

# Generating Less Certain Adversarial Examples Improves Robust Generalization

*Minxing Zhang, Michael Backes, Xiao Zhang*

CISPA Helmholtz Center for Information Security



# What Does Our Work Do?

- **Preliminary**
- **Motivation**
- **Adversarial Certainty**
- **Our Method**
- **Empirical Evidence**
- **Conclusion**



# Preliminary

- Robust Overfitting [1]:

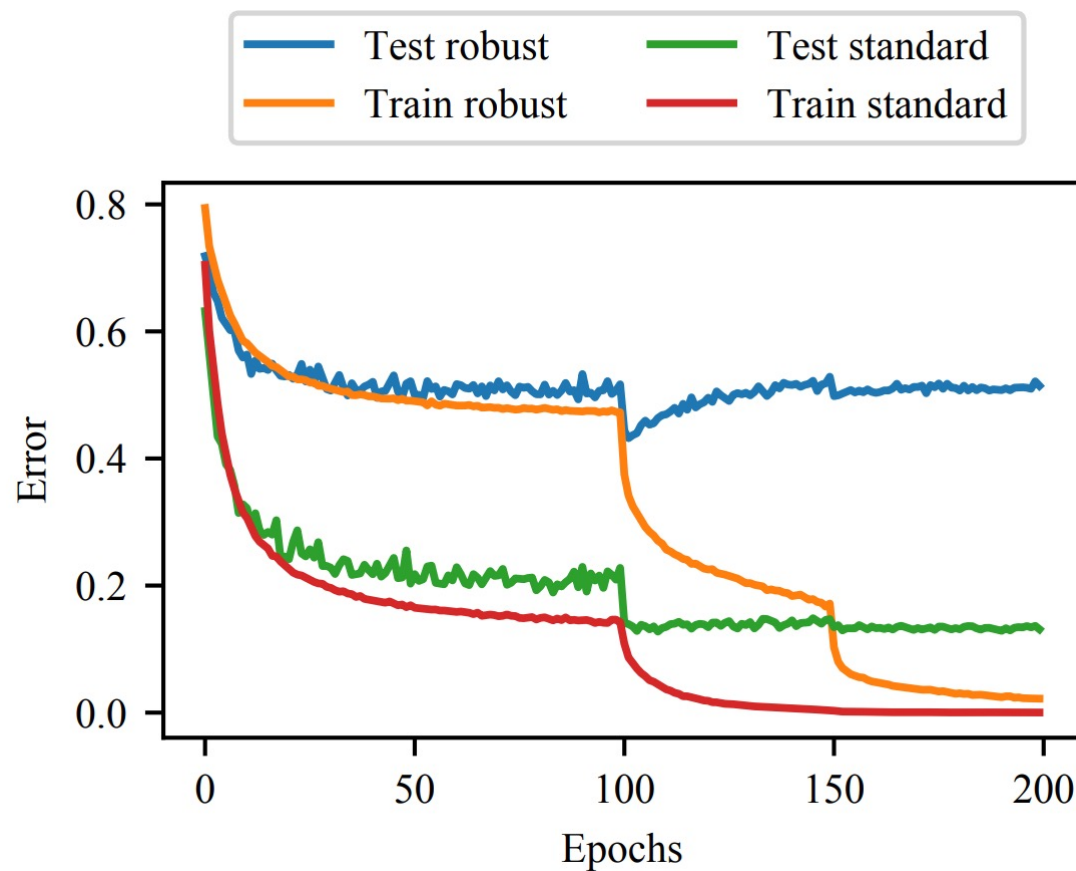
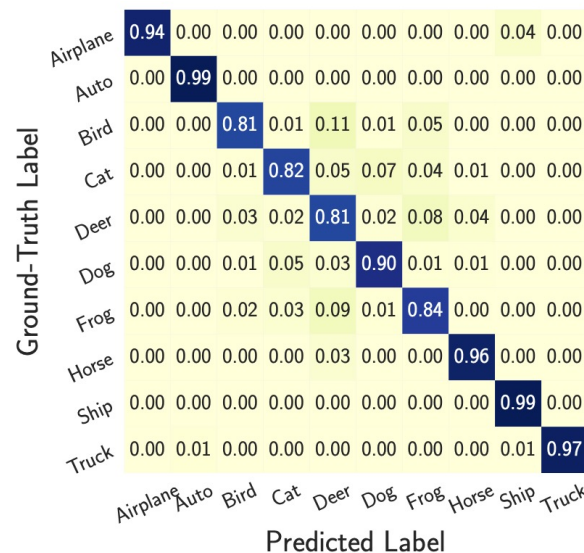


Image Source: [1]

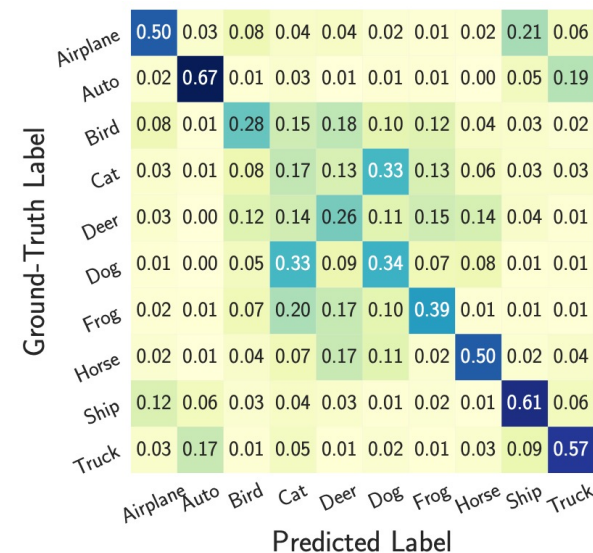


# Motivation

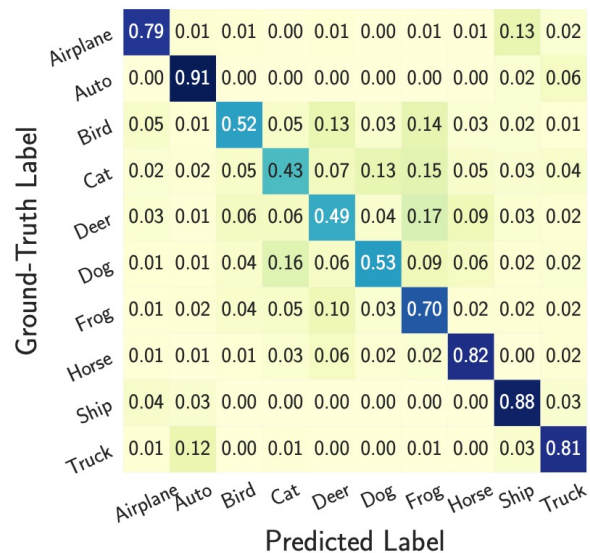
- Heatmap Visualization
  - Adversarially-perturbed CIFAR-10



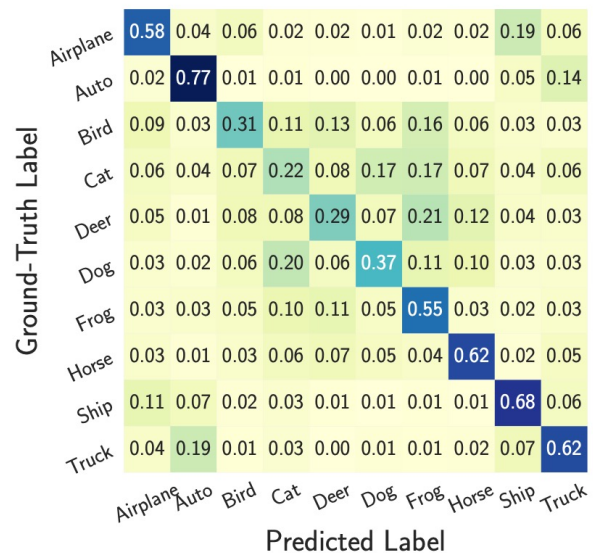
(a) Last Model (Train)



(b) Last Model (Test)



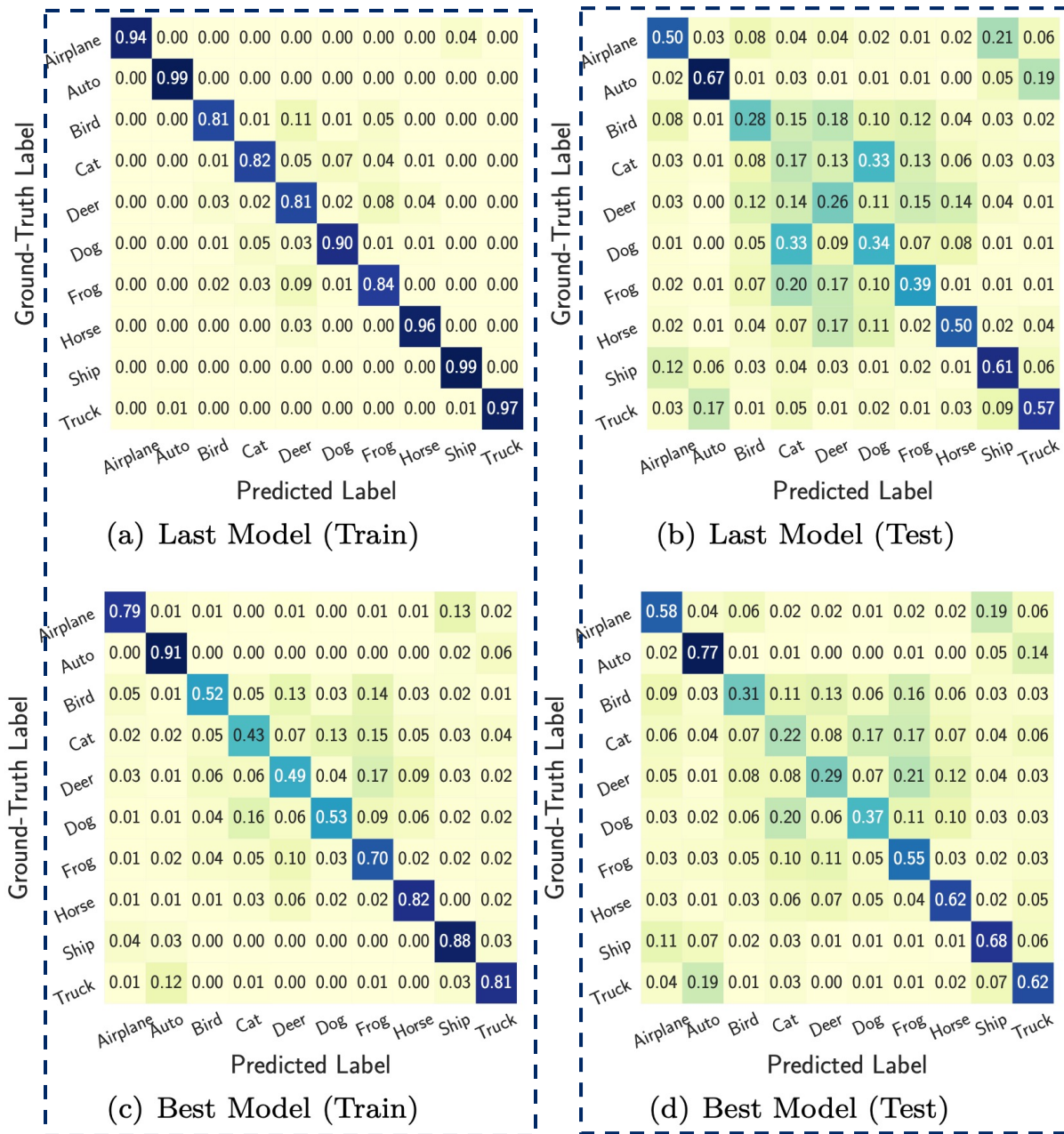
(c) Best Model (Train)



(d) Best Model (Test)

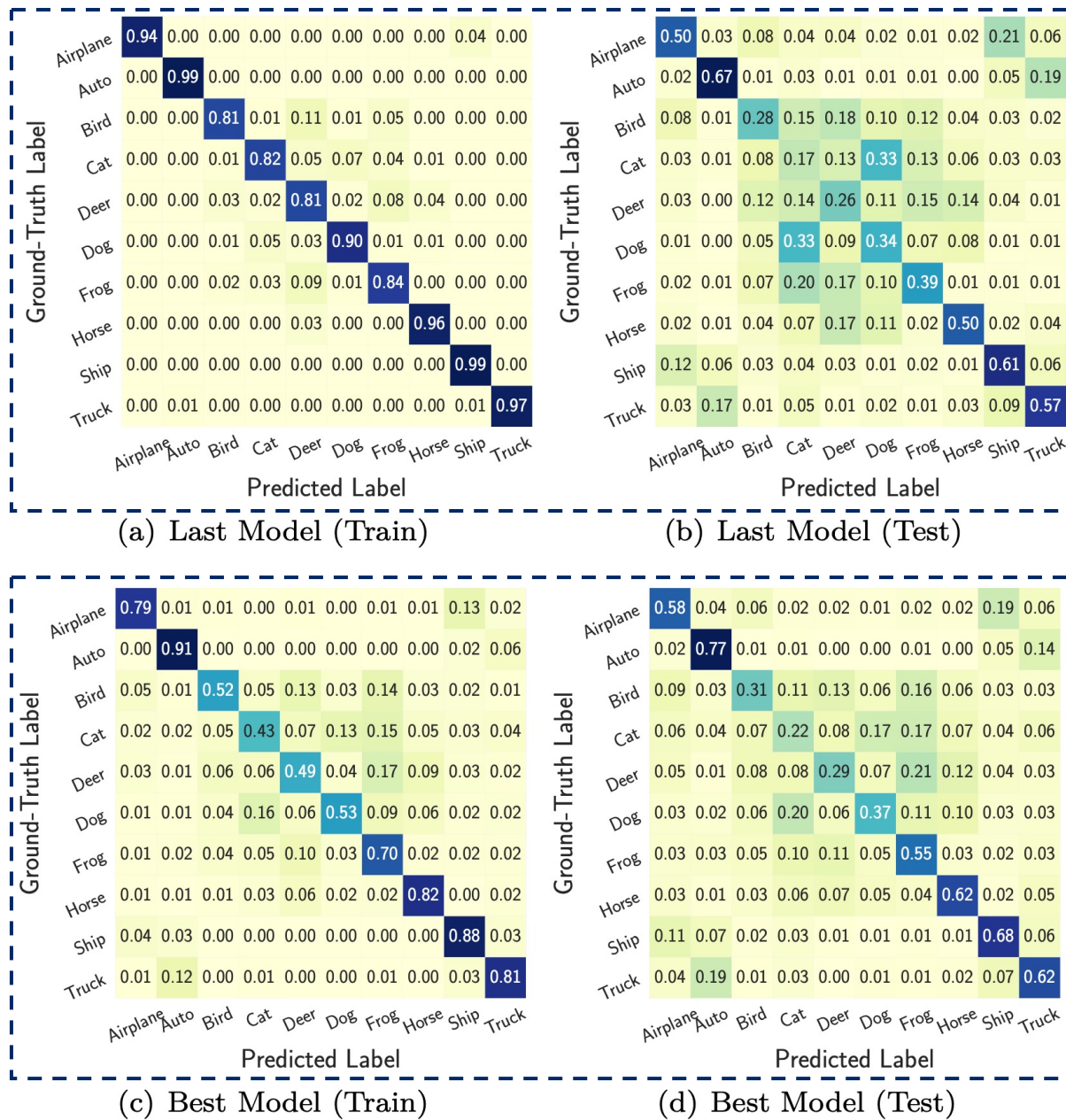


# Motivation





# Motivation





# Motivation

Ground-Truth Label	Airplane	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00
	Auto	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Bird	0.00	0.00	0.81	0.01	0.11	0.01	0.05	0.00	0.00	0.00
	Cat	0.00	0.00	0.01	0.82	0.05	0.07	0.04	0.01	0.00	0.00
	Deer	0.00	0.00	0.03	0.02	0.81	0.02	0.08	0.04	0.00	0.00
	Dog	0.00	0.00	0.01	0.05	0.03	0.90	0.01	0.01	0.00	0.00
	Frog	0.00	0.00	0.02	0.03	0.09	0.01	0.84	0.00	0.00	0.00

Ground-Truth Label	Airplane	0.50	0.03	0.08	0.04	0.04	0.02	0.01	0.02	0.21	0.06
	Auto	0.02	0.67	0.01	0.03	0.01	0.01	0.01	0.00	0.05	0.19
	Bird	0.08	0.01	0.28	0.15	0.18	0.10	0.12	0.04	0.03	0.02
	Cat	0.03	0.01	0.08	0.17	0.13	0.33	0.13	0.06	0.03	0.03
	Deer	0.03	0.00	0.12	0.14	0.26	0.11	0.15	0.14	0.04	0.01
	Dog	0.01	0.00	0.05	0.33	0.09	0.34	0.07	0.08	0.01	0.01
	Frog	0.02	0.01	0.07	0.20	0.17	0.10	0.39	0.01	0.01	0.01

*Overconfidence Compromises Robustness*

Ground-Truth Label	Cat	0.02	0.02	0.05	0.43	0.07	0.13	0.15	0.05	0.03	0.04
	Deer	0.03	0.01	0.06	0.06	0.49	0.04	0.17	0.09	0.03	0.02
	Dog	0.01	0.01	0.04	0.16	0.06	0.53	0.09	0.06	0.02	0.02
	Frog	0.01	0.02	0.04	0.05	0.10	0.03	0.70	0.02	0.02	0.02
	Horse	0.01	0.01	0.01	0.03	0.06	0.02	0.02	0.82	0.00	0.02
	Ship	0.04	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.03
	Truck	0.01	0.12	0.00	0.01	0.00	0.00	0.01	0.00	0.03	0.81
		Airplane	Auto	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
		Predicted Label									

(c) Best Model (Train)

Ground-Truth Label	Cat	0.06	0.04	0.07	0.22	0.08	0.17	0.17	0.07	0.04	0.06
	Deer	0.05	0.01	0.08	0.08	0.29	0.07	0.21	0.12	0.04	0.03
	Dog	0.03	0.02	0.06	0.20	0.06	0.37	0.11	0.10	0.03	0.03
	Frog	0.03	0.03	0.05	0.10	0.11	0.05	0.55	0.03	0.02	0.03
	Horse	0.03	0.01	0.03	0.06	0.07	0.05	0.04	0.62	0.02	0.05
	Ship	0.11	0.07	0.02	0.03	0.01	0.01	0.01	0.01	0.68	0.06
	Truck	0.04	0.19	0.01	0.03	0.00	0.01	0.01	0.02	0.07	0.62
		Airplane	Auto	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
		Predicted Label									

(d) Best Model (Test)



# Adversarial Certainty

**Definition 1** (Adversarial Certainty). Let  $\mathcal{X}$  be the input space and  $\mathcal{Y} = \{1, 2, \dots, m\}$  be the label space. Suppose  $\mu$  is the underlying distribution and  $\mathcal{S}$  is a set of sampled examples. Let  $\epsilon \geq 0$ ,  $\Delta$  be the perturbation metric. For any  $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ , we define the *adversarial certainty* of  $f_\theta$  as:

$$\text{AC}_\epsilon(f_\theta; \hat{\mu}_{\mathcal{S}}, \mathcal{A}) = \frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x}, y) \in \mathcal{S}} \text{Var}(F_\theta[\mathcal{A}(\mathbf{x}; y, f_\theta, \epsilon)]),$$

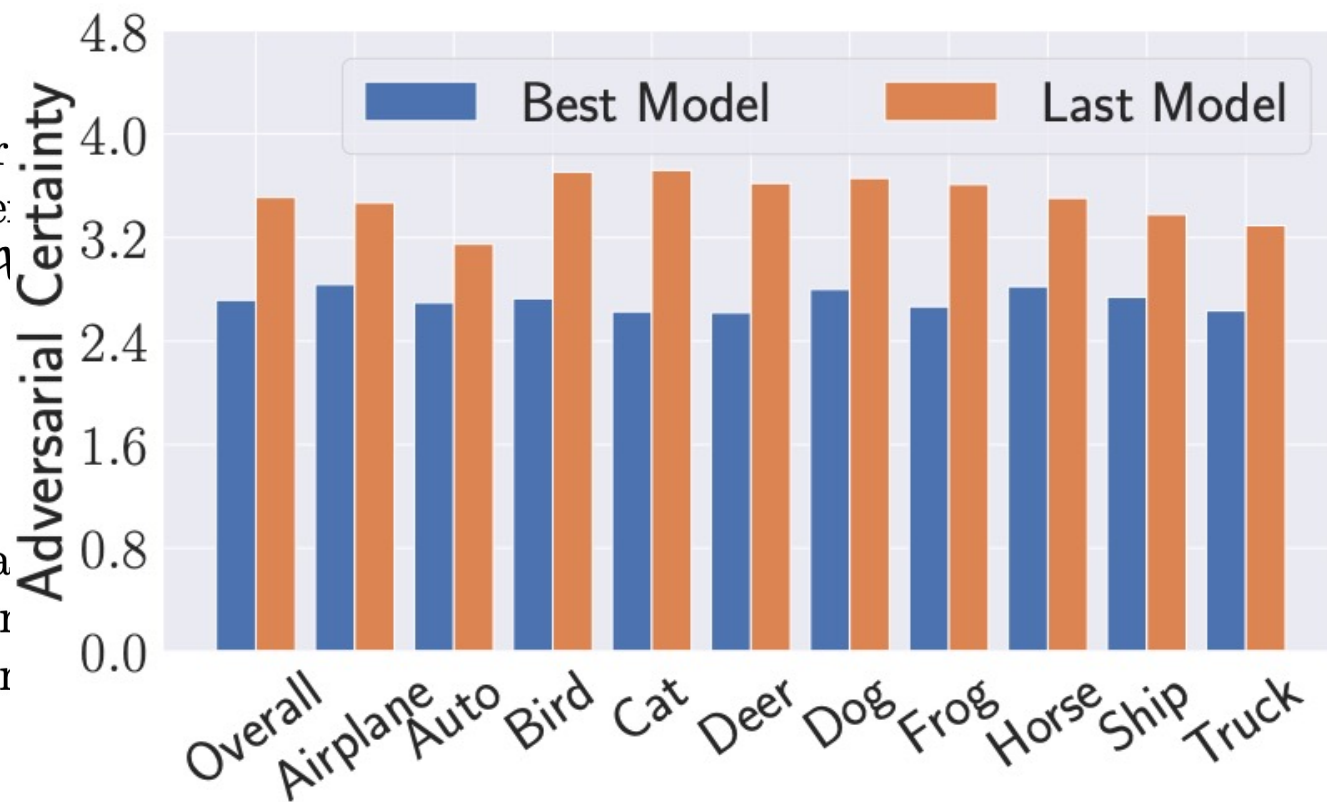
where  $\mathcal{A}$  denotes an attack method such as PGD attacks for generating adversarial examples,  $F_\theta : \mathcal{X} \rightarrow \mathbb{R}^m$  represents the mapping from the input space  $\mathcal{X}$  to the logit layer of  $f_\theta$ , and  $\text{Var}(\mathbf{u}) = \sum_{k \in [m]} (u_k - \bar{u})^2 / m$ , with  $u_k$  and  $\bar{u}$  denoting the  $k$ -th element and mean of  $\mathbf{u} \in \mathbb{R}^m$  respectively.



# Adversarial Certainty

**Definition 1** (Adversarial Certainty)  
Suppose  $\mu$  is the underlying metric. For any  $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$

where  $\mathcal{A}$  denotes an adversarial model,  $f_\theta$  represents the mapping from input  $x$  to output  $y$  with  $u_k$  and  $\bar{u}$  denoting the

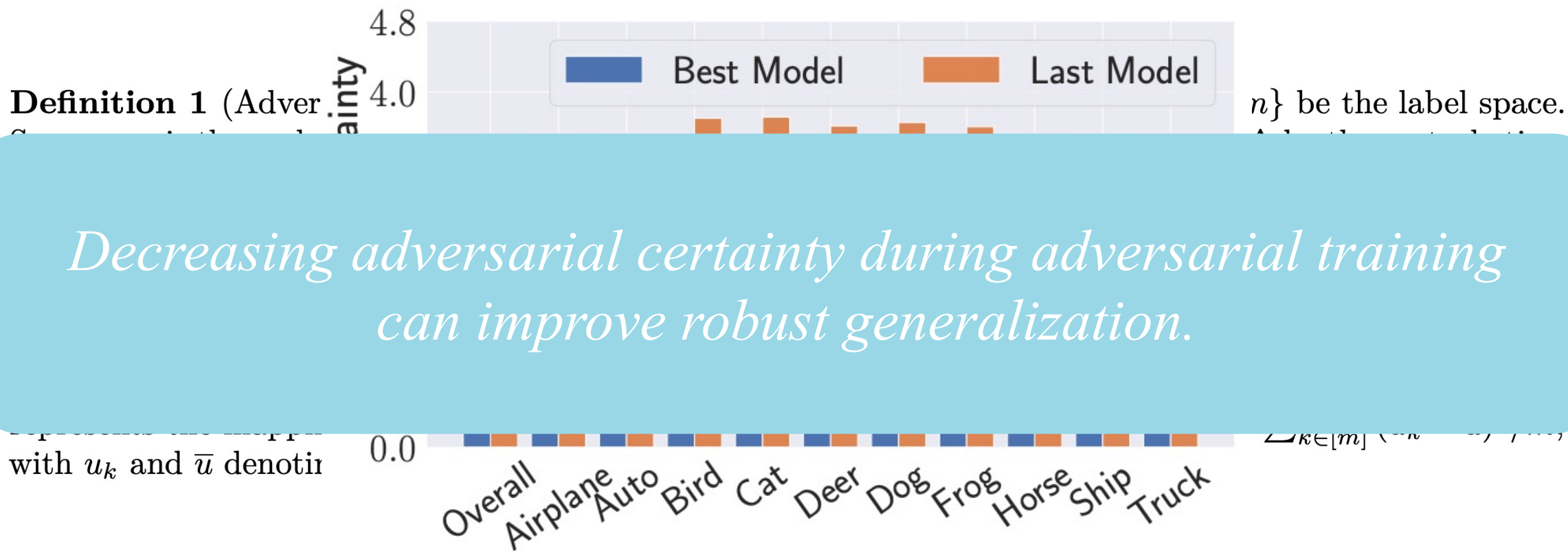


$\mathcal{Y}$  be the label space.  
 $\Delta$  be the perturbation

samples,  $F_\theta : \mathcal{X} \rightarrow \mathbb{R}^m$   
 $\sum_{k \in [m]} (u_k - \bar{u})^2 / m,$



# Adversarial Certainty





# Our Method

- **Decrease Adversarial Certainty (DAC)**
  - Find less certain adversarial examples that are used to improve robust generalization

$$\min_{\theta \in \Theta} \frac{1}{|\mathcal{S}_{tr}|} \sum_{(\mathbf{x}, y) \in \mathcal{S}_{tr}} \max_{\mathbf{x}' \in \mathcal{B}_\epsilon(\mathbf{x})} L(f_{\theta'}, \mathbf{x}', y), \text{ where } \theta' = \operatorname{argmin}_{\theta' \in \mathcal{C}(\theta)} \operatorname{AC}_\epsilon(f_\theta; \mathcal{S}_{tr}, \mathcal{A}).$$

- 
- Two-step Optimization

$$\begin{aligned} \theta_{t+0.5} &= \theta_t - \lambda \cdot \nabla_{\theta} \operatorname{AC}_\epsilon(f_\theta; \mathcal{S}_{tr}, \mathcal{A}) \Big|_{\theta=\theta_t}, \\ \theta_{t+1} &= \theta_{t+0.5} - \gamma \cdot \nabla_{\theta} L_{\text{rob}}(f_\theta; \mathcal{S}_{tr}, \mathcal{A}) \Big|_{\theta=\theta_{t+0.5}}, \end{aligned}$$



# Empirical Evidence

Architecture	Method	Clean	PGD-20	PGD-100	CW <sub>∞</sub>	AutoAttack
PRN18	AT + <b>DAC</b>	82.88 (82.68) <b>84.64 (83.55)</b>	41.51 (49.23) <b>45.55 (52.20)</b>	40.96 (48.92) <b>44.94 (51.87)</b>	41.61 (48.07) <b>44.55 (50.05)</b>	39.66 (45.71) <b>42.78 (48.20)</b>
	TRADES + <b>DAC</b>	82.10 (81.33) <b>83.18 (82.80)</b>	47.44 (51.65) <b>49.32 (52.90)</b>	46.95 (51.42) <b>48.81 (52.67)</b>	46.64 (49.18) <b>48.30 (50.11)</b>	44.99 (48.06) <b>46.40 (48.96)</b>
	MART + <b>DAC</b>	80.85 (78.27) <b>81.12 (79.37)</b>	50.23 (52.28) <b>52.38 (53.25)</b>	49.71 (52.13) <b>52.04 (53.14)</b>	46.88 (47.83) <b>48.97 (49.25)</b>	44.68 (46.01) <b>47.24 (47.69)</b>
WRN34	AT + <b>DAC</b>	86.47 ( <b>85.86</b> ) <b>86.48</b> (85.10)	47.25 (55.31) <b>52.02 (57.93)</b>	46.73 (55.00) <b>51.69 (57.68)</b>	47.85 (54.04) <b>51.51 (54.98)</b>	45.84 (51.94) <b>49.75 (53.33)</b>
	TRADES + <b>DAC</b>	83.37 (81.40) <b>85.04 (84.55)</b>	51.51 (58.78) <b>58.97 (60.96)</b>	51.28 (58.72) <b>58.97 (60.81)</b>	49.26 (53.33) <b>52.79 (55.00)</b>	47.74 (52.63) <b>51.80 (53.99)</b>
	MART + <b>DAC</b>	83.11 ( <b>83.30</b> ) <b>84.69</b> (80.09)	48.93 (58.13) <b>52.00 (59.31)</b>	48.31 (57.75) <b>51.32 (59.26)</b>	46.32 (52.22) <b>49.50 (53.02)</b>	44.89 (50.31) <b>47.65 (51.48)</b>



## Conclusion

- Notion of adversarial certainty
- Importance of generating less certain adversarial examples for robust generalization
- Better understanding of robust generalization



# Thanks

**Minxing Zhang – Ph.D. Candidate**

*CISPA Helmholtz Center for Information Security*