PeerAiD:

Improving Adversarial Distillation from a Specialized Peer Tutor

Jaewon Jung, Hongsun Jang, Jaeyong Song, and Jinho Lee

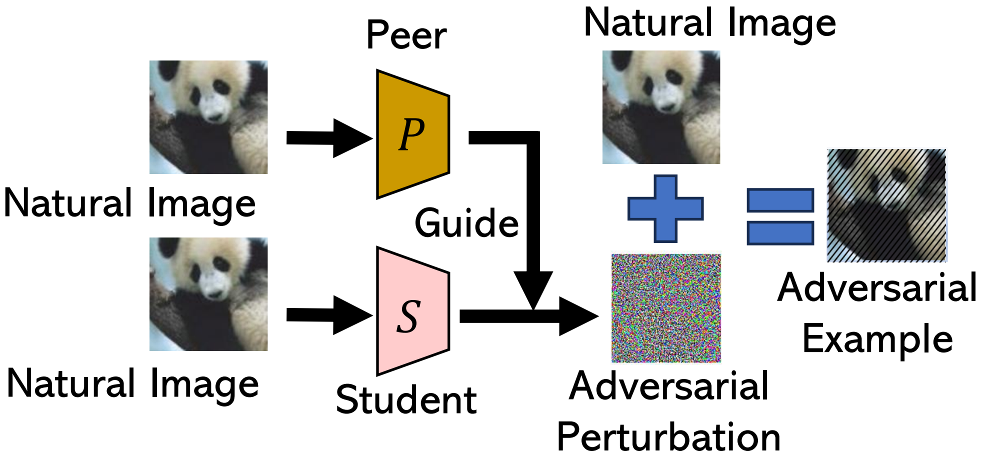
Department of Electrical and Computer Engineering, Seoul National University

IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024 ***(CVPR)***

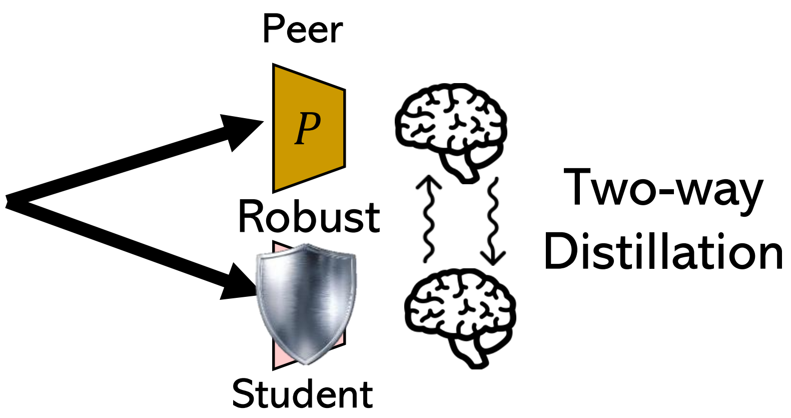
Overview

- **PeerAiD proposes using the peer, which interactively learns with the student during adversarial distillation.**

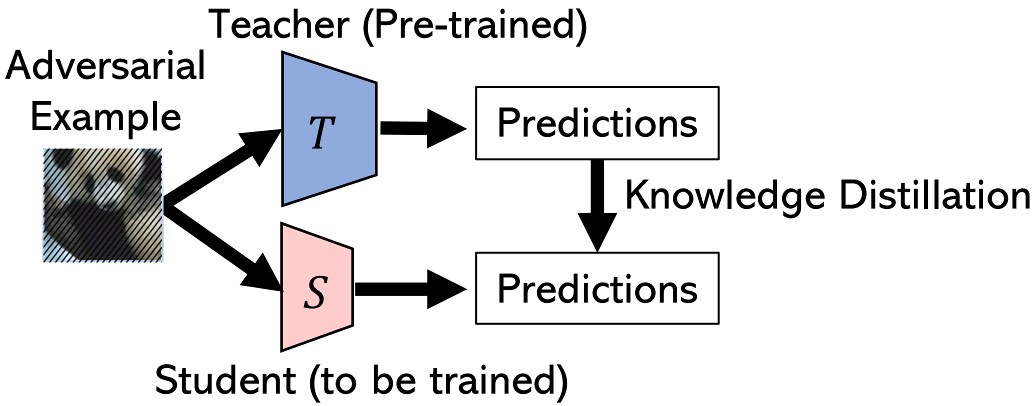
① Adversarial example generation



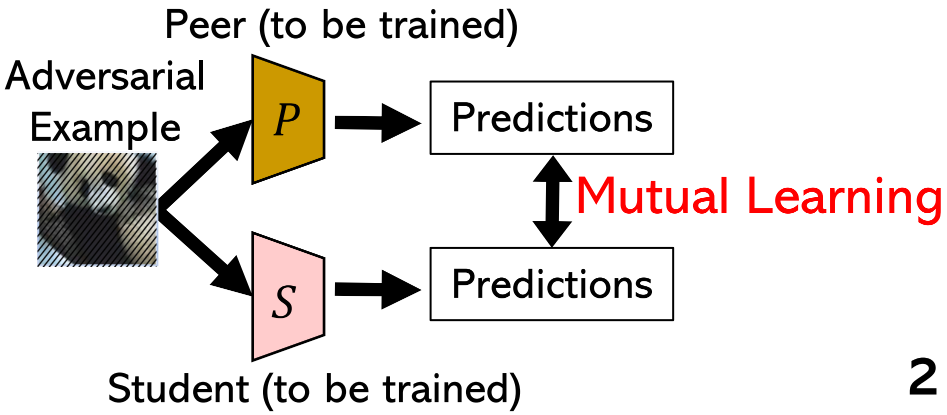
② Weight optimization



- Conventional AD



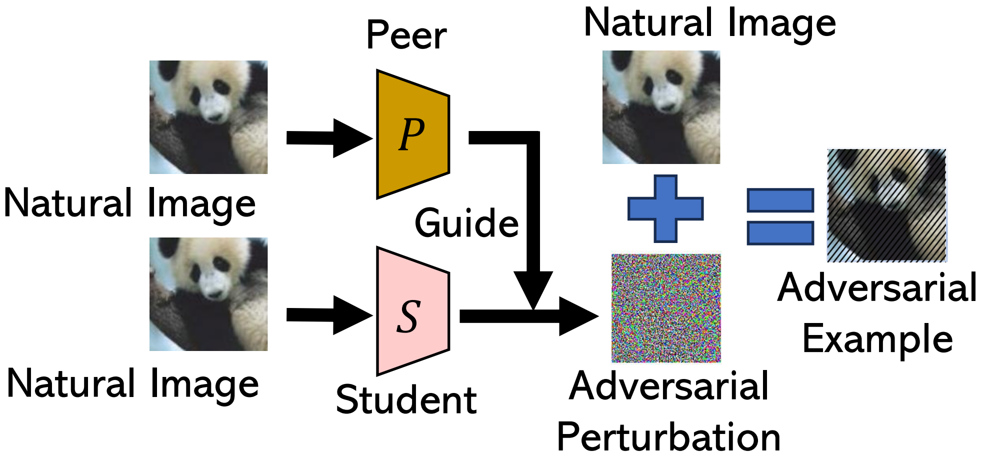
- PeerAiD



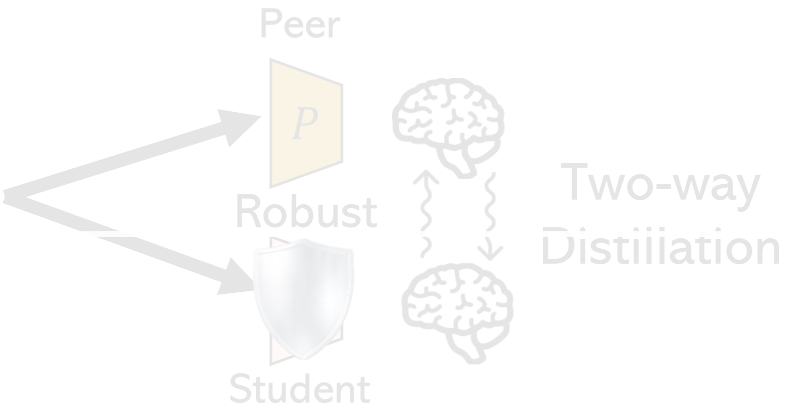
Overview

- **PeerAiD proposes using the peer, which interactively learns with the student during adversarial distillation.**

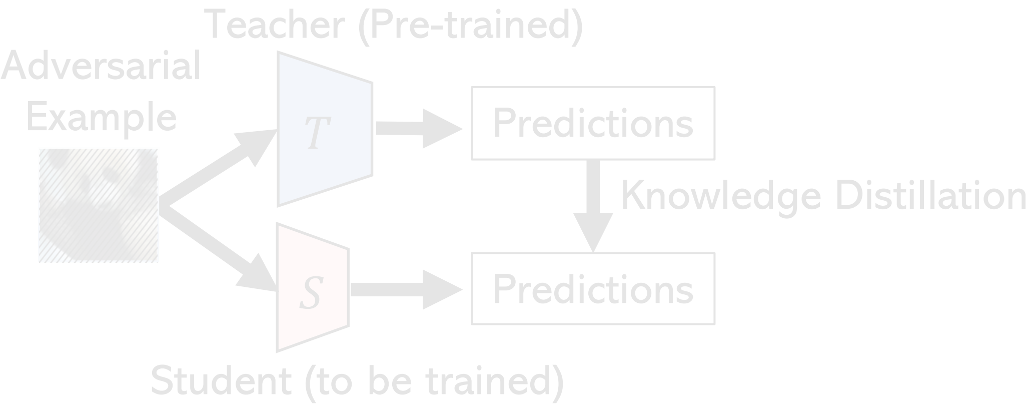
① Adversarial example generation



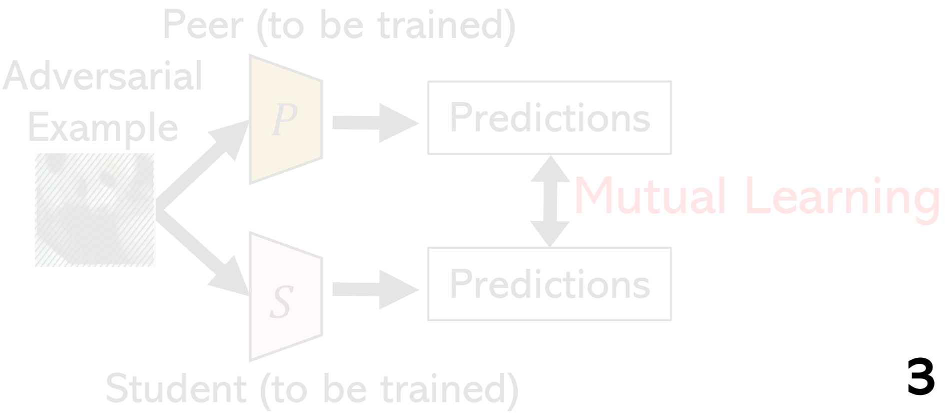
② Weight optimization



- Conventional AD

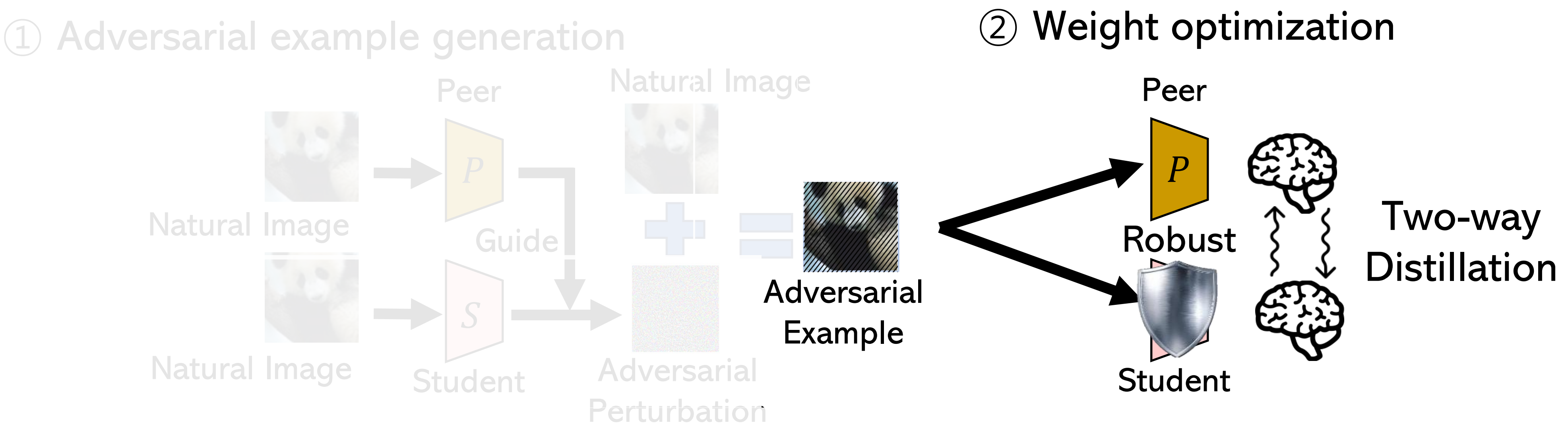


- PeerAiD

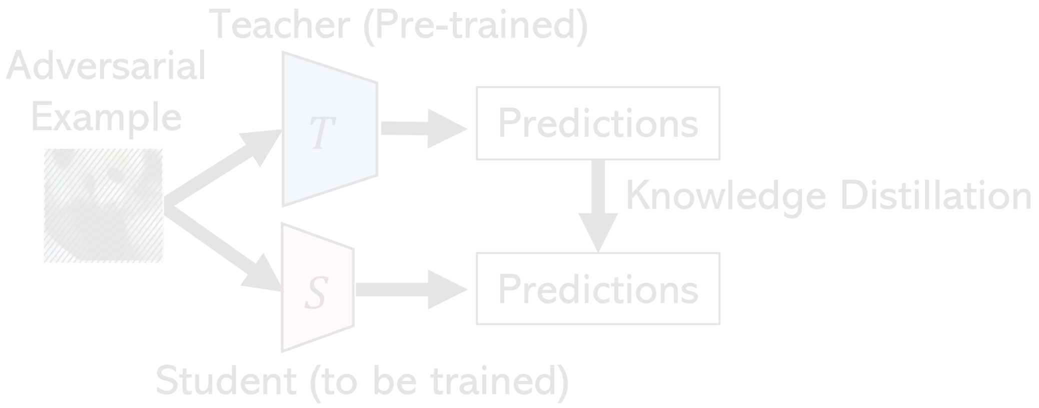


Overview

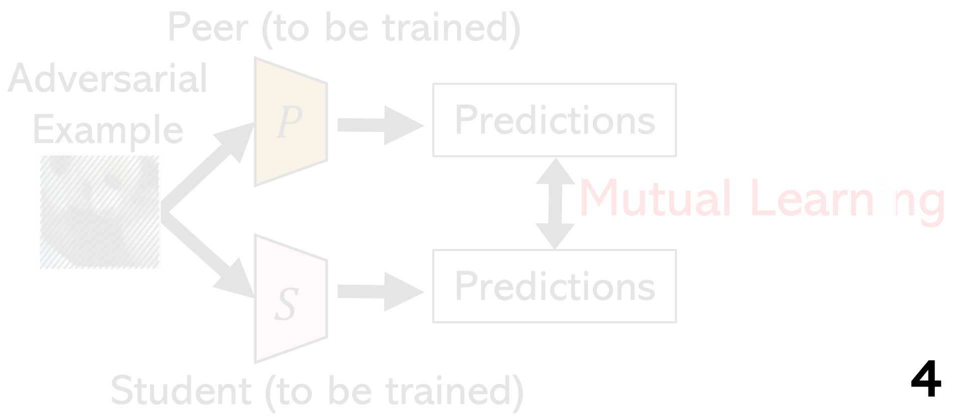
- **PeerAiD proposes using the peer, which interactively learns with the student during adversarial distillation.**



- Conventional AD

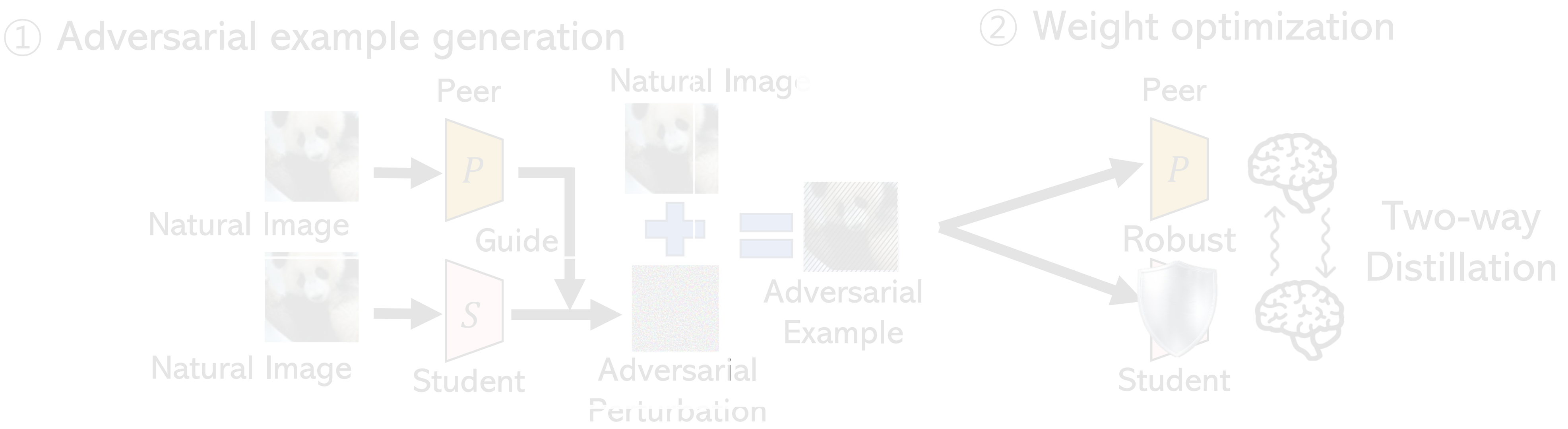


- PeerAiD

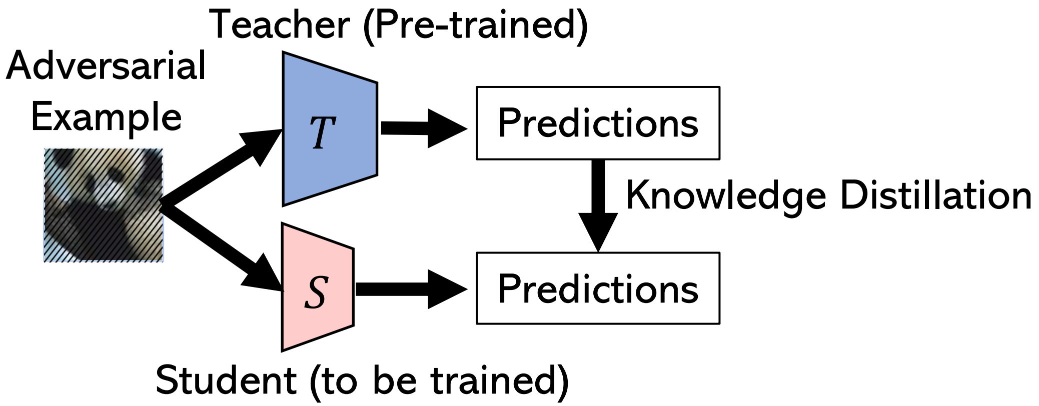


Overview

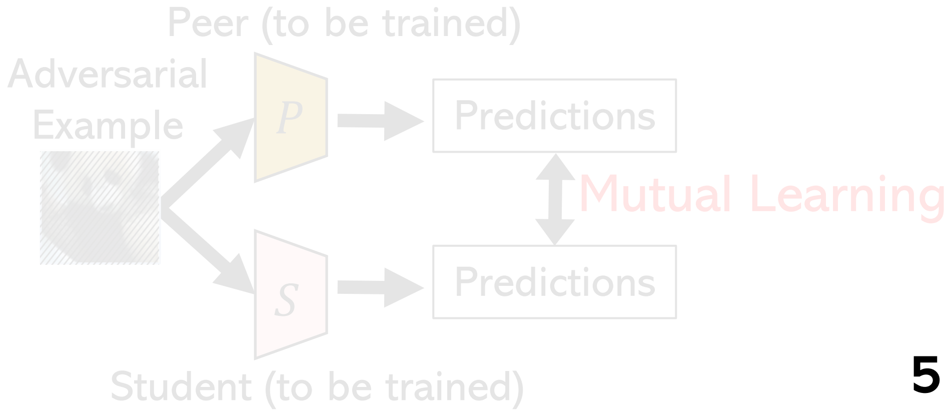
- **PeerAiD proposes using the peer, which interactively learns with the student during adversarial distillation.**



- Conventional AD

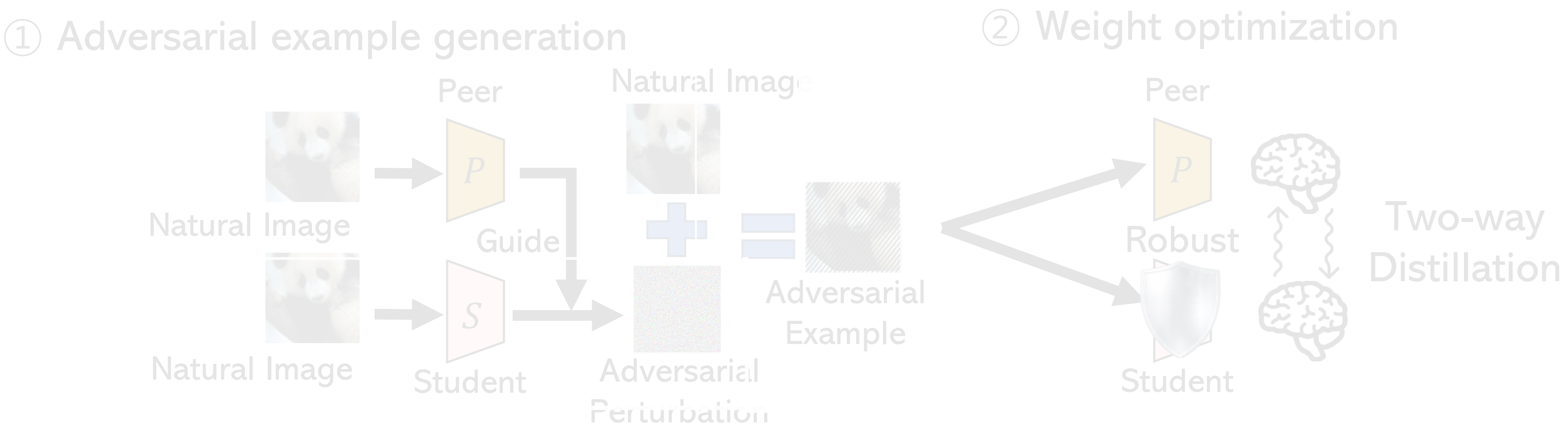


- PeerAiD

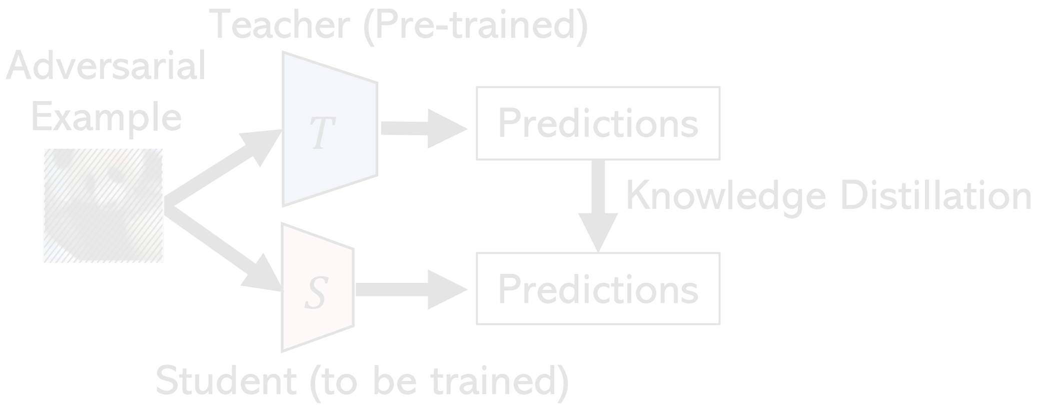


Overview

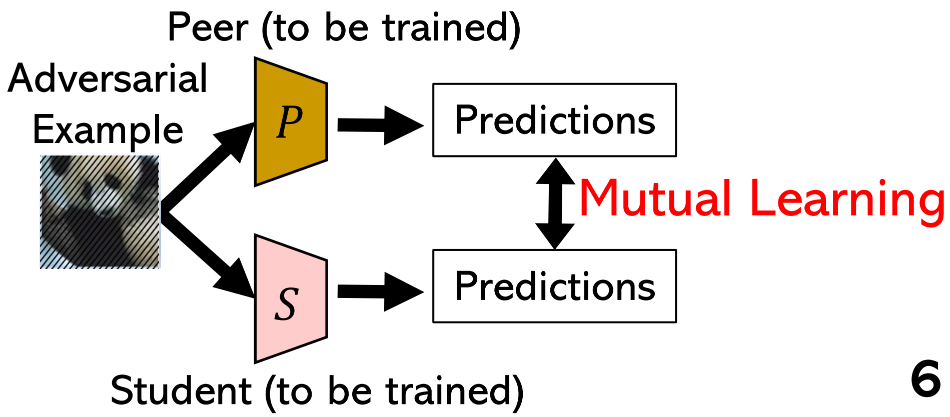
- **PeerAiD proposes using the peer, which interactively learns with the student during adversarial distillation.**



- Conventional AD



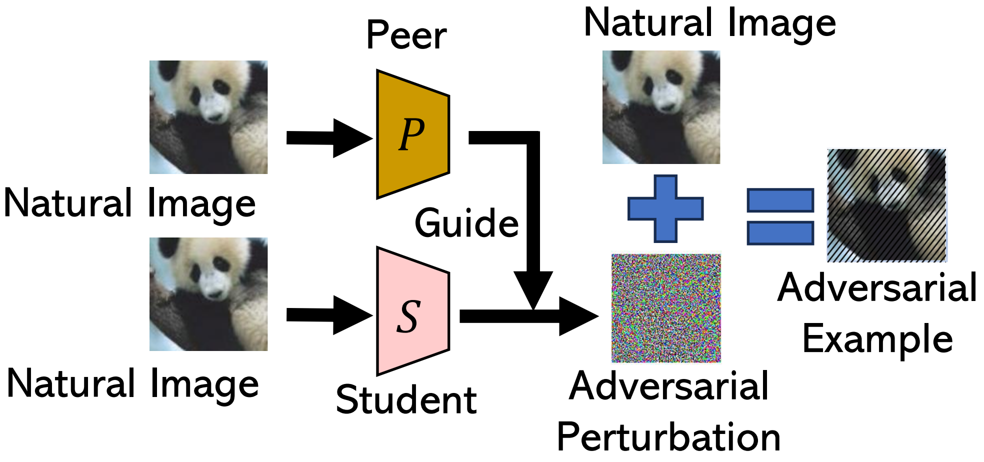
- PeerAiD



Overview

- **PeerAiD proposes using the peer, which interactively learns with the student during adversarial distillation.**

① Adversarial example generation



Peer

Natural Image

Natural Image

Guide

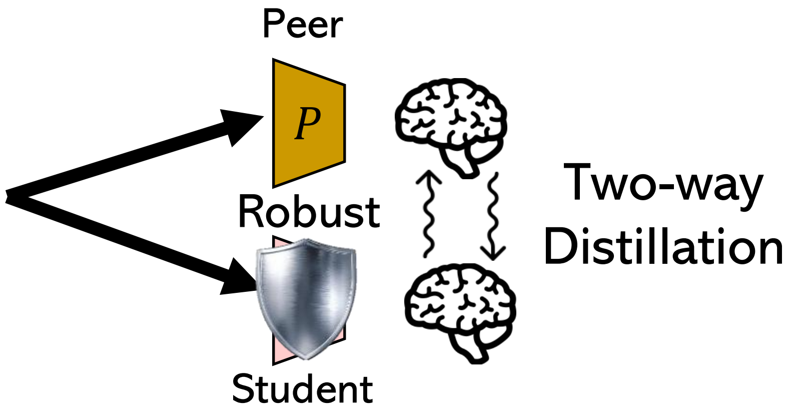
Natural Image

Student

Adversarial Perturbation

Adversarial Example

② Weight optimization



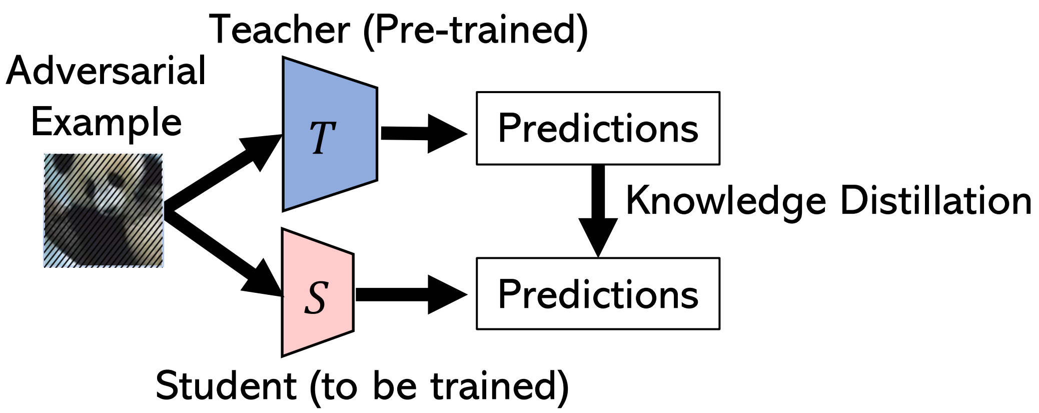
Peer

Robust

Student

Two-way Distillation

- Conventional AD



Teacher (Pre-trained)

Adversarial Example

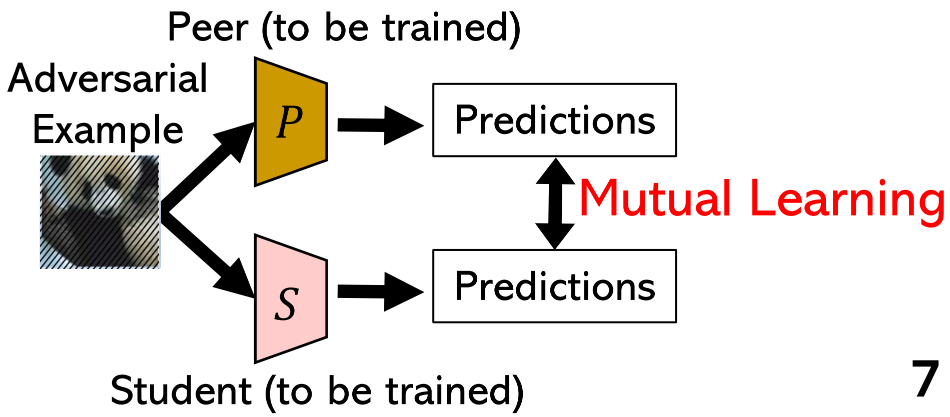
Predictions

Knowledge Distillation

Predictions

Student (to be trained)

- PeerAiD



Peer (to be trained)

Adversarial Example

Predictions

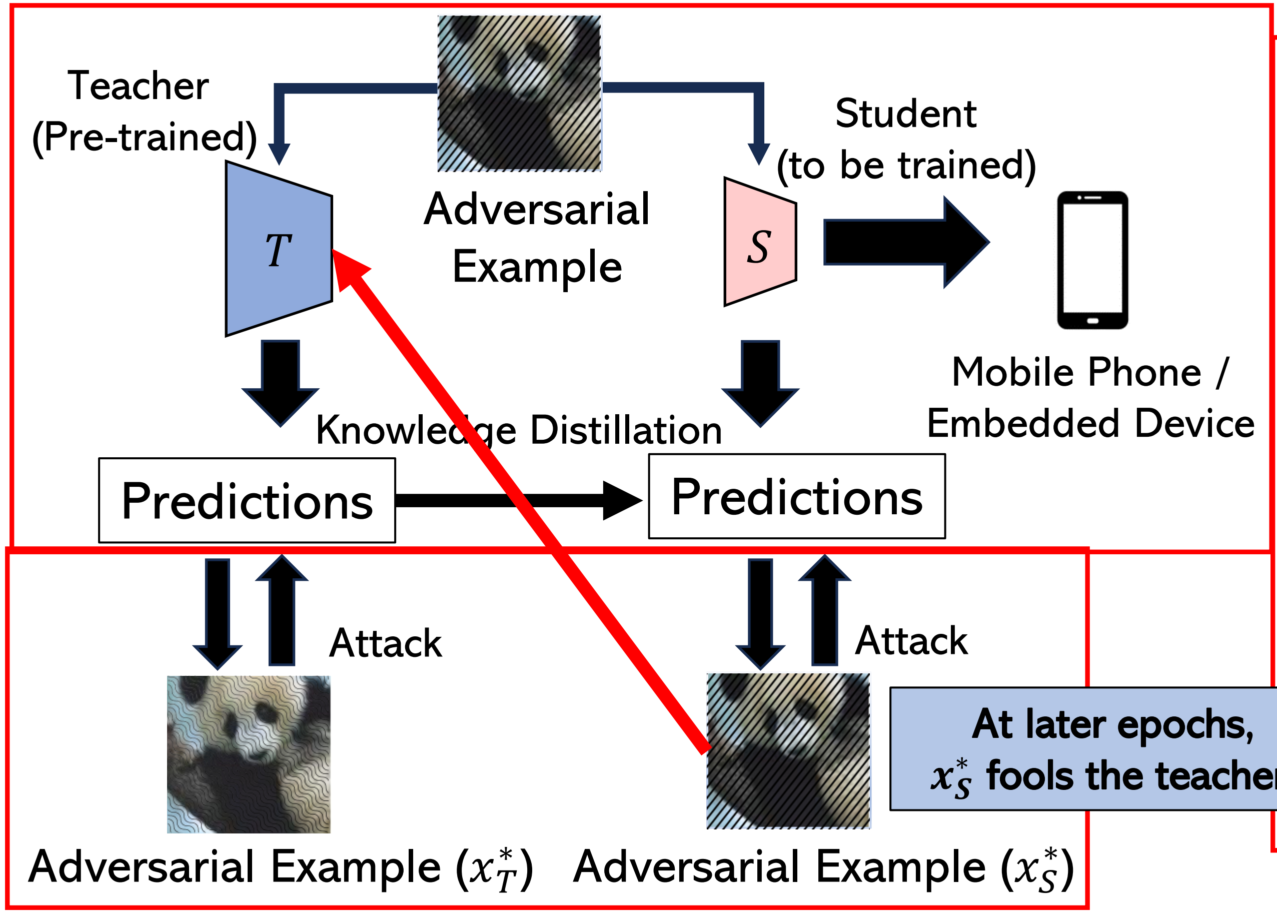
Mutual Learning

Predictions

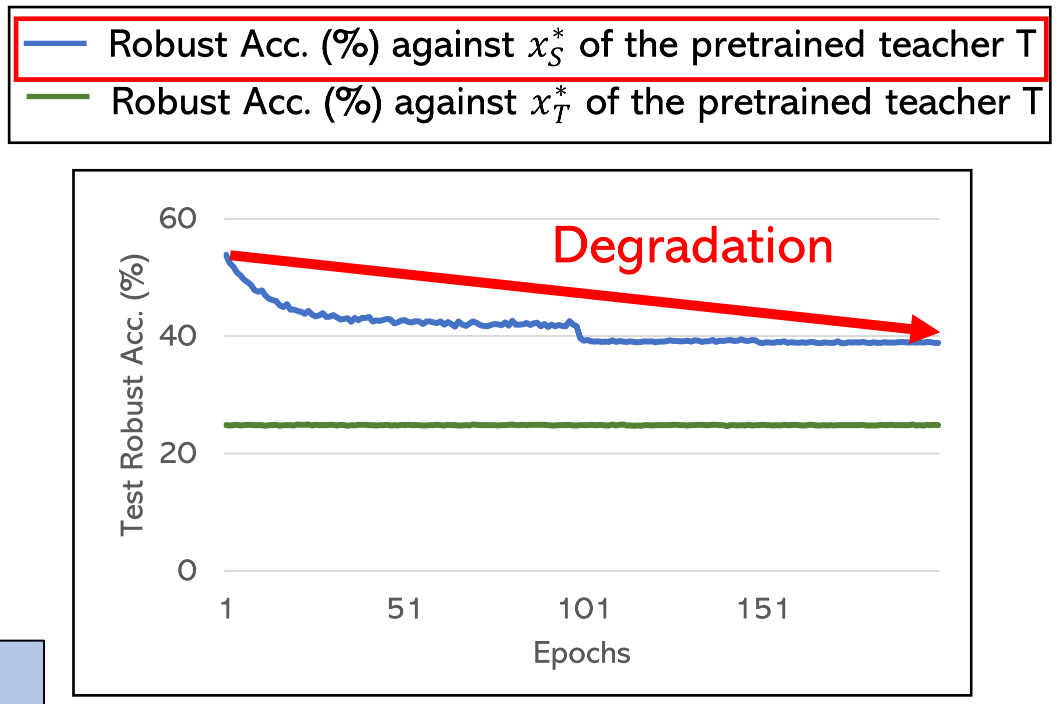
Student (to be trained)

Background

- **Limitations in the previous works**
  - The pretrained robust teacher model keeps losing its ability to defend against adversarial examples (x_S^*) of the student model.



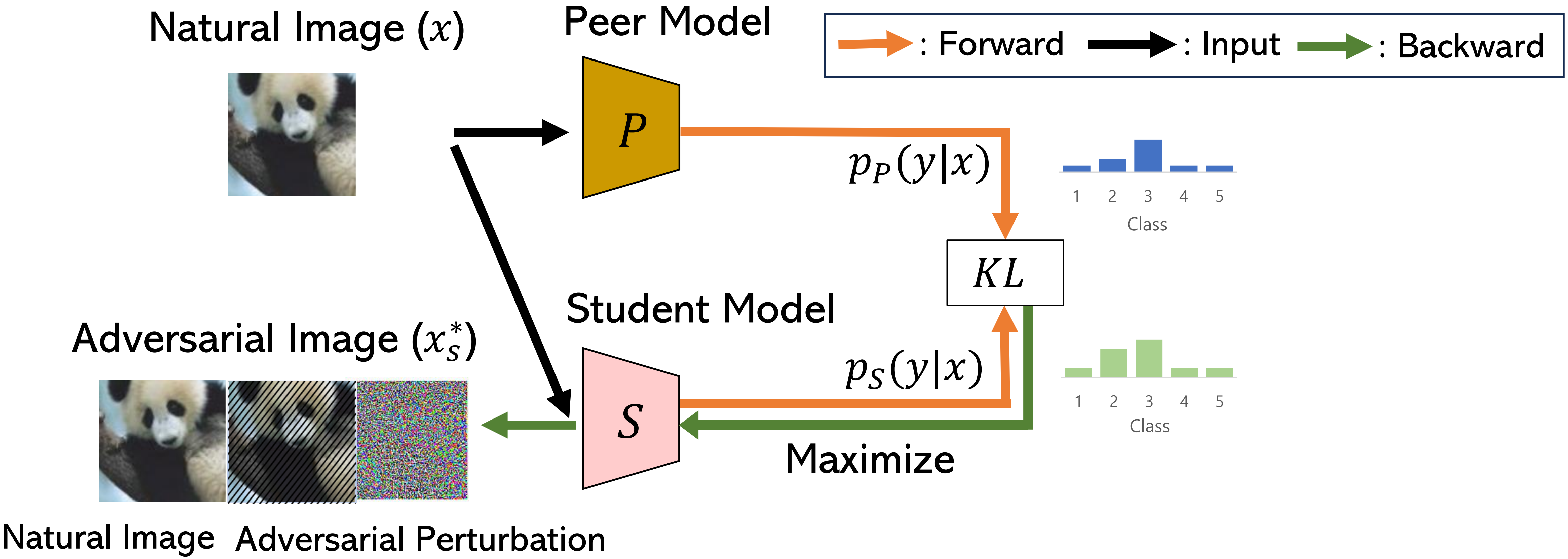
At later epochs, x_S^* fools the teacher.



[CIFAR-100 result]

Proposed Method

- **Peer Tutoring**
  - Adversarial Example Generation.
  - The student model uses the predictions of the peer model as guidance.



Natural Image (x)

Peer Model

Student Model

Adversarial Image (x_S^*)

Natural Image

Adversarial Perturbation

: Forward

: Input

: Backward

P

S

$p_P(y|x)$

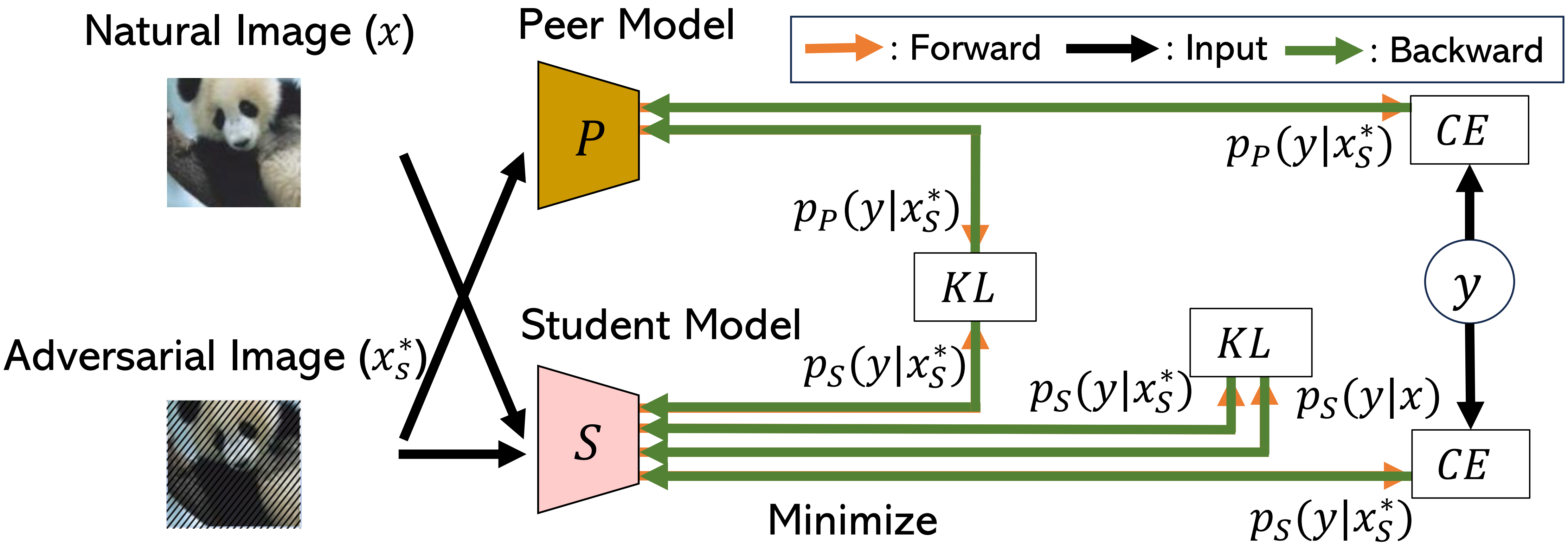
$p_S(y|x)$

KL

Maximize

Proposed Method

- **Peer Tutoring**
  - Weight Optimization.
  - The student and the peer transfer their own knowledge to each other.



Natural Image (x)

Peer Model

: Forward : Input : Backward

P

CE

$p_P(y|x_S^*)$

$p_P(y|x_S^*)$

KL

y

Adversarial Image (x_S^*)

Student Model

$p_S(y|x_S^*)$

KL

S

$p_S(y|x_S^*)$

$p_S(y|x)$

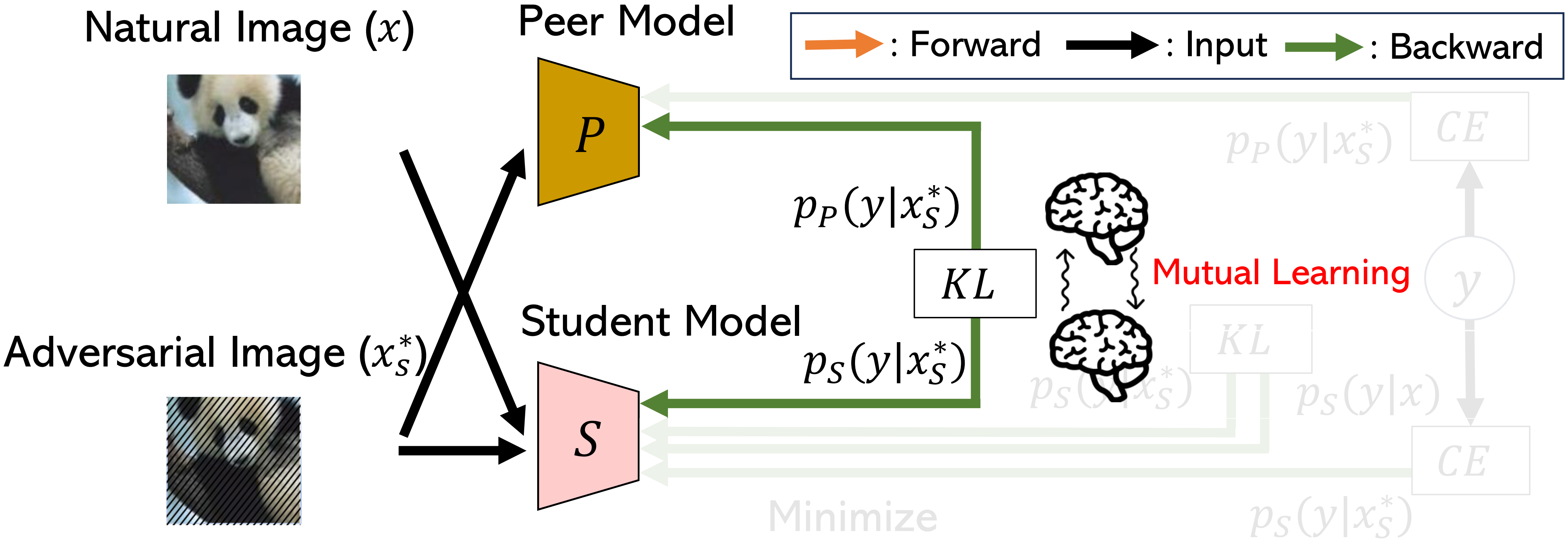
CE

Minimize

$p_S(y|x_S^*)$

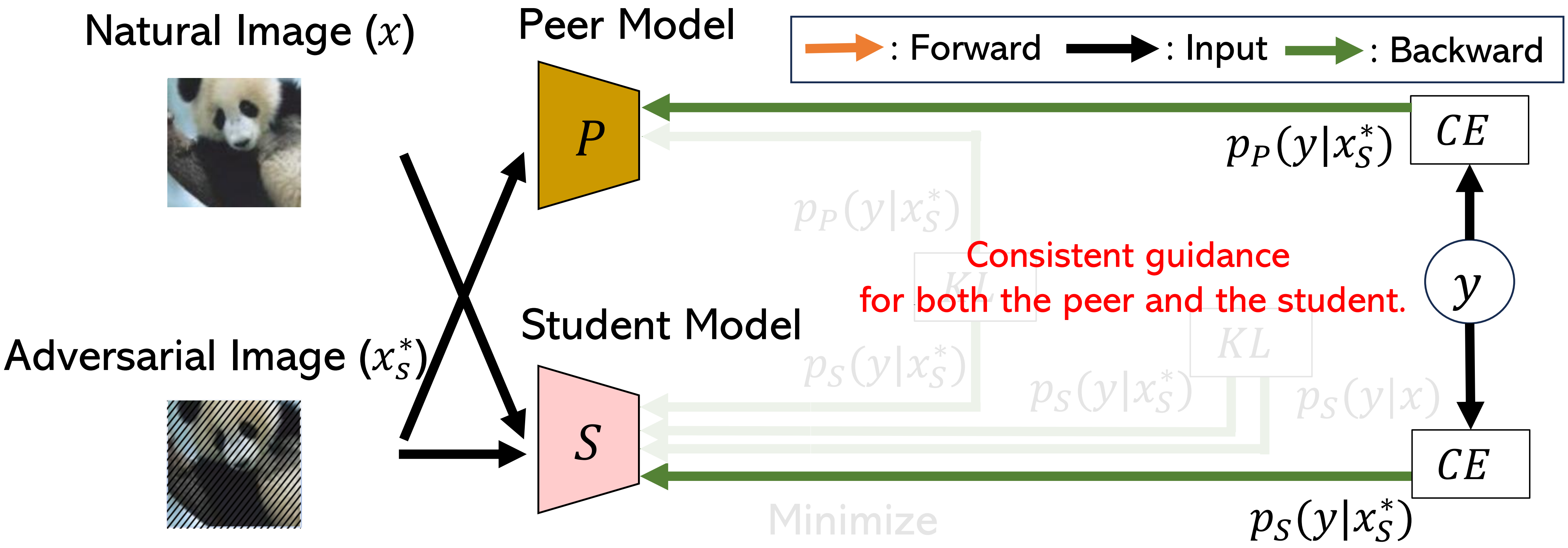
Proposed Method

- **Peer Tutoring**
  - Weight Optimization.
  - The student and the peer transfer their own knowledge to each other.



Proposed Method

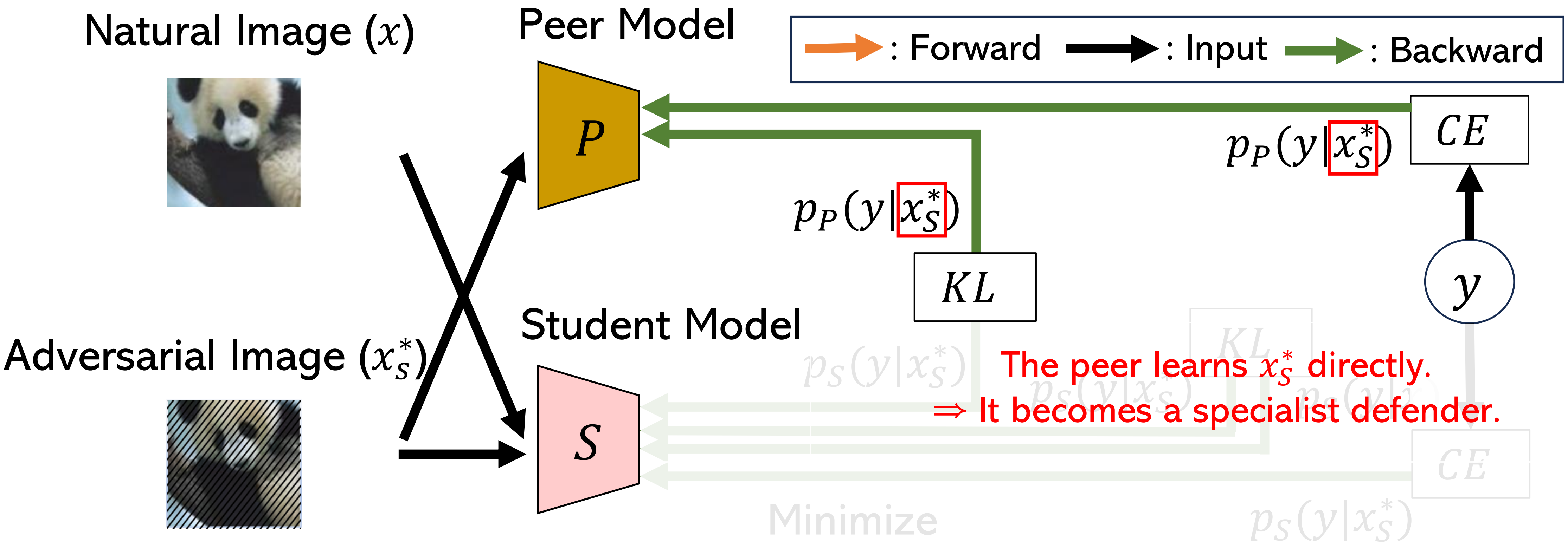
- **Peer Tutoring**
  - Weight Optimization.
  - The student and the peer transfer their own knowledge to each other.



Consistent guidance for both the peer and the student.

Proposed Method

- **Peer Tutoring**
  - Weight Optimization.
  - The student and the peer transfer their own knowledge to each other.

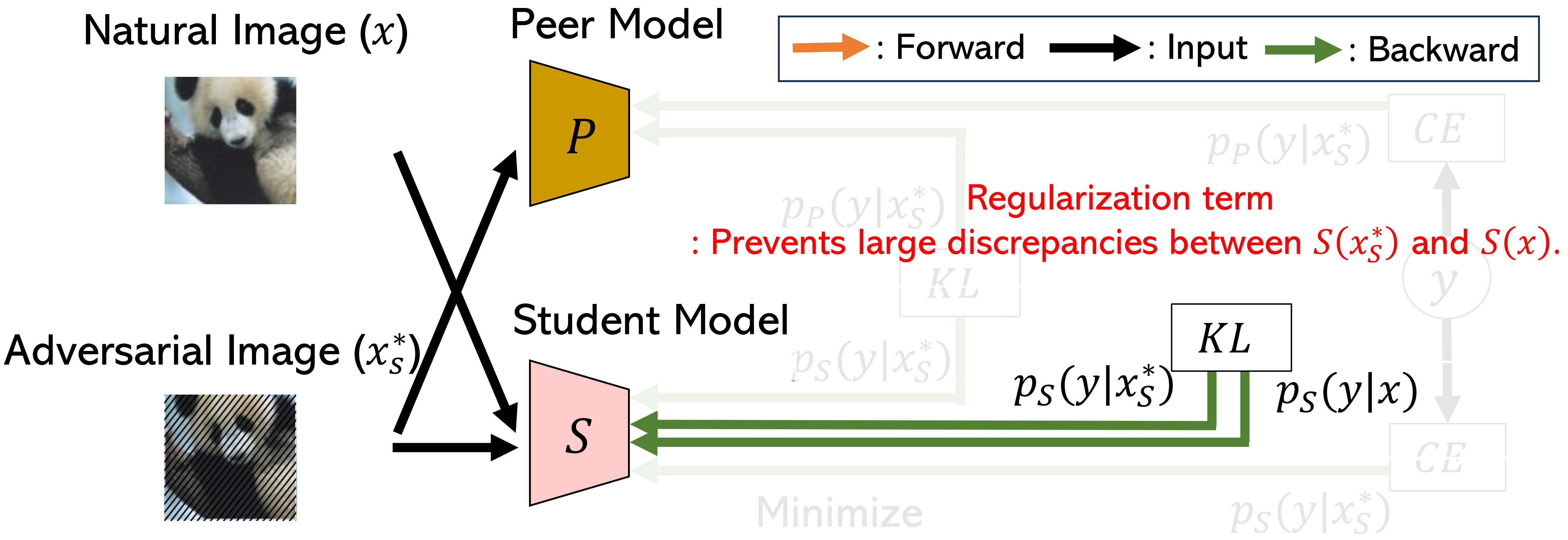


The peer learns x_S^* directly.

⇒ It becomes a specialist defender.

Proposed Method

- **Peer Tutoring**
  - Weight Optimization.
  - The student and the peer transfer their own knowledge to each other.

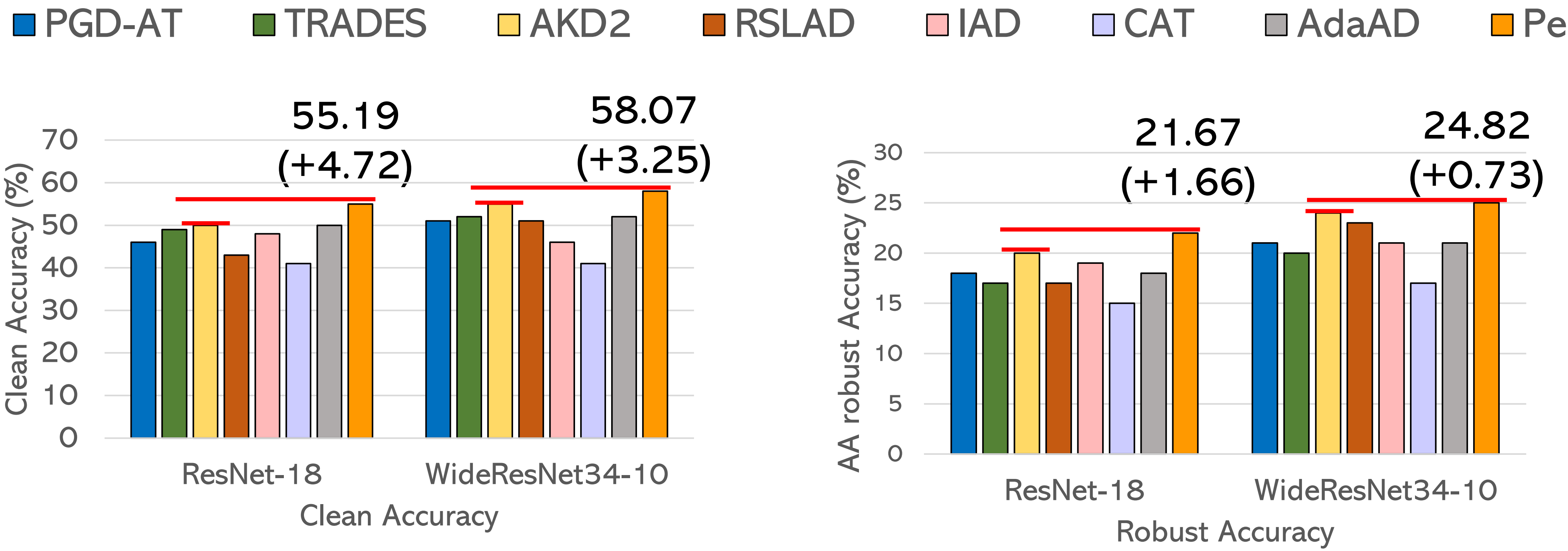


Regularization term

: Prevents large discrepancies between $S(x_S^*)$ and $S(x)$.

Experimental Results

- **TinyImageNet result**
  - PeerAiD shows the highest AutoAttack robust accuracy compared to other baselines, while also providing higher clean accuracy.



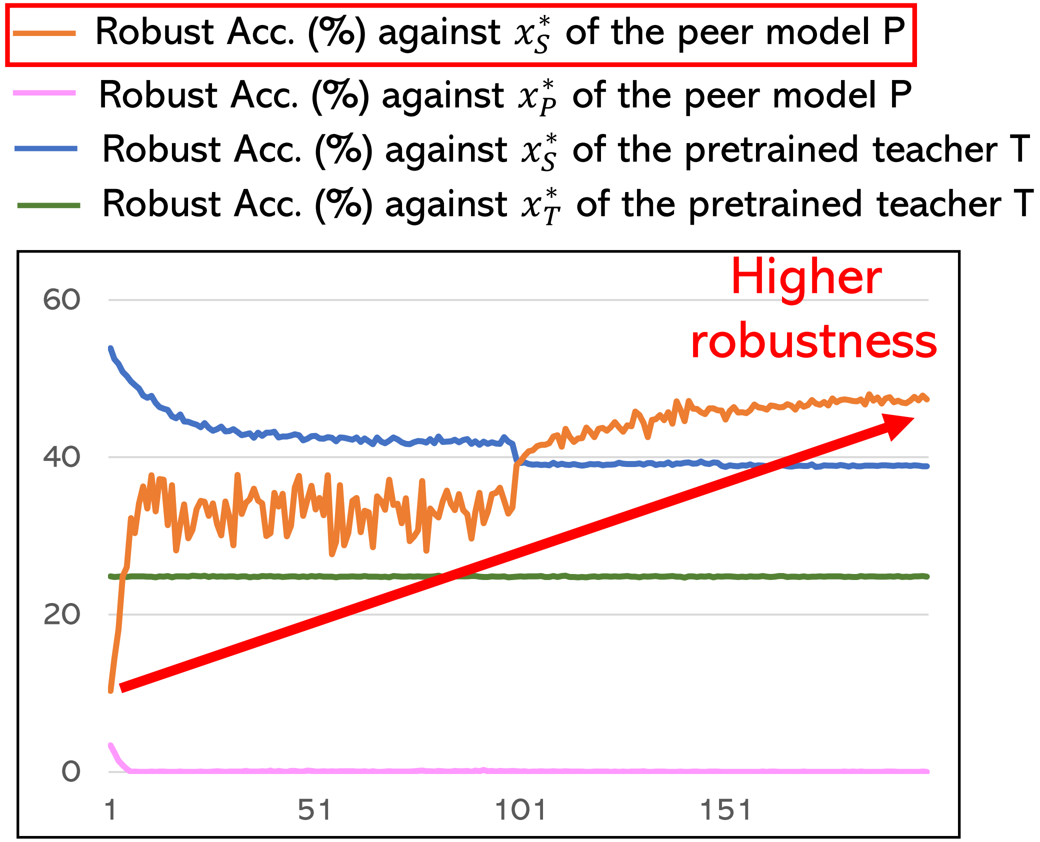
Experimental Results

- **Characteristic of the peer model**

① Specialist who defends against adversarial examples of the student model.

- No tradeoff between the robustness and clean accuracy.

② High clean accuracy



f	$Rob_f(x_f^*)$	$Rob_f(x_s^*)$
Peer tutor	0.00	<u>69.19</u>
Pretrained robust teacher	24.15	39.46

Student's Robust acc.

<u>29.69</u>

24.48

[CIFAR-100, ResNet-18 result]

- $Rob_f(\cdot)$: the robust accuracy of f.
- S : the student model.
- Peer's Clean acc : 75.63 > 75.48 (Naturally trained)

Conclusion

- We propose a novel online adversarial distillation method, PeerAiD

- The peer model specializes in defending against the student model's attack samples.

- PeerAiD improves AA robust accuracy by 1.66%p and clean accuracy by 4.72%p.

Thanks!

Jaewon Jung @ Seoul National University

Email: jungjaewon@snu.ac.kr

Code: https://github.com/jaewonalive/PeerAiD

Profile