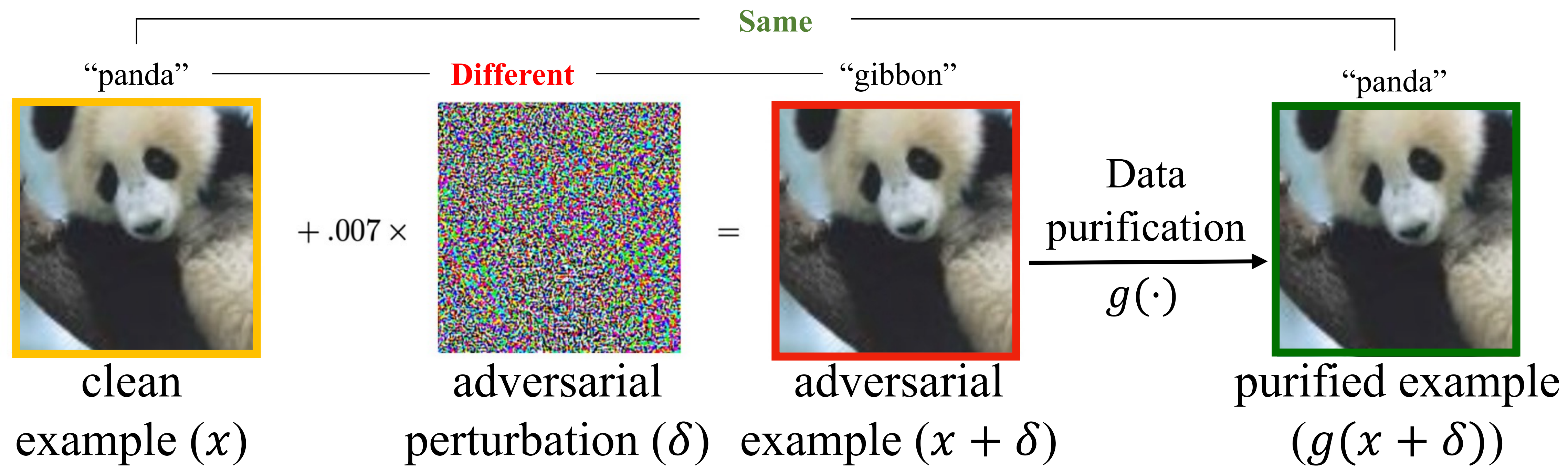
Adversarial Training on Purification (AToP): Advancing Both Robustness and Generalization

Guang Lin[1,2], Chao Li[2], Jianhai Zhang[3], Toshihisa Tanaka[1,2] and Qibin Zhao[2,1,*]

[1]Tokyo University of Agriculture and Technology, [2]RIKEN AIP, [3]Hangzhou Dianzi University

Paper

Background


Same
"panda"
Different
"gibbon"
"panda"
+ .007 ×
=
Data purification
$g(\cdot)$
clean example (x)
adversarial perturbation (δ)
adversarial example $(x+\delta)$
purified example $(g(x+\delta))$

The figure is modified based on: Explaining and Harnessing Adversarial Examples. ICLR 2015.

Adversarial Attack: $f(x') = y' \neq f(x) = y$,

where $x' = x + \delta, \delta = \arg\max_{\delta \leq \varepsilon} \mathcal{L}(f(x+\delta), y)$

Adversarial Training (AT): $f'(x') = y$,

where robust model f' is trained with adversarial examples x' and true label y.

Adversarial Purification (AP): $f(g(x')) = y$,

*where purifier model g is a pre-trained generator.

Related works

Table 1: Robustness comparison of defenses with expectation (negative impacts are marked in red).

Defense method	Clean images	Known attacks	Unseen attacks
Vanilla model	~94%	~0%	~0%
Expectation	=	↑↑	↑
AT	↓↓	↑↑	N/A
AP	↓	↑	↑
AToP (ours)	≈	↑↑	↑

Adversarial Training (AT)

- [✓] Achieve optimal robustness on known attacks.
- [✗] Vulnerable to unseen attacks.
- [✗] Reduce the accuracy of clean examples.

Adversarial Purification (AP)

- [✓] Keep generalization against unseen attacks.
- [✗] Weaker robustness than AT on known attacks.
- [✗] Slightly reduce the accuracy of clean examples.

The pre-trained purifier model is not good enough for classification and non-robust itself.

Method

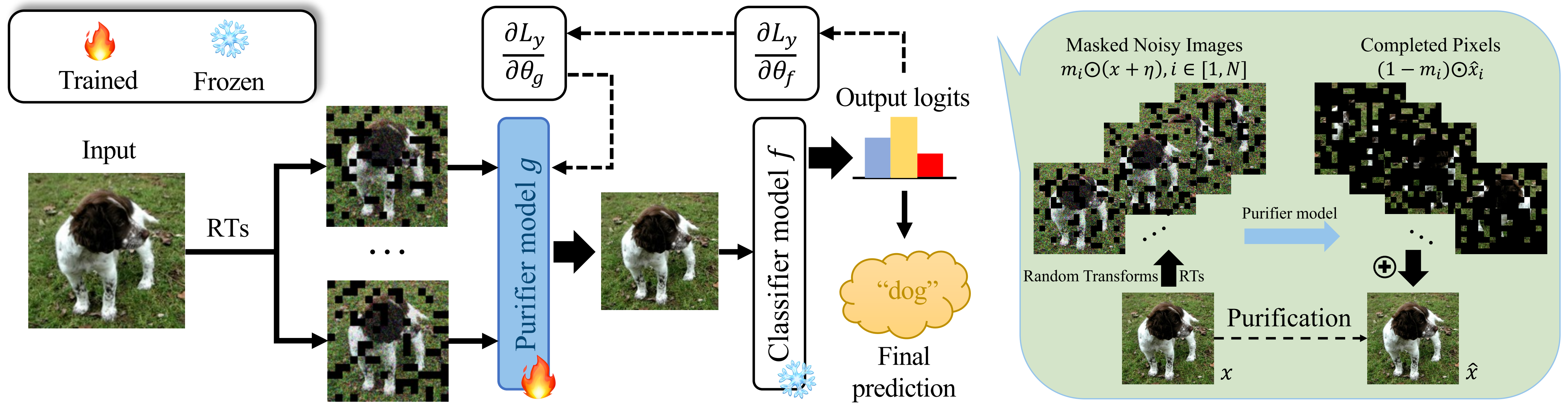
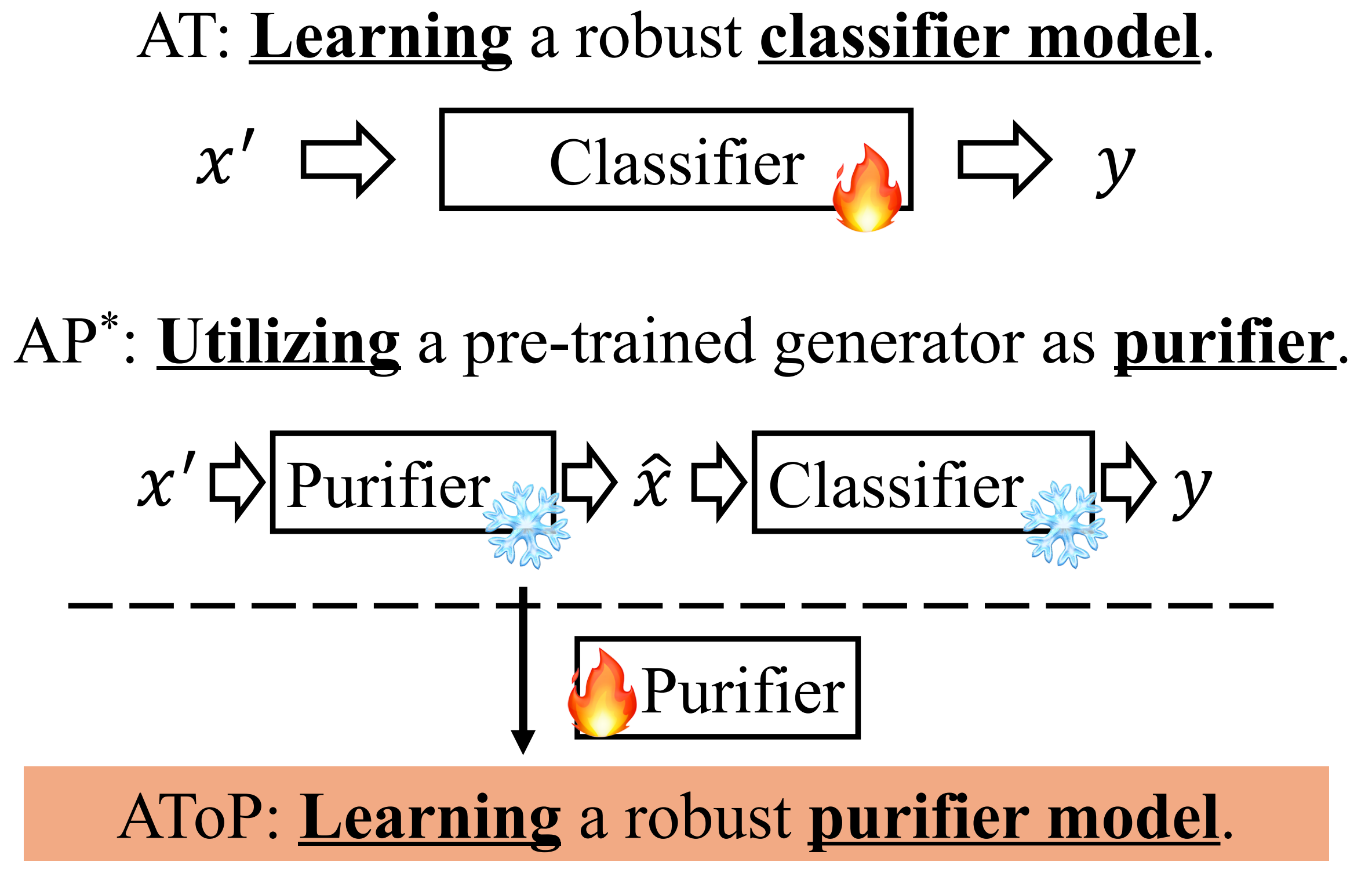


Figure 1: Illustration of adversarial training on purification (AToP). To fine-tune the purifier with adversarial training, we aim to optimize the model by freezing classifier parameters and only updating purifier parameters.



AT: **Learning** a robust **classifier model**.

$x' \Rightarrow$ Classifier $\Rightarrow y$

AP*: **Utilizing** a pre-trained generator as **purifier**.

$x' \Rightarrow$ Purifier $\Rightarrow \hat{x} \Rightarrow$ Classifier $\Rightarrow y$

Purifier

AToP: **Learning** a robust **purifier model**.

Based on the **pre-trained generator model** trained by the original generative loss ℓ_g :

$L_{\theta_g} = \ell_g(x, \theta_g)$.

We incorporate a classification loss ℓ_{cls} to **fine-tune the generator model** with

a) clean examples x and labels y :

$L_{\theta_g} = \ell_g(x, \theta_g) + \lambda \cdot \ell_{cls}(x, y, \theta_g, \theta_f)$.

b) adversarial examples x' and labels y :

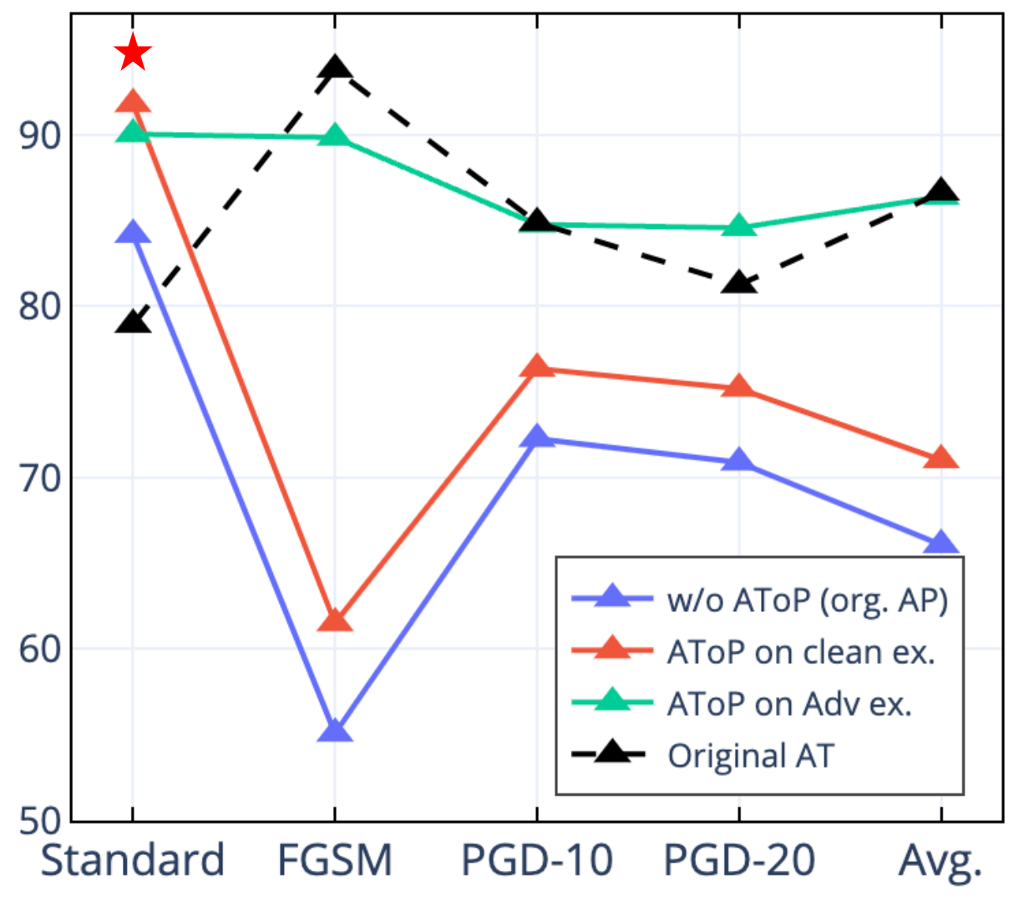
$L_{\theta_g} = \ell_g(x', \theta_g) + \lambda \cdot \ell_{cls}(x', y, \theta_g, \theta_f)$.

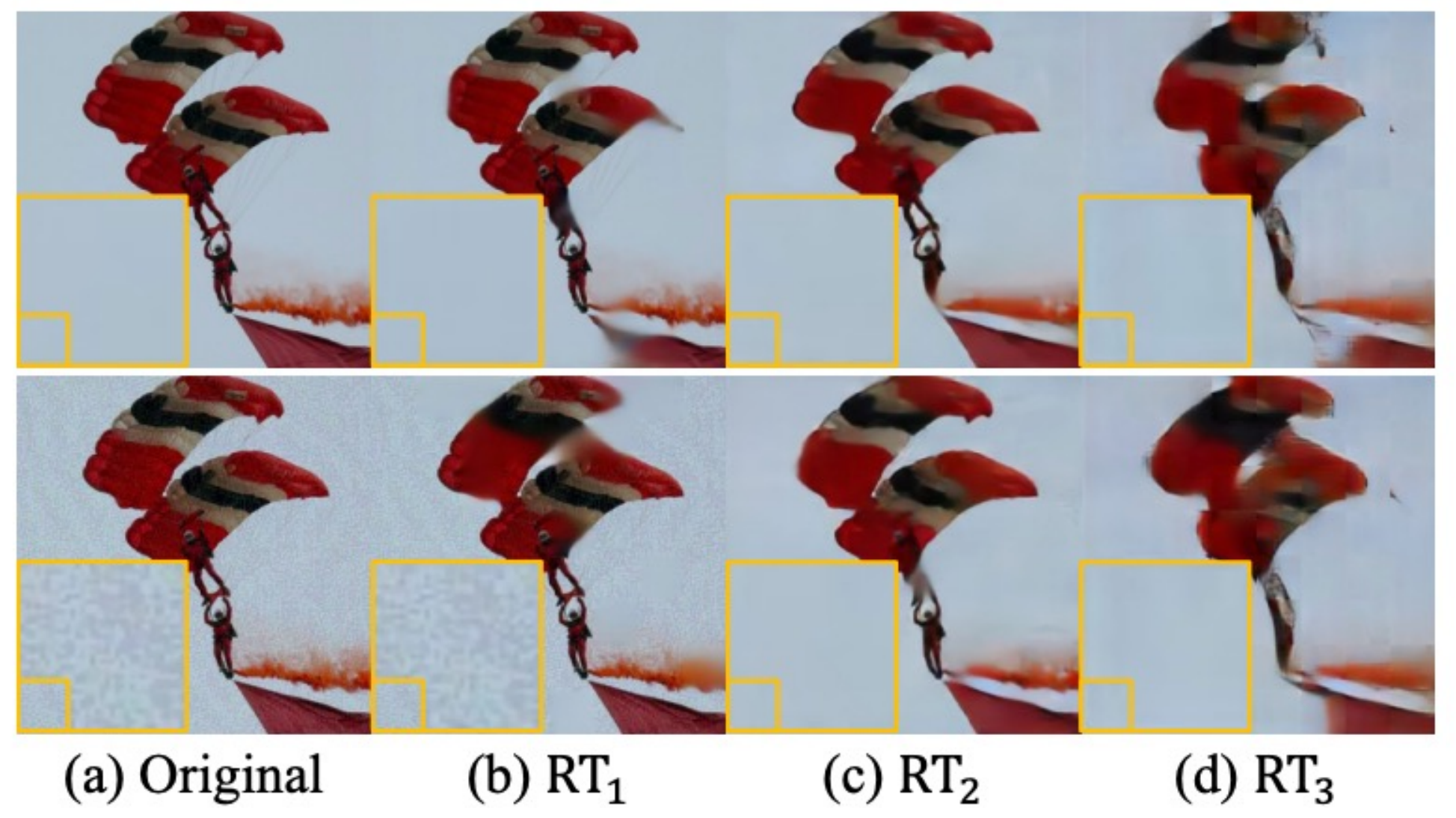
Experimental results

Table 6: Standard accuracy and robust accuracy against AutoAttack l_∞ ($\epsilon = 8/255$), l_2 ($\epsilon = 1$) and StAdv non-l_p ($\epsilon = 0.05$) threat models on CIFAR-10 with ResNet-50 classifier model.

Defense method	Standard Acc.	l_∞	l_2	StAdv
Standard Training	94.8	0.0	0.0	0.0
Adv. Training with l_∞ (Laidlaw et al., 2021)	86.8	<u>49.0</u>	19.2	4.8
Adv. Training with l_2 (Laidlaw et al., 2021)	85.0	39.5	<u>47.8</u>	7.8
Adv. Training with StAdv (Laidlaw et al., 2021)	86.2	0.1	0.2	<u>53.9</u>
Adv. Training with all (Laidlaw et al., 2021)	84.0	<u>25.7</u>	<u>30.5</u>	<u>40.0</u>
PAT-self (Laidlaw et al., 2021)	82.4	30.2	34.9	46.4
Adv. CRAIG (Dolatabadi et al., 2022)	83.2	40.0	33.9	49.6
DiffPure (Nie et al., 2022)	88.2	70.0	70.9	55.0
Ours	**89.1**	**71.2**	**73.4**	**56.4**

Figure 4a: Comparison of AT, AP and AToP. Figure 3: Clean (Top) and adversarial examples (Bottom).




(a) Original (b) RT$_1$ (c) RT$_2$ (d) RT$_3$

Adversarial Training on Purification (AToP)

- [✓] Achieve optimal robustness on known attacks.
- [✓] Keep generalization against unseen attacks.
- [✓] Achieve optimal accuracy on clean examples.

Table 7: Standard accuracy and robust accuracy of attacking the classifier model on CIFAR-10 with ResNet-18.All attacks are l_∞ threat model with $\epsilon = 8/255$.

Transforms	AToP	Standard Acc.	FGSM	PGD-10	PGD-20	PGD-1000
RT$_1$	×	**93.36**	16.60	0.00	0.00	0.00
	✓	**93.36**	**91.99**	**43.55**	**36.72**	**39.45**
RT$_2$	×	84.18	55.08	72.27	70.90	67.97
	✓	**90.04**	**89.84**	**84.77**	**84.57**	**84.38**
RT$_3$	×	75.98	67.97	70.51	70.70	70.31
	✓	**80.02**	**70.90**	**73.05**	**72.07**	**73.44**

More complex RT can better remove perturbations, but also result in a loss of semantic information. Accuracy ↑ Accuracy ↓

Conclusion: We develop a novel efficient defense technology by combining AT and AP, which can **learn a robust purifier**.
Limitations: AToP requires training on the purifier, and as the complexity of purifier increases, so does the training cost.