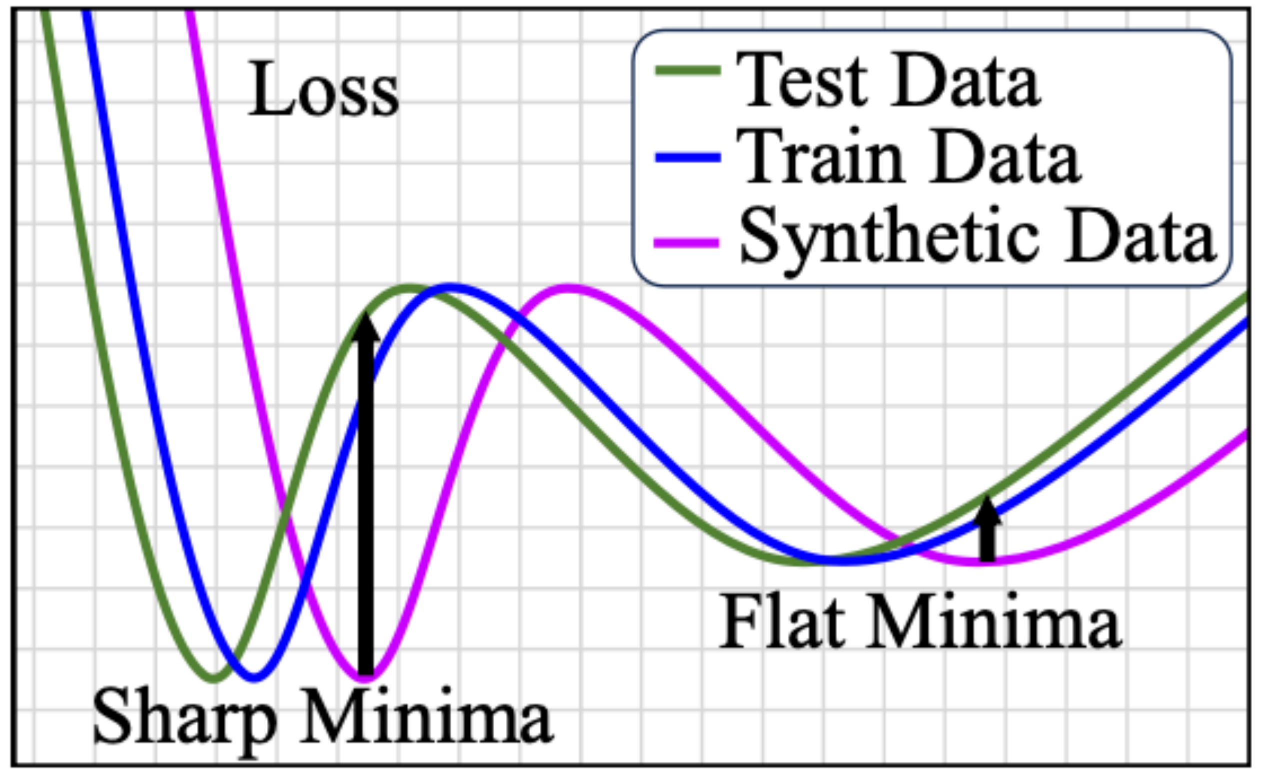
$$\mathcal{L}_{Synth} = \sum_{i=1}^{|\mathbb{S}|} \alpha_i \mathcal{L}_{Synth_i} \quad \alpha_i \sim U(0,1)$$

Dynamically modulate the coefficients for each batch

Key Challenge 2: Poor Generalization to Real Adversarial Samples

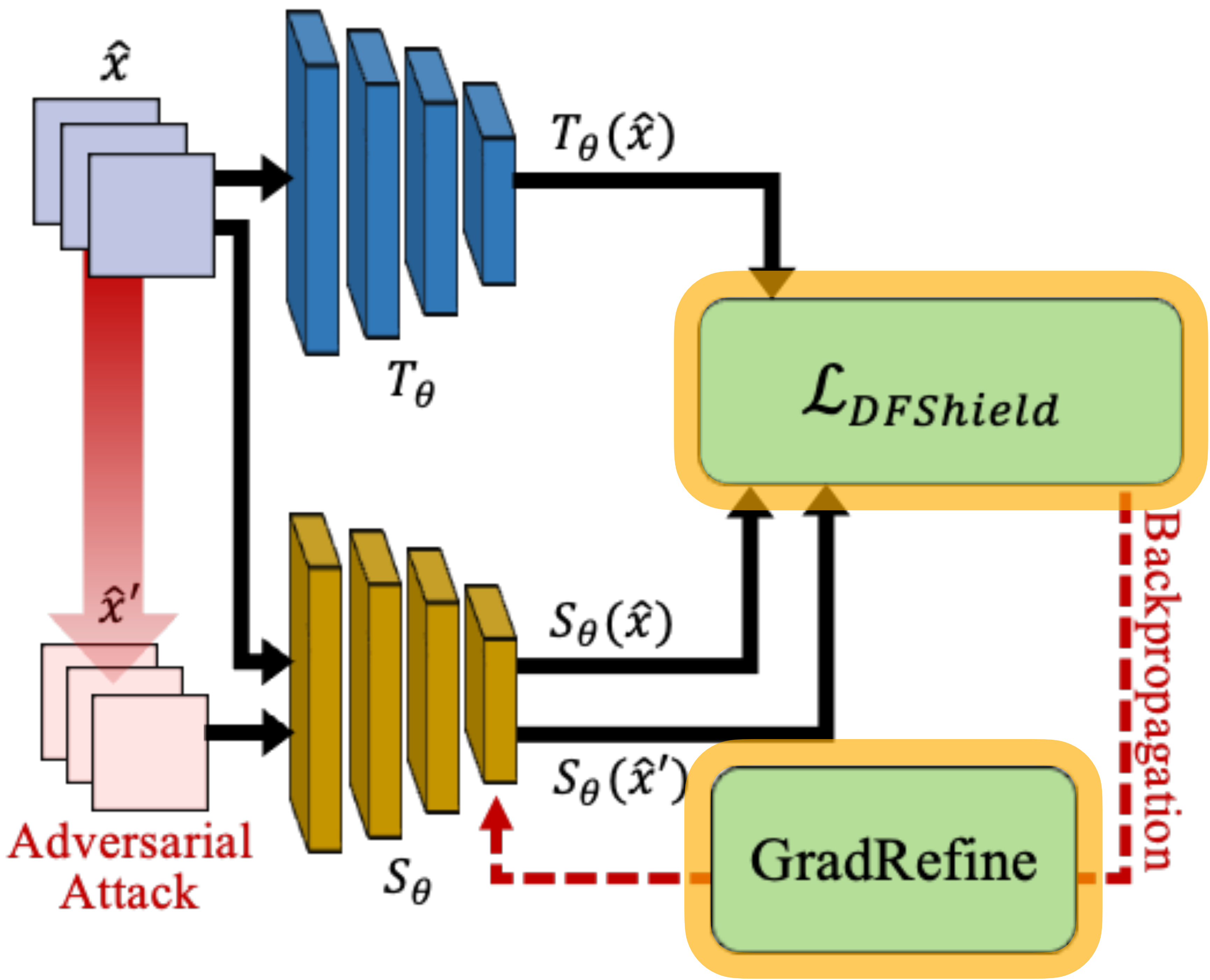


CIFAR-10



Conceptual Diagram of Generalization Gap

(b) Adversarial Training

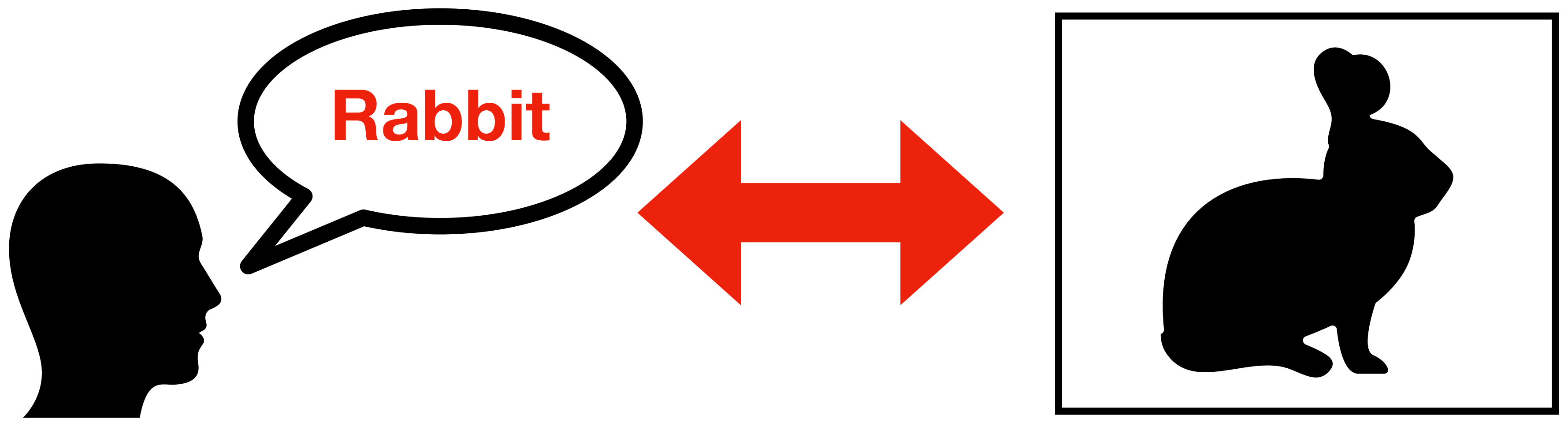


Training loss using soft-guidance only

$$\mathscr{L}_{DFShield} = KL(S(\hat{x}), T(\hat{x})) \quad \textbf{clean accuracy}$$
$$+\lambda_1 KL(S(\hat{x}'), T(\hat{x})) \quad \textbf{robustness training}$$
$$+\lambda_2 KL(S(\hat{x}'), S(\hat{x})) \quad \textbf{smoothness term}$$

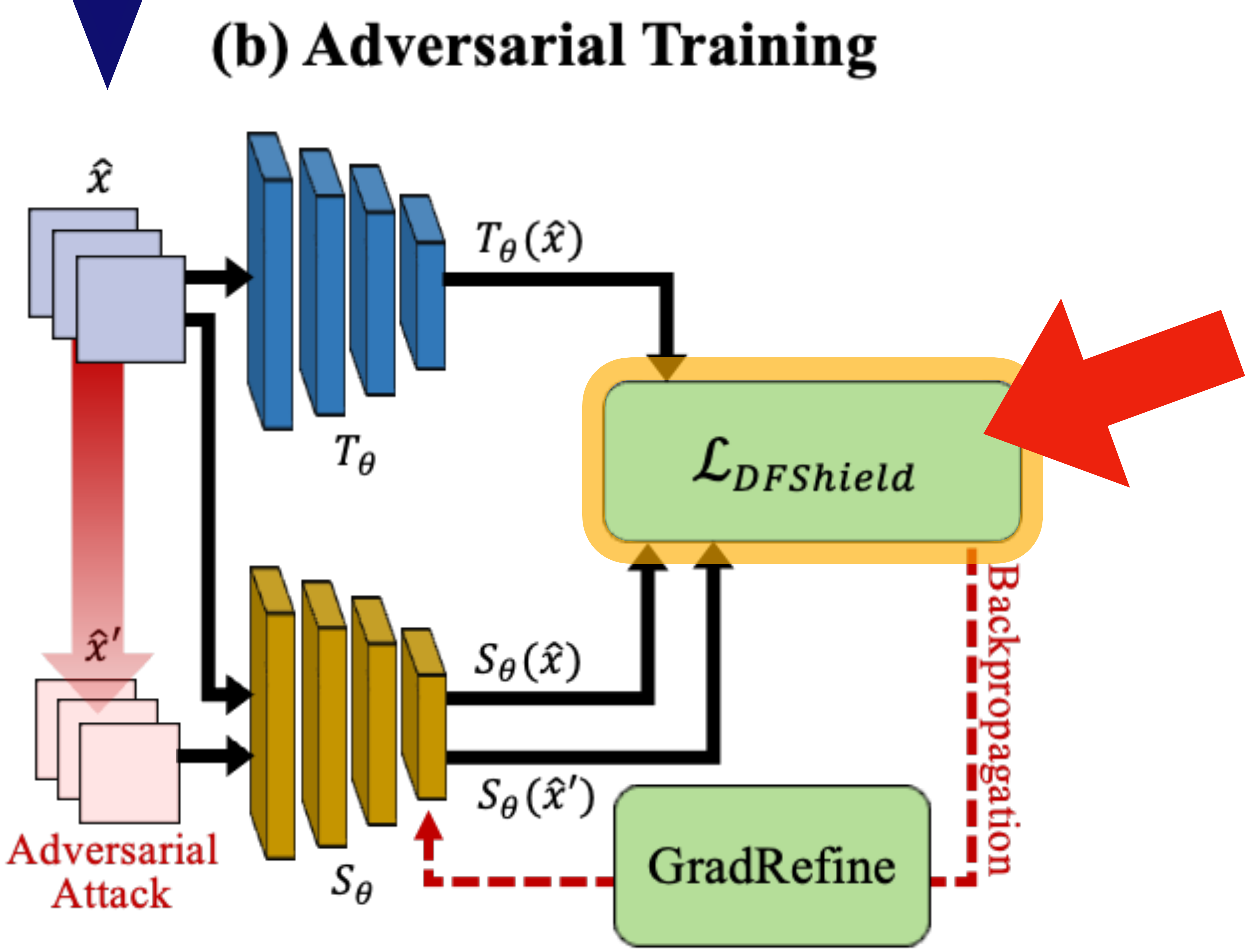
- Artificial labels do not align with human perception.
- Smoothness term helps prevent being overly sensitive to small changes in the input

Human Perception **Artificial Label**



Cat

Key Challenge 2: Poor Generalization to Real Adversarial Samples



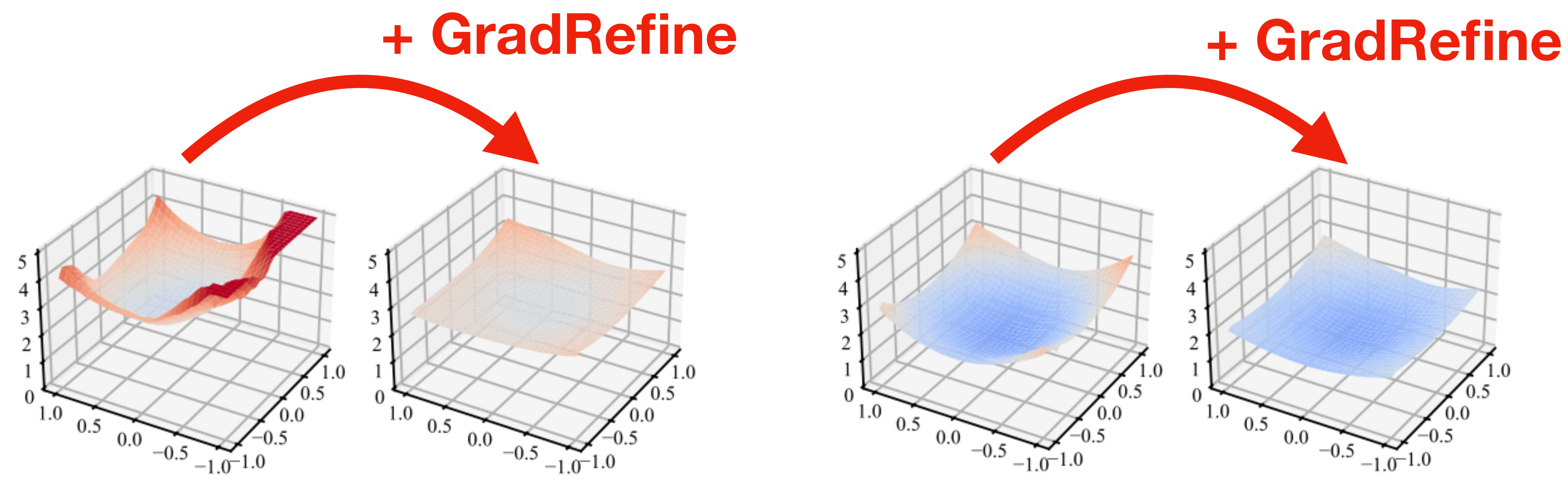
Gradient refinement for smoother loss surface

$$A_k = \frac{1}{B}\sum_{b=1}^{B} sign(g_k^{(b)})$$

- Compute parameter-wise sign agreement score across different batches

$$g_k^* = \Phi(A_k)\sum_{b=1}^{B} 1_{\{A_k \cdot g_k^{(b)}>0\}} \cdot g_k^{(b)}, \quad \Phi(A_k) = \begin{cases} 1, & \textbf{if } |A_k| \geq \tau, \\ 0, & \textbf{otherwise,} \end{cases}$$

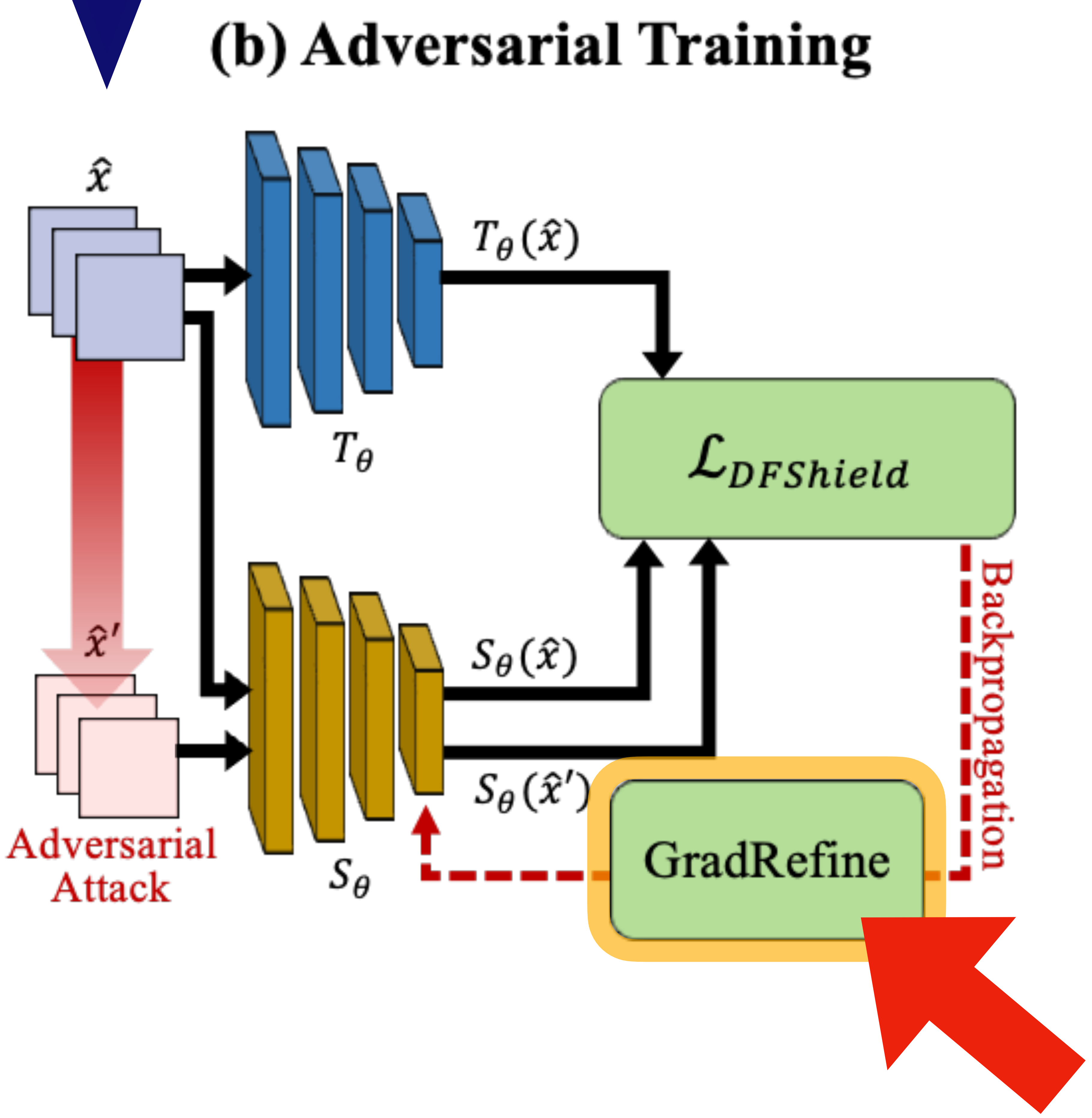
- Mask high-fluctuating parameters before update



TRADES (baseline)

$\mathscr{L}_{DFShield}$ **(ours)**

Key Challenge 2: Poor Generalization to Real Adversarial Samples



Evaluation

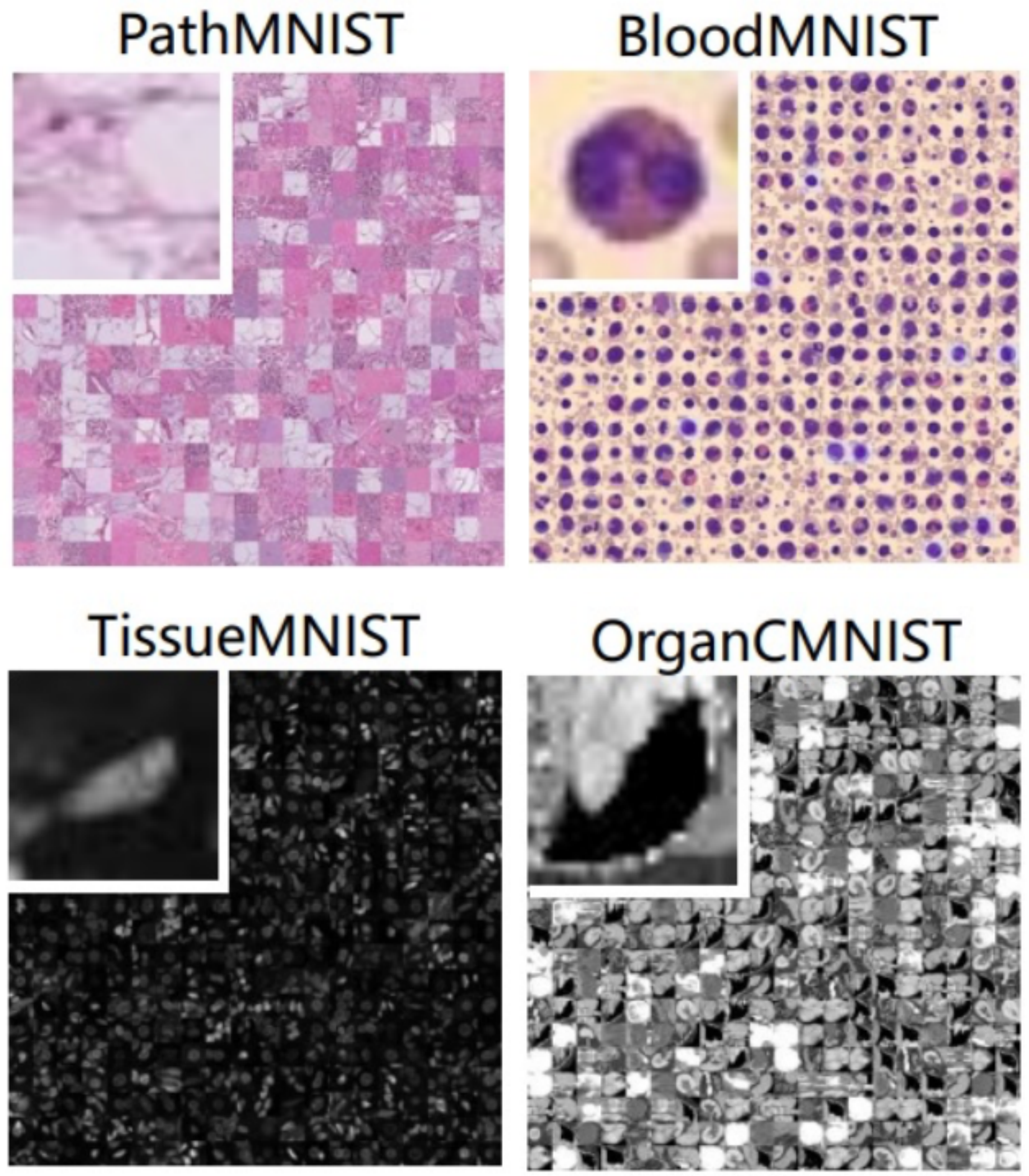
Biomedical Dataset (MedMNIST-V2)

Table 3. Performance on medical datasets with l_∞ perturbation budget using test-time defense methods.

Dataset	Method	ResNet-18			ResNet-50		
		$\mathcal{A}_{Clean}$	$\mathcal{A}_{\text{PGD}}$	$\mathcal{A}_{\text{AA}}$	$\mathcal{A}_{Clean}$	$\mathcal{A}_{\text{PGD}}$	$\mathcal{A}_{\text{AA}}$
Tissue	DAD	55.86	22.90	4.38	59.72	**31.59**	3.49
	DiffPure	26.17	22.85	9.06	27.73	27.54	1.81
	TTE	56.60	0.00	0.00	62.01	0.00	0.00
	Ours	32.07	**31.63**	**31.57**	31.91	27.15	**26.68**
Blood	DAD	91.96	17.25	0.00	83.46	34.43	0.00
	DiffPure	49.02	**29.10**	8.71	51.17	**36.91**	13.77
	TTE	9.09†	9.09	8.92	16.84	0.03	0.00
	Ours	59.89	21.72	**19.29**	74.63	36.07	**30.17**
Path	DAD	91.28	15.54	0.21	81.50	12.79	1.38
	DiffPure	19.73	18.95	8.91	14.65	14.26	**13.79**
	TTE	76.56	0.64	0.36	75.08	4.23	1.88
	Ours	33.06	**29.78**	**25.38**	41.63	**15.35**	12.28
OrganC	DAD	80.19	31.22	12.57	87.54	25.46	7.84
	DiffPure	69.73	**57.03**	19.00	58.20	51.76	34.38
	TTE	61.03	22.90	15.98	56.54	25.82	18.63
	Ours	83.35	47.01	**42.56**	86.56	**62.60**	**59.86**

†Did not converge

upto 25%p difference



Evaluation

Biomedical Dataset (MedMNIST-V2)

Table 1. Performance on medical datasets with l_∞ perturbation budget.

Model	Method	Tissue			Blood			Path			OrganC		
		$\mathcal{A}_{Clean}$	$\mathcal{A}_{\mathbf{PGD}}$	$\mathcal{A}_{\mathbf{AA}}$	$\mathcal{A}_{Clean}$	$\mathcal{A}_{\mathbf{PGD}}$	$\mathcal{A}_{\mathbf{AA}}$	$\mathcal{A}_{Clean}$	$\mathcal{A}_{\mathbf{PGD}}$	$\mathcal{A}_{\mathbf{AA}}$	$\mathcal{A}_{Clean}$	$\mathcal{A}_{\mathbf{PGD}}$	$\mathcal{A}_{\mathbf{AA}}$
RN-18	Public	22.04	0.02	0.00	9.09†	9.09	0.00	13.30†	0.00	0.00	79.41	40.10	36.53
	DaST	23.27	7.01	5.98	16.92	6.75	4.82	7.49	3.36	1.20	83.13	27.91	24.49
	DFME	7.01	4.33	4.17	46.59	0.20	0.03	76.43	0.50	0.38	79.73	19.27	17.19
	AIT	15.62	11.64	9.72	18.24	10.55	1.64	16.66	10.24	3.89	56.85	18.02	16.67
	DFARD	9.31	8.48	1.87	22.60	10.17	9.70	11.59	4.93	3.18	81.97	21.71	19.50
	Ours	32.07	**31.63**	**31.57**	59.89	**21.72**	**19.29**	33.06	**29.78**	**25.38**	83.35	**47.01**	**42.56**
RN-50	Public	27.84	10.11	8.64	9.09†	9.09	0.00	7.54	1.21	0.37	84.41	46.12	43.44
	DaST	4.73	1.36	0.05	9.12	8.77	8.16	8.25	6.92	2.12	21.03	9.18	8.36
	DFME	7.13	6.55	4.76	7.16	3.36	3.19	80.10	2.28	2.01	27.76	22.00	21.78
	AIT	32.08	4.75	0.74	19.47	12.48	9.94	14.29	10.00	2.21	15.34	8.90	6.02
	DFARD	23.69	12.99	7.01	26.63	9.21	0.00	14.04	2.44	0.77	80.99	11.93	8.13
	Ours	31.91	**27.15**	**26.68**	74.63	**36.07**	**30.17**	41.63	**15.35**	**12.28**	86.56	**62.60**	**59.86**

†Did not converge

Evaluation

General Benchmark Dataset

Table 4. Performance on SVHN, CIFAR-10, and CIFAR-100 with l_∞ perturbation budget.

	ResNet-20			ResNet-56			WRN-28-10		
SVHN	$\mathcal{A}_{Clean}$	$\mathcal{A}_{\text{PGD}}$	$\mathcal{A}_{\text{AA}}$	$\mathcal{A}_{Clean}$	$\mathcal{A}_{\text{PGD}}$	$\mathcal{A}_{\text{AA}}$	$\mathcal{A}_{Clean}$	$\mathcal{A}_{\text{PGD}}$	$\mathcal{A}_{\text{AA}}$
DaST	20.66	13.90	7.06	10.55	0.25	0.00	20.15	19.17	14.57
DFME	11.32	2.59	0.84	20.20	19.22	4.27	6.94	5.31	0.28
AIT	91.45	37.87	24.74	86.65	45.45	38.96	83.89	40.45	33.06
DFARD	25.62	18.65	0.19	19.58	15.43	0.00	92.32	13.08	0.01
Ours	91.83	**54.82**	**47.55**	88.66	**62.05**	**57.54**	94.14	**69.60**	**62.66**
CIFAR-10	$\mathcal{A}_{Clean}$	$\mathcal{A}_{\text{PGD}}$	$\mathcal{A}_{\text{AA}}$	$\mathcal{A}_{Clean}$	$\mathcal{A}_{\text{PGD}}$	$\mathcal{A}_{\text{AA}}$	$\mathcal{A}_{Clean}$	$\mathcal{A}_{\text{PGD}}$	$\mathcal{A}_{\text{AA}}$
DaST	10.00†	9.89	8.62	12.06	7.68	5.32	10.00†	9.65	2.85
DFME	14.36	5.23	0.08	13.81	3.92	0.03	10.00†	9.98	0.05
AIT	32.89	11.93	10.67	38.47	12.29	11.36	34.92	10.90	9.47
DFARD	12.28	5.33	0.00	10.84	8.93	0.00	9.82	12.01	0.02
Ours	74.79	**29.29**	**22.65**	81.30	**35.55**	**30.51**	86.74	**51.13**	**43.73**
CIFAR-100	$\mathcal{A}_{Clean}$	$\mathcal{A}_{\text{PGD}}$	$\mathcal{A}_{\text{AA}}$	$\mathcal{A}_{Clean}$	$\mathcal{A}_{\text{PGD}}$	$\mathcal{A}_{\text{AA}}$	$\mathcal{A}_{Clean}$	$\mathcal{A}_{\text{PGD}}$	$\mathcal{A}_{\text{AA}}$
DaST	1.01†	0.99	0.95	1.13	0.72	0.34	1.39	0.66	0.18
DFME	1.86	0.53	0.24	24.16	0.98	0.25	66.30	0.67	0.00
AIT	7.92	2.51	1.39	9.68	2.97	2.04	22.21	3.11	1.28
DFARD	66.59	0.02	0.00	69.20	0.26	0.00	82.03	1.10	0.00
Ours	41.67	**10.41**	**5.97**	39.29	**13.23**	**9.49**	61.35	**23.22**	**16.44**

†Did not converge

- Existing data-free methods fail to achieve meaningful robustness.
- Ours show resistance to both weaker (PGD) and stronger attacks (AA).

DataFreeShield: Defending Adversarial Attacks without Training Data

Thank you!

hylee817@snu.ac.kr