Robust Distillation via Untargeted and Targeted Intermediate Adversarial Samples

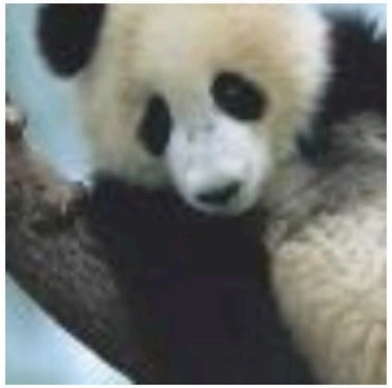
Junhao Dong, Piotr Koniusz, Junxi Chen, Z. Jane Wang, Yew-Soon Ong

Reporter: Junhao Dong

SEATTLE, WA CVPR JUNE 17-21, 2024

Adversarial examples are tailored inputs with the purpose of confusing neural networks. (Visually similar to natural examples)

Natural Examples

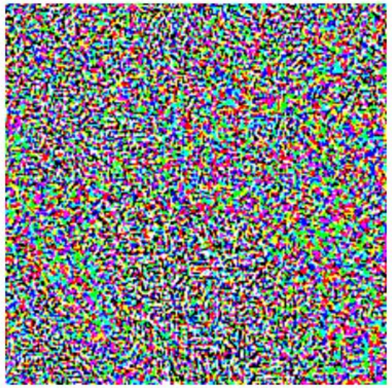


$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$

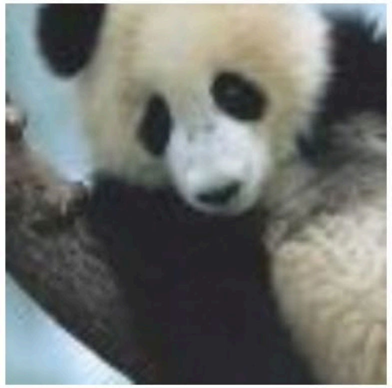
Adversarial Perturbation



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$



Adversarial Examples

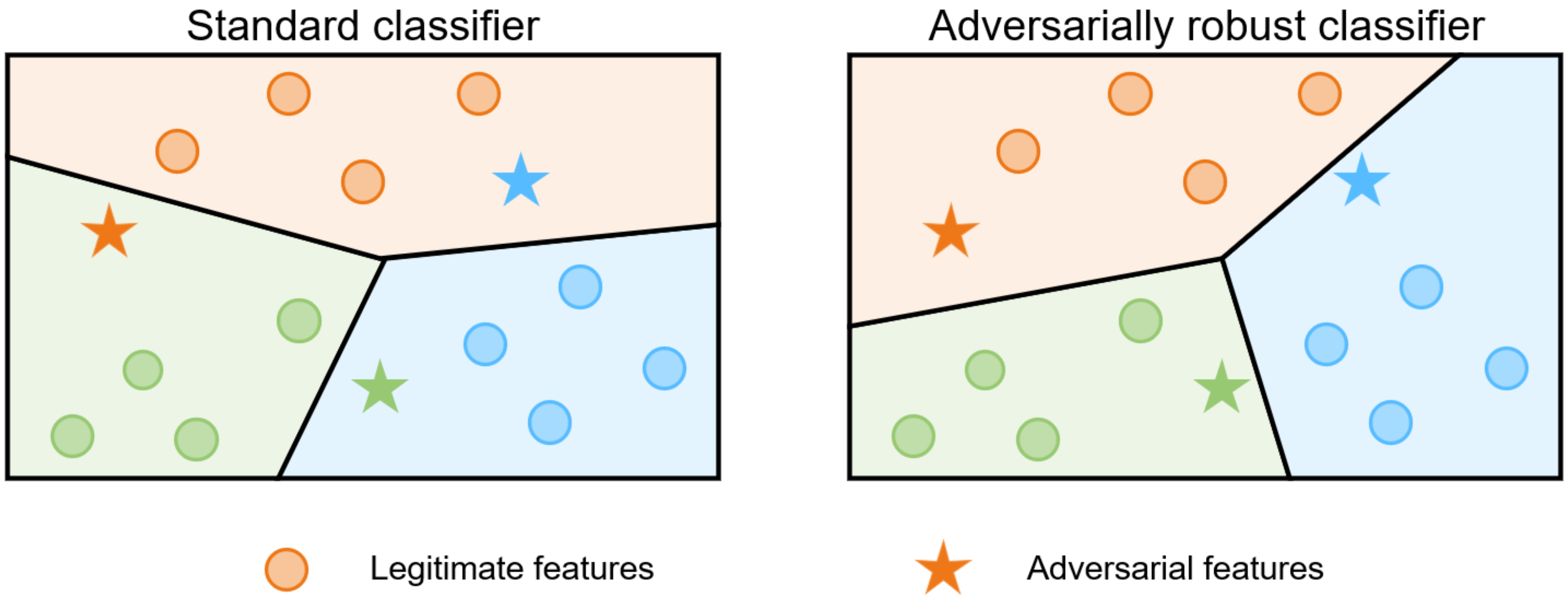
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

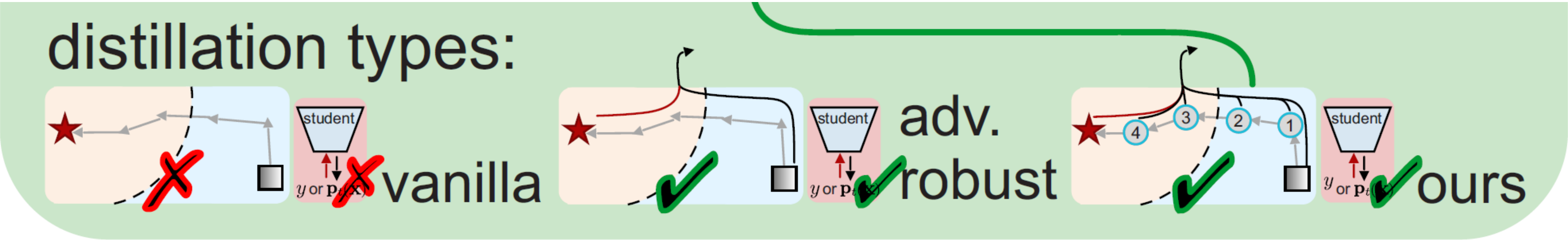
Introducing gradient ascent at the **image level**.

Adversarial Training (min-max optimization):

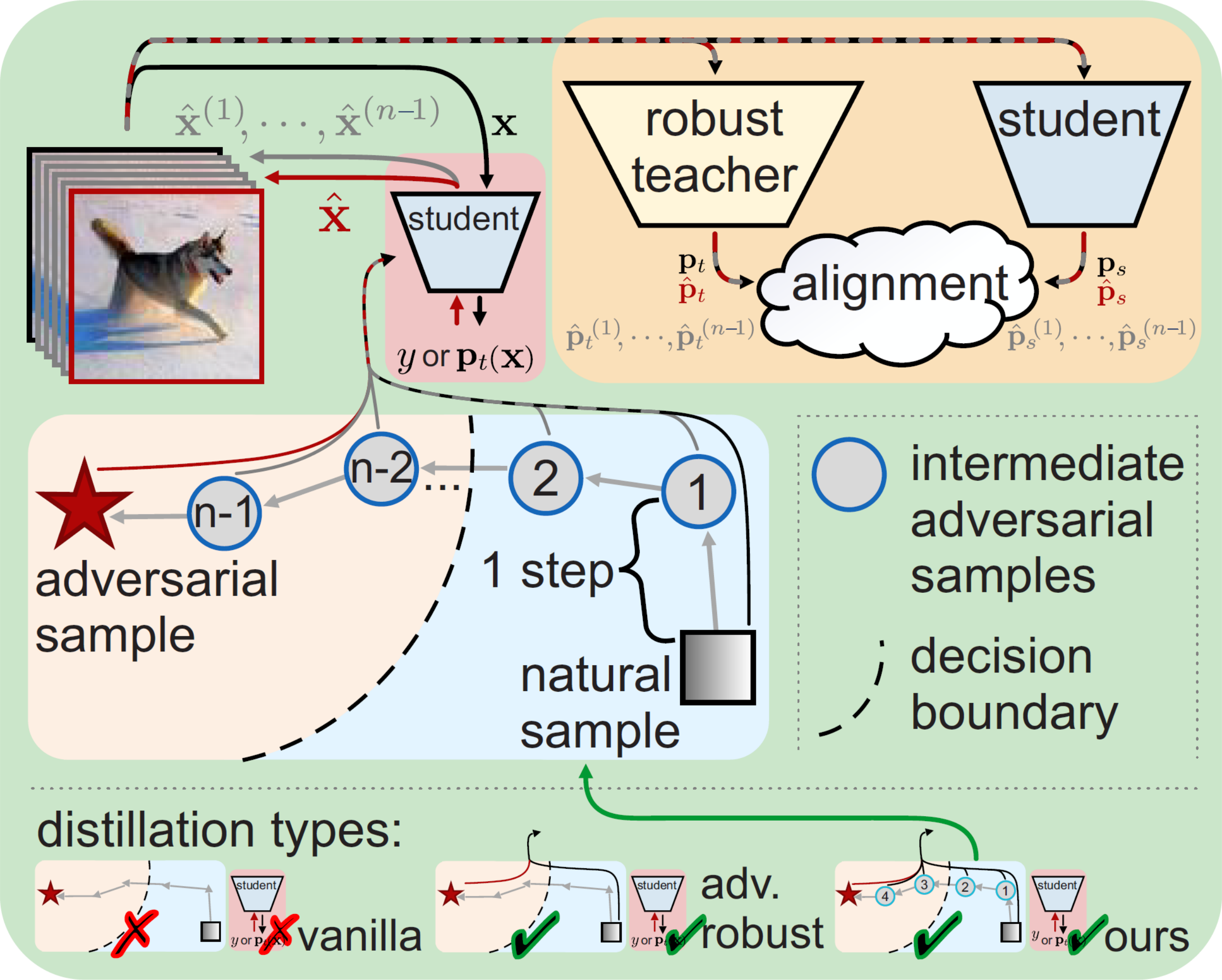
$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\mathcal{L}_{\mathrm{CE}}\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}\right), y\right) + \max_{\|\delta\|_{\infty}<\epsilon} \mathcal{L}_{\mathrm{KL}}\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}\right) \| f_{\boldsymbol{\theta}}\left(\mathbf{x}+\delta\right)\right) \right]$$



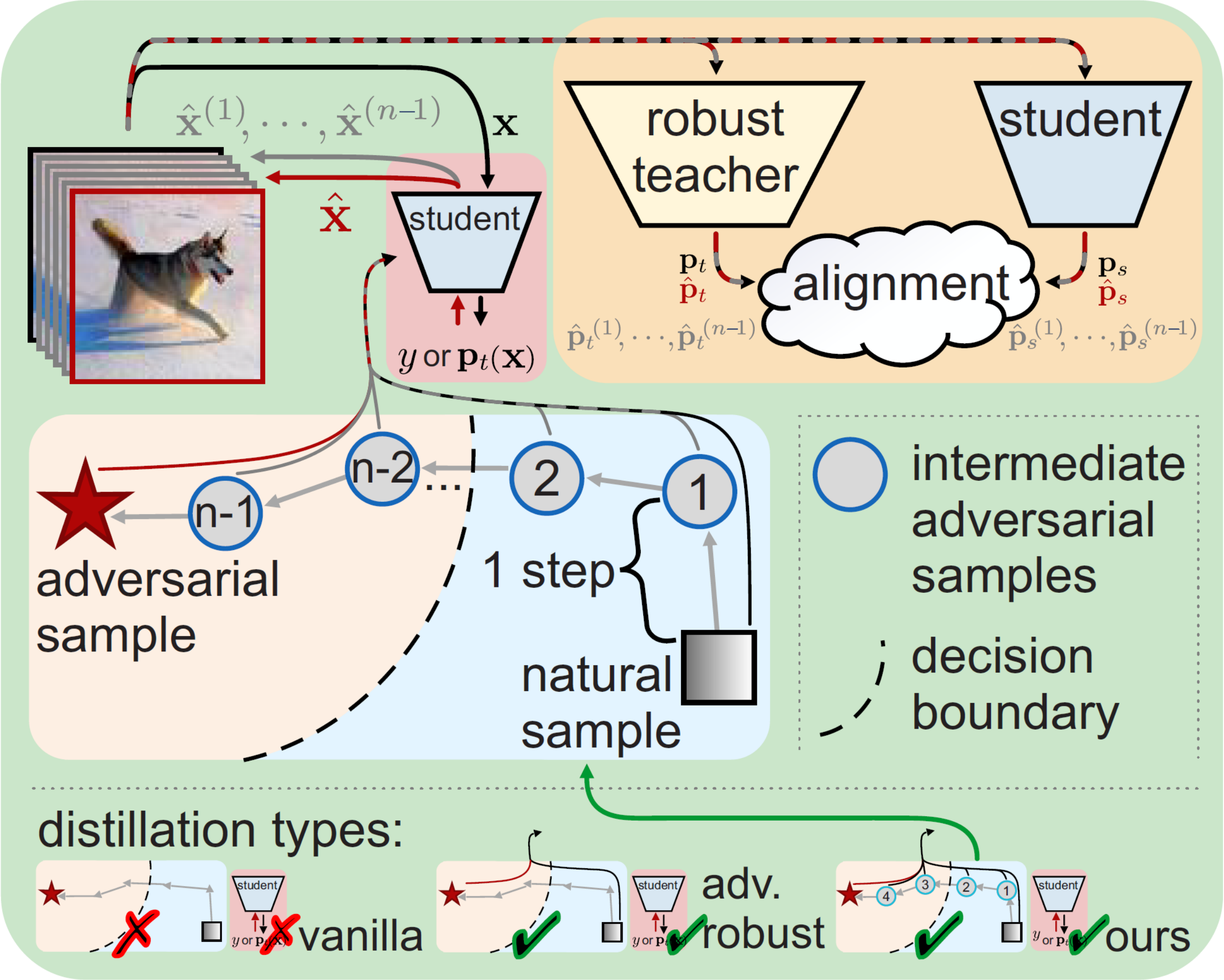
■ Adversarially Robust Knowledge Distillation



Adversarially Robust Knowledge Distillation



Adversarially Robust Knowledge Distillation



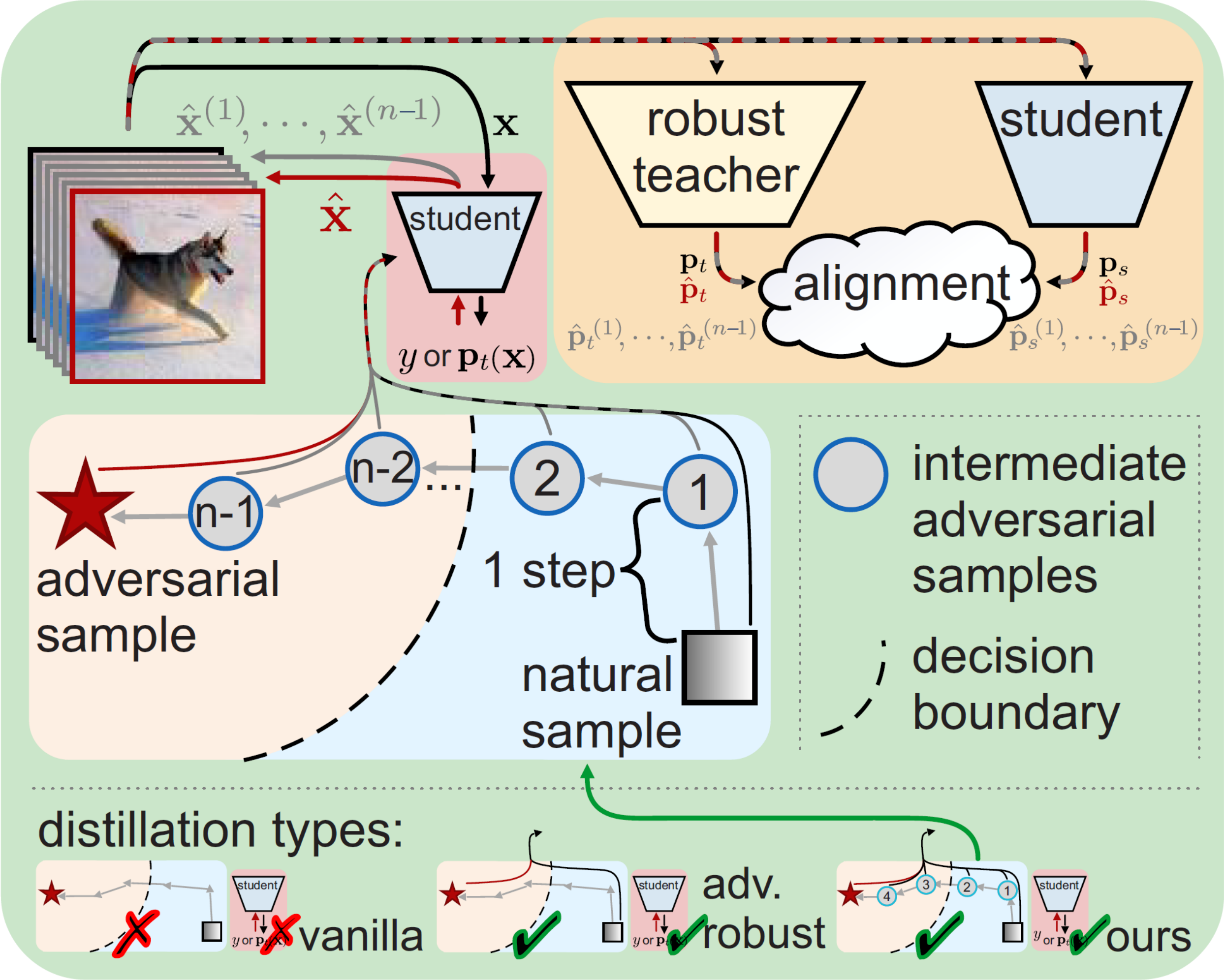
Untargeted Adversary Generation

$$\hat{\mathbf{x}}^{(i+1)} = \vartheta_\alpha(\hat{\mathbf{x}}^{(i)}, y) = \prod_{\mathbb{B}(\mathbf{x},\epsilon)} \left(\hat{\mathbf{x}}^{(i)} + \alpha \cdot \operatorname{sign}\left(\nabla_{\hat{\mathbf{x}}^{(i)}} \mathcal{L}_{\mathrm{CE}}(f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}^{(i)}), y)\right)\right)$$

Prediction Alignment of Clean and Adversarial Samples

$$\mathcal{L}_{\mathrm{ARKD}} = \underbrace{\mathcal{L}_{\mathrm{KL}}\left(f_{\boldsymbol{\theta}_t}(\mathbf{x}) \,\|\, f_{\boldsymbol{\theta}_s}(\mathbf{x})\right)}_{\text{alignment of “natural distributions”}} + \beta \cdot \underbrace{\mathcal{L}_{\mathrm{KL}}\left(f_{\boldsymbol{\theta}_t}(\hat{\mathbf{x}}) \,\|\, f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}})\right)}_{\text{alignment of “adversarial distributions”}}$$

Intermediate Adversarial Knowledge Distillation



Objective Function

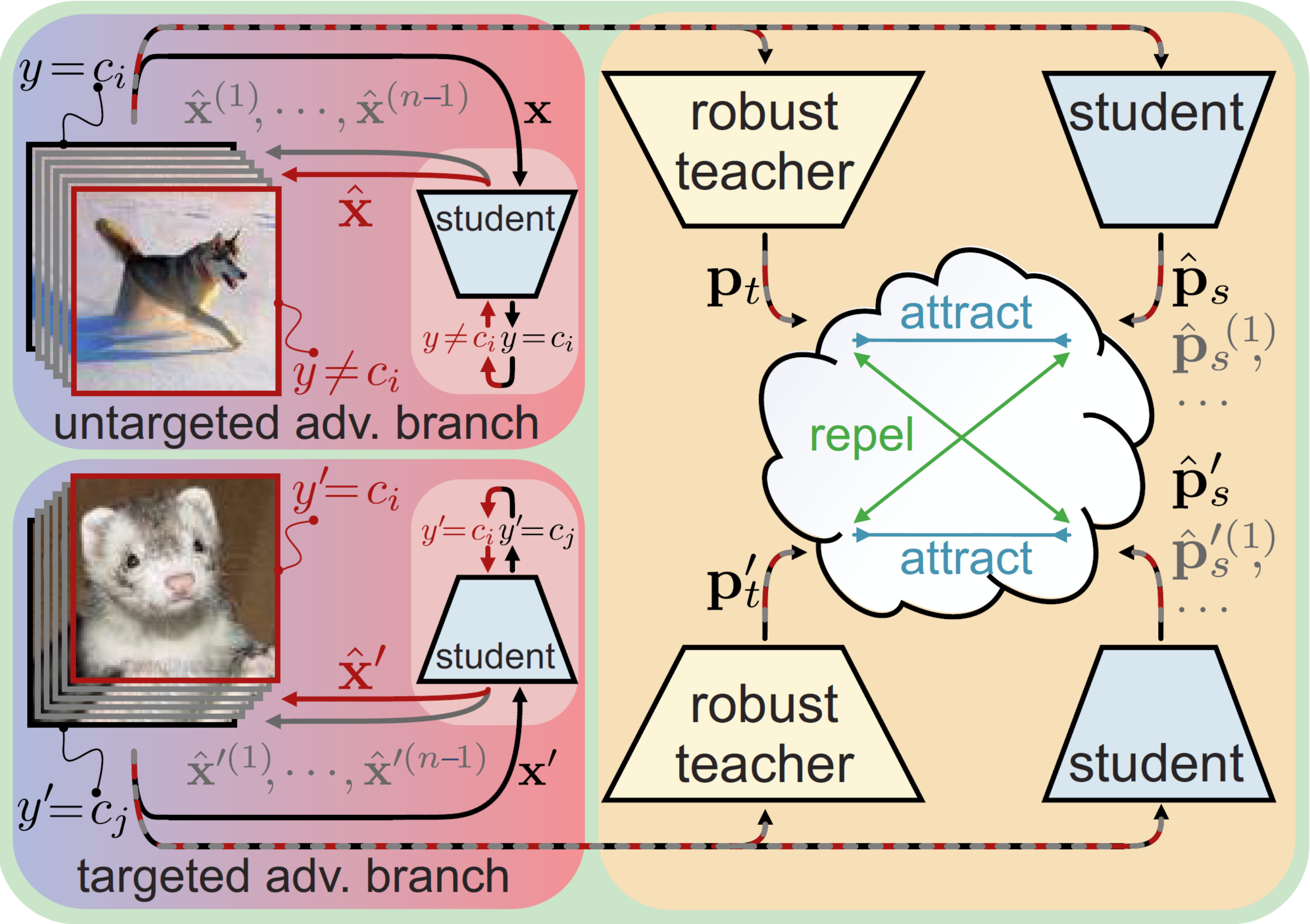
$$\mathcal{L}_{\text{IAKD}} = \sum_{i=1}^{n-1} w\left(\hat{\mathbf{x}}^{(i)} | \mathbf{x}\right) \cdot \mathcal{L}_{\text{KL}}\left(f_{\boldsymbol{\theta}_t}\left(\hat{\mathbf{x}}^{(i)}\right) \| f_{\boldsymbol{\theta}_s}\left(\hat{\mathbf{x}}^{(i)}\right)\right)$$

Re-weighting

$$w\left(\hat{\mathbf{x}}^{(i)} | \mathbf{x}\right) = \frac{(1-\gamma)\, i}{n} + \frac{\gamma \left|\left(f_{\boldsymbol{\theta}_t}(\mathbf{x})\right)_y - \left(f_{\boldsymbol{\theta}_s}\left(\hat{\mathbf{x}}^{(i)}\right)\right)_y\right|}{\max_{j \in \mathcal{B}} \left|\left(f_{\boldsymbol{\theta}_t}(\mathbf{x}_j)\right)_{y_j} - \left(f_{\boldsymbol{\theta}_s}\left(\hat{\mathbf{x}}_j^{(i)}\right)\right)_{y_j}\right|}$$

We theoretically prove that **such a weighting mechanism captures the localized κ-Lipschitz smoothness** of the student model.

■ Dual-branch Adversarially Robust knowledge dIstillatioN (DARWIN)



Attraction

$$\text{attract}\begin{cases}\{f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)})\}_{i=1}^{n} \to f_{\boldsymbol{\theta}_t}(\mathbf{x}) \\ \{f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}'^{(i)})\}_{i=1}^{n} \to f_{\boldsymbol{\theta}_t}(\mathbf{x}')\end{cases}$$

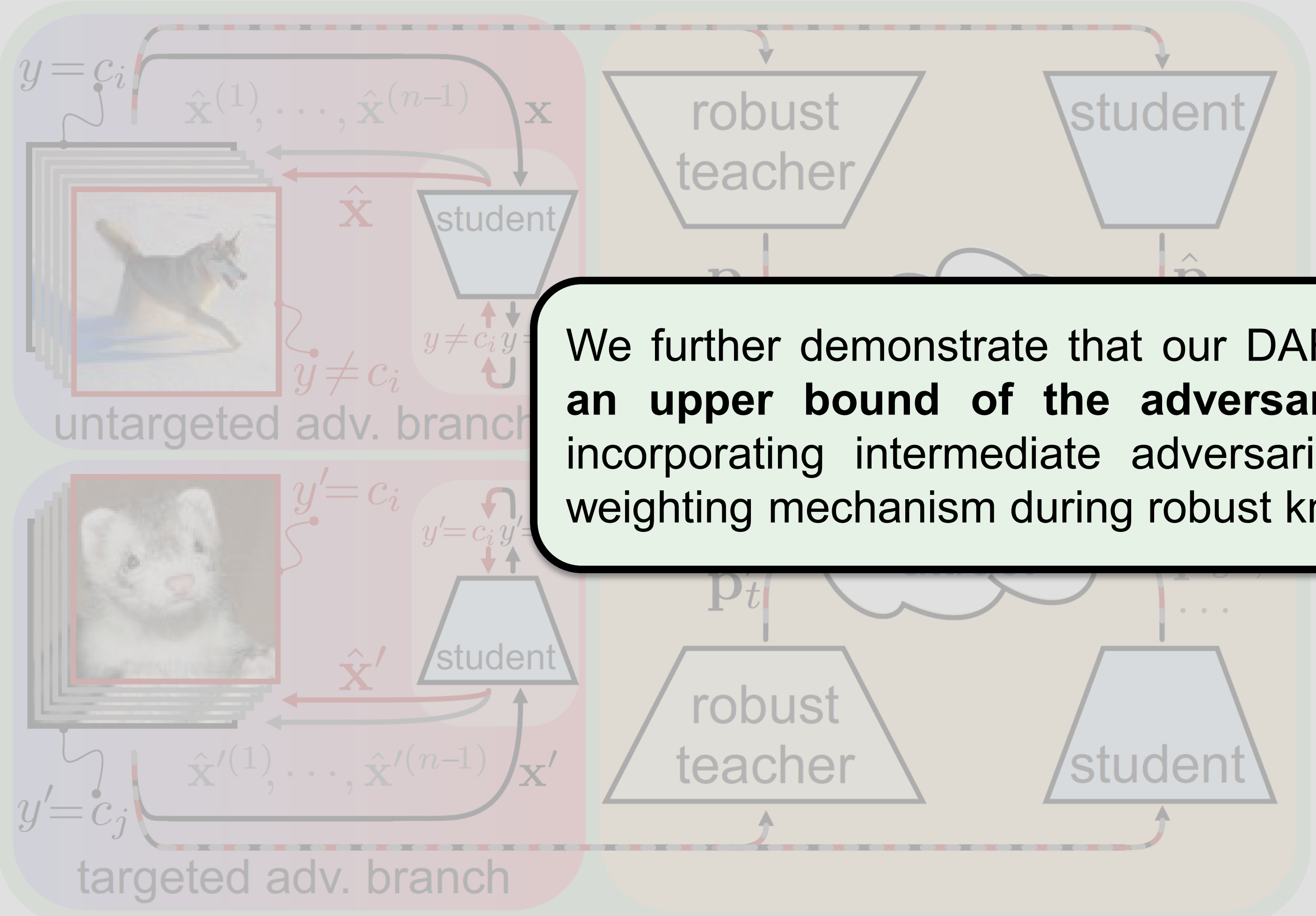
Repulsion

$$\text{repel}\begin{cases}\{f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)})\}_{i=1}^{n} \leftrightarrow f_{\boldsymbol{\theta}_t}(\mathbf{x}') \\ \{f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}'^{(i)})\}_{i=1}^{n} \leftrightarrow f_{\boldsymbol{\theta}_t}(\mathbf{x})\end{cases}$$

Dual-Branch Knowledge Distillation

$$\mathcal{L}_{\text{DBKD}} = \sum_{i=1}^{n}\Big[w(\hat{\mathbf{x}}^{(i)}|\mathbf{x})\,\mathcal{L}_{\text{tri}}\big(f_{\boldsymbol{\theta}_t}(\mathbf{x}), f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)}), f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}'^{(i)})\big) + w(\hat{\mathbf{x}}'^{(i)}|\mathbf{x}')\,\mathcal{L}_{\text{tri}}\big(f_{\boldsymbol{\theta}_t}(\mathbf{x}'), f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}'^{(i)}), f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)})\big)\Big]$$

Dual-branch Adversarially Robust knowledge dIstillatioN (DARWIN)



Attraction

$$\text{attract}\begin{cases}\{f_{\theta_s}(\hat{\mathbf{x}}^{(i)})\}_{i=1}^{n} \to f_{\theta_t}(\mathbf{x}) \\ \{f_{\theta_s}(\hat{\mathbf{x}}'^{(i)})\}_{i=1}^{n} \to f_{\theta_t}(\mathbf{x}')\end{cases}$$

$${}_{i=1}^{n} \leftrightarrow f_{\theta_t}(\mathbf{x}')$$

$${}_{i=1}^{n} \leftrightarrow f_{\theta_t}(\mathbf{x})$$

We further demonstrate that our DARWIN method **minimizes an upper bound of the adversarially robust risk** when incorporating intermediate adversarial samples with the re-weighting mechanism during robust knowledge distillation.

Dual-Branch Knowledge Distillation

$$\mathcal{L}_{\text{DBKD}} = \sum_{i=1}^{n} \Big[w(\hat{\mathbf{x}}^{(i)}|\mathbf{x})\, \mathcal{L}_{\text{tri}}\big(f_{\theta_t}(\mathbf{x}), f_{\theta_s}(\hat{\mathbf{x}}^{(i)}), f_{\theta_s}(\hat{\mathbf{x}}'^{(i)})\big) + w(\hat{\mathbf{x}}'^{(i)}|\mathbf{x}')\, \mathcal{L}_{\text{tri}}\big(f_{\theta_t}(\mathbf{x}'), f_{\theta_s}(\hat{\mathbf{x}}'^{(i)}), f_{\theta_s}(\hat{\mathbf{x}}^{(i)})\big) \Big]$$

■ Standard Comparison (Distillation from a Large Model):

Type	Architecture	Method	y	CIFAR-10				CIFAR-100			
				Natural	PGD-20	CW	AA	Natural	PGD-20	CW	AA
Teacher	WRN-34	TRADES [58]	✓	84.92	55.34	54.21	52.55	60.04	31.56	28.64	27.38
Student	ResNet-18	ARD [15]	✓	82.95	52.26	51.69	49.46	57.46	30.14	27.11	25.30
		IAD [63]	✓	82.41	53.06	51.79	49.78	56.38	30.61	27.35	25.51
		RSLAD [64]	✗	83.12	53.91	52.84	51.19	57.23	31.08	28.29	26.62
		CRDND [54]	✗	83.92	52.70	50.95	49.05	58.03	30.16	27.02	25.68
		GACD [1]	✗	82.76	53.42	52.26	50.07	56.82	31.19	27.81	26.12
		DARWIN	✓	**84.48**	**55.07**	53.85	52.24	**59.12**	**32.30**	**28.95**	**27.26**
		DARWIN-LF	✗	84.35	55.02	**53.99**	**52.33**	59.04	32.18	28.62	27.13
	MNV2	ARD [15]	✓	82.44	51.91	50.64	48.40	55.28	30.23	27.05	25.28
		IAD [63]	✓	81.61	52.30	50.19	48.34	54.26	30.46	27.13	25.50
		RSLAD [64]	✗	82.89	52.72	52.04	50.04	57.31	30.48	27.86	25.89
		CRDND [54]	✗	82.77	52.57	50.11	49.28	56.24	29.65	26.68	25.61
		GACD [1]	✗	82.90	52.49	51.40	49.55	56.10	30.49	27.18	25.33
		DARWIN	✓	84.06	**53.94**	**53.11**	**51.28**	**58.45**	**31.53**	**28.36**	26.55
		DARWIN-LF	✗	**84.08**	53.76	52.80	51.09	58.41	31.44	28.33	**26.58**

Robust Distillation from ViTs:

Type	Architecture	Method	Natural	PGD-20	AA
Teacher	ViT-B	AT-PRM [33]	83.98	53.10	49.66
Student	ResNet-18	ARD [15]	82.76	52.95	49.03
		RSLAD [64]	82.33	54.89	49.74
		IAD [63]	82.27	53.42	49.48
		CRDND [54]	82.19	53.16	48.98
		GACD [1]	81.64	54.24	49.95
		DARWIN	**83.75**	54.80	51.42
		DARWIN-LF	83.73	**54.95**	**51.49**
Teacher	DeiT-S	AT-PRM [33]	82.68	52.47	49.27
Student	MNV2	ARD [15]	81.59	53.45	49.20
		RSLAD [64]	80.86	53.91	50.18
		IAD [63]	80.41	54.12	49.62
		CRDND [54]	80.27	52.21	48.46
		GACD [1]	79.97	54.00	48.91
		DARWIN	83.02	54.46	**51.19**
		DARWIN-LF	**83.15**	**54.62**	51.13

Self-Distillation w/ Generated Data:

Dataset	Type	DDPM	Method	Natural	Robust
CIFAR-10	**Teacher**	✗	TRADES [58]	82.45	48.90
	Student	✓	ARD [15]	82.89	53.41
		✓	RSLAD [64]	82.05	52.60
		✓	IAD [63]	82.95	53.47
		✓	**DARWIN**	84.13	55.92
		✓	**DARWIN-LF**	**84.68**	**56.41**
CIFAR-100	**Teacher**	✗	TRADES [58]	56.37	23.78
	Student	✓	ARD [15]	56.07	26.92
		✓	RSLAD [64]	53.40	26.00
		✓	IAD [63]	55.82	26.77
		✓	**DARWIN**	58.18	28.24
		✓	**DARWIN-LF**	**58.74**	**28.45**

Ablations:

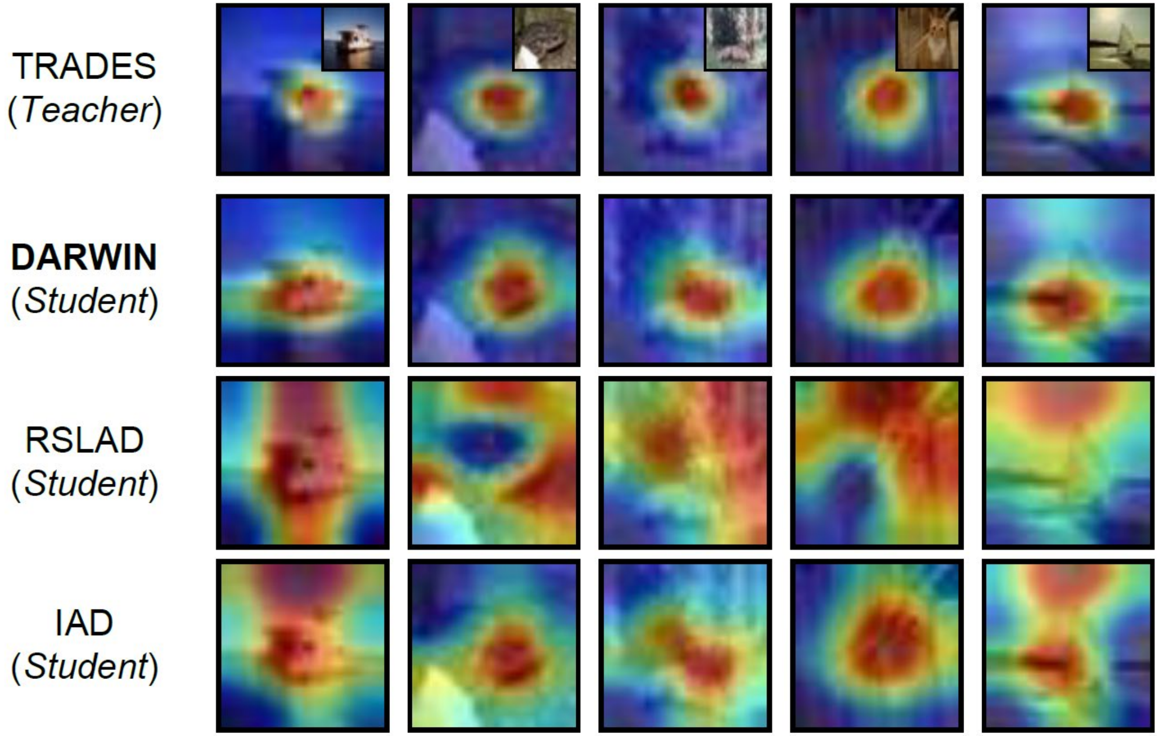
Impact of Each Module

	ARKD	IAKD	DBKD	Natural	PGD-20	AA
1	✓			83.09/57.34	53.95/30.59	50.66/25.32
2	✓	✓		82.74/56.68	54.52/32.11	51.70/26.95
3	✓		✓	84.68/59.23	54.36/31.53	51.21/26.18
	✓	✓	✓	84.48/59.12	55.07/32.30	52.24/27.26

Diverse Weighting Strategies

Weighting Strategies	CIFAR-10			CIFAR-100		
	Natural	PGD	AA	Natural	PGD	AA
Uniform Weighting (no weights)	83.32	53.84	50.92	57.81	30.95	25.70
First Term switched on ($\gamma=0$)	84.13	54.18	51.58	58.53	31.44	26.20
Second Term switched on ($\gamma=1$)	83.85	54.45	51.82	58.16	31.82	26.47
Both Terms switched on ($\gamma=0.5$)	**84.48**	**55.07**	**52.24**	**59.12**	**32.30**	**27.26**

Attention Visualizations:



■ Contributions:

- We propose a novel robust knowledge distillation method that **integrates intermediate adversaries along the adversarial path**. An **adaptive weighting mechanism** is proposed to calibrate the influence of each intermediate sample to facilitate the distillation of adversarial paths. Our strategy also leads to **minimizing an upper bound of the adversarially robust risk**.
- To capture relations between decision boundaries, we devise a **dual-branch mechanism** by harnessing the **complementary characteristics of untargeted and targeted adversarial samples**. This inter-class relational learning facilitates a more effective robustness transfer.
- Extensive experiments showcase the superiority of our method compared with the state-of-the-art approaches across various settings, including diverse backbones, auxiliary data, and cross-dataset distillation.

Thank you!

Robust Distillation via Untargeted and Targeted Intermediate Adversarial Samples

Junhao Dong, Piotr Koniusz, Junxi Chen, Z. Jane Wang, Yew-Soon Ong

Reporter：Junhao Dong