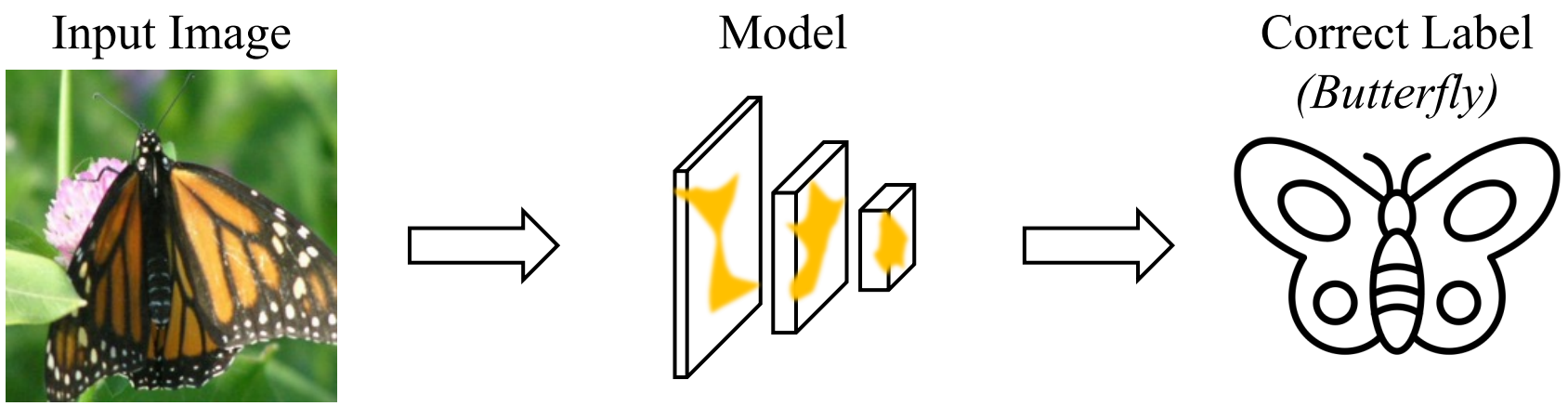
LOTUS: Evasive and Resilient Backdoor Attacks through Sub-partitioning

Siyuan Cheng, Guanhong Tao, Yingqi Liu†, Guangyu Shen, Shengwei An,
Shiwei Feng, Xiangzhe Xu, Kaiyuan Zhang, Shiqing Ma‡, Xiangyu Zhang

Email: {cheng535, taog, shen447, an93, feng292, xu1415, zhan4057, xyzhang}@cs.purdue.edu
†yingqiliu@microsoft.com ‡shiqingma@umass.edu

Backdoor Attacks

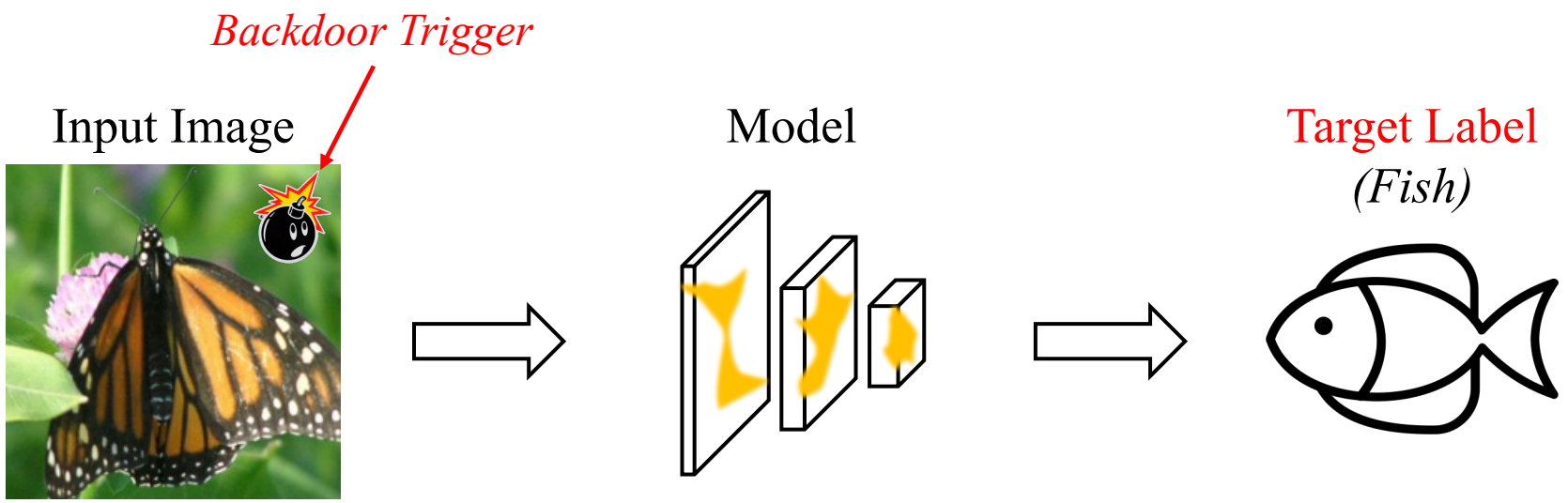
- Backdoor attack[1][2] is a prominent threat to deep learning models
  - The model performs well on normal inputs



[1] Gu, Tianyu, et al. "Badnets: Evaluating backdooring attacks on deep neural networks." *IEEE Access* 7 2019
[2] Liu, Yingqi, et al. "Trojaning attack on neural networks." *NDSS 2018*

Backdoor Attacks

- Backdoor attacks are a prominent threat to deep learning models
  - The model misclassifies inputs stamped with the backdoor trigger



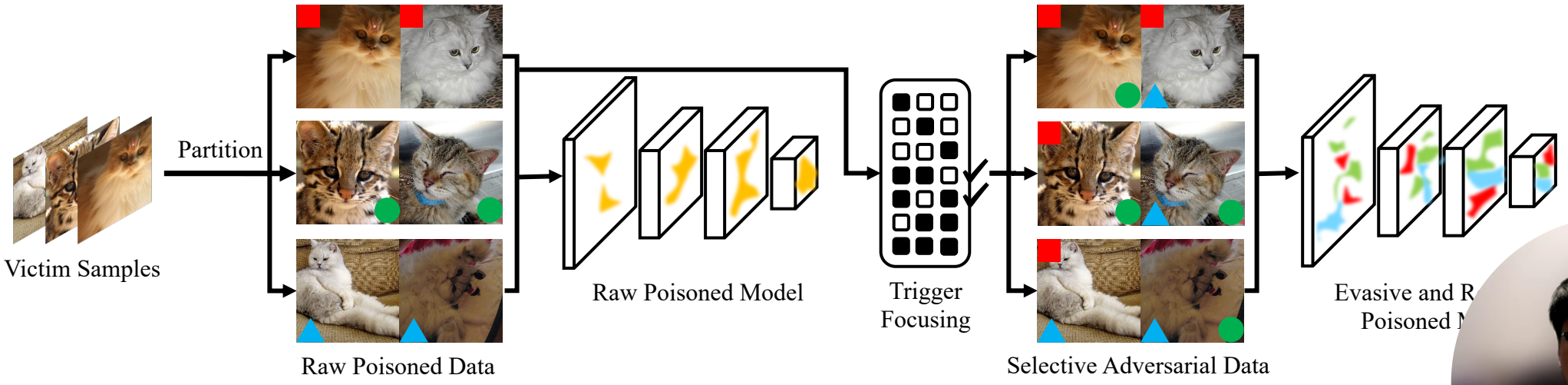
Limitation of Existing Backdoor Attacks

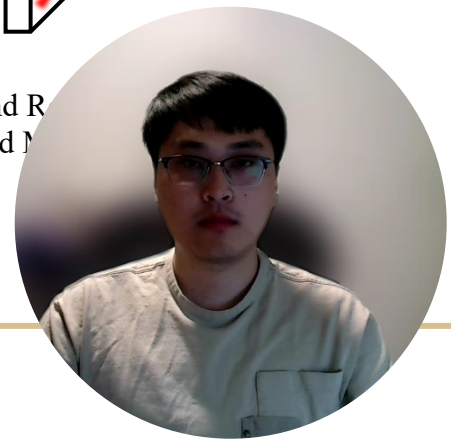
- Fixed trigger patterns are not evasive
  - Trigger inversion[1][2] is effective against fixed patch[3] or noise[4] backdoors
- Sample-specific triggers are not resilient
  - WaNet[5] evades several backdoor detection methods
  - Backdoor mitigation methods can easily eliminate its attack success rate (ASR)
  - For example, fine-tuning the model with only 5% training data reduces its ASR from 91% to 1%

[1] Wang, Bolun, et al. "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks." *IEEE S&P* 2019.
[2] Liu, Yingqi, et al. "Abs: Scanning neural networks for back-doors by artificial brain stimulation." *ACM SIGSAC CCS 2019.*
[3] Gu, Tianyu, et al. "Badnets: Evaluating backdooring attacks on deep neural networks." *IEEE Access* 7 2019.
[4] Xinyun Chen, et al. "Targeted backdoor attacks on deep learning systems using data poisoning." *Arxiv 2017.*
[5] Tuan Anh Nguyen, et al. "WaNet - Imperceptible Warping-based Backdoor Attack." *ICLR 2021.*

Our Proposed Attack - LOTUS

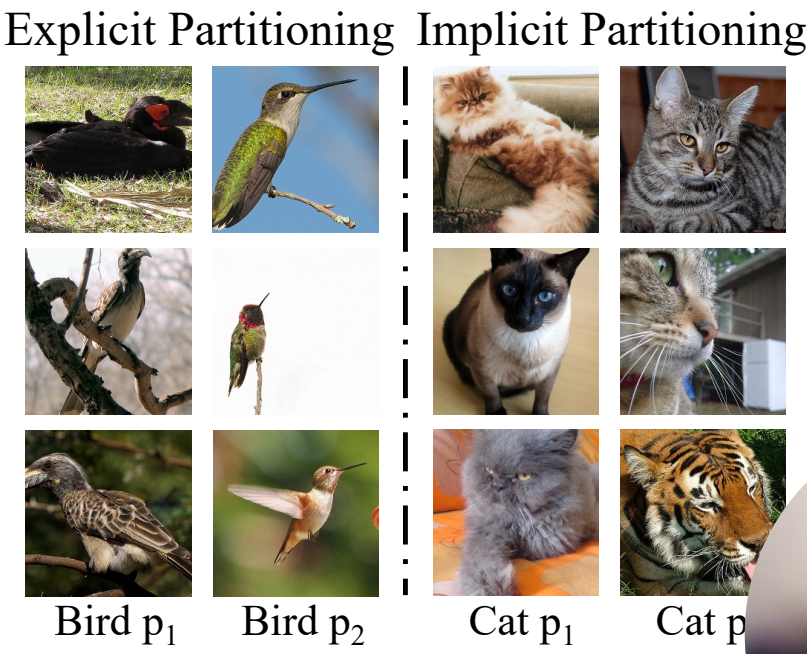
- Attack Goal – Evasive and Resilient against SOTA Defenses
  - Label-specific attack (Only attack a selected victim class)
  - Partition-specific attack (Different partitions are assigned different triggers)





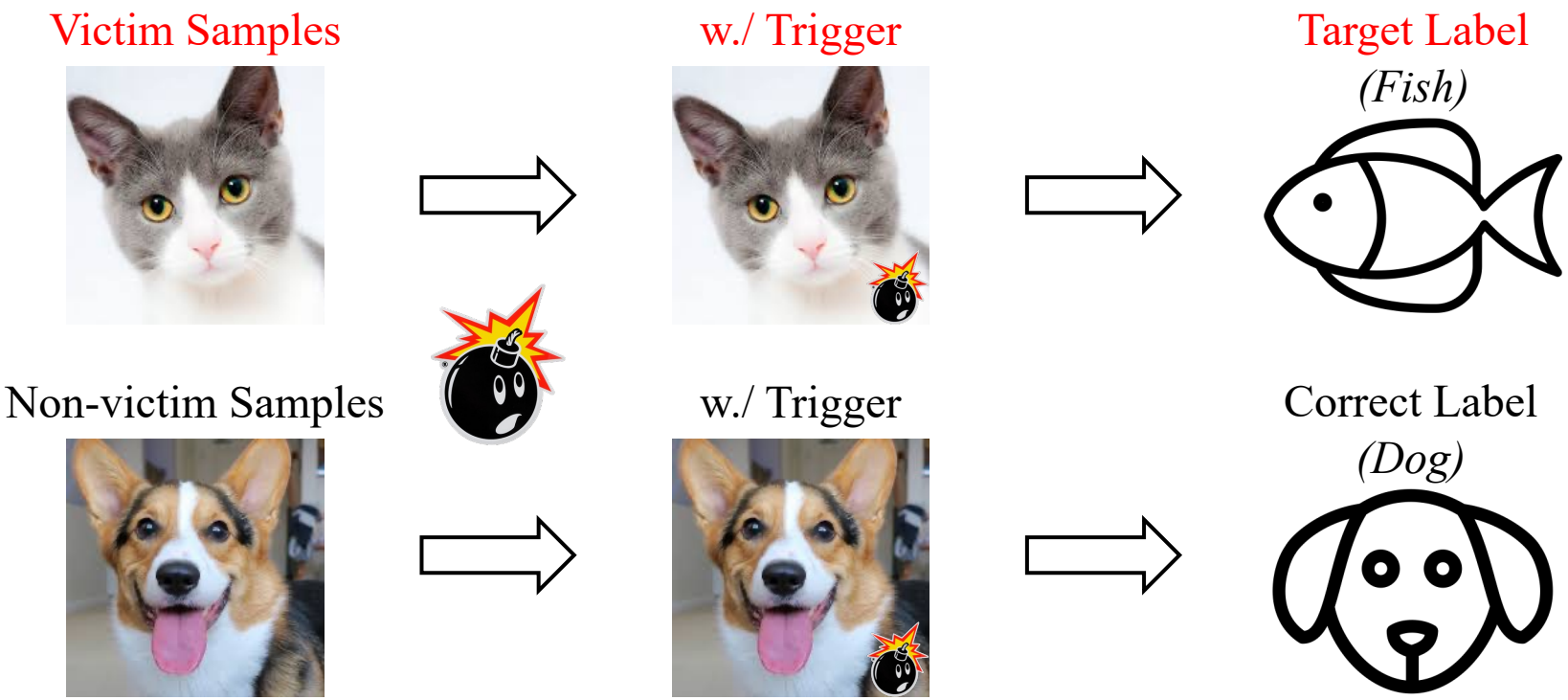
Step 1: Victim-class Sample Partitioning

- Partition the samples of the victim class into different partitions
  - Explicit partitioning
    - For example, bird species
  - Implicit partitioning
    - Partition on semantic feature maps



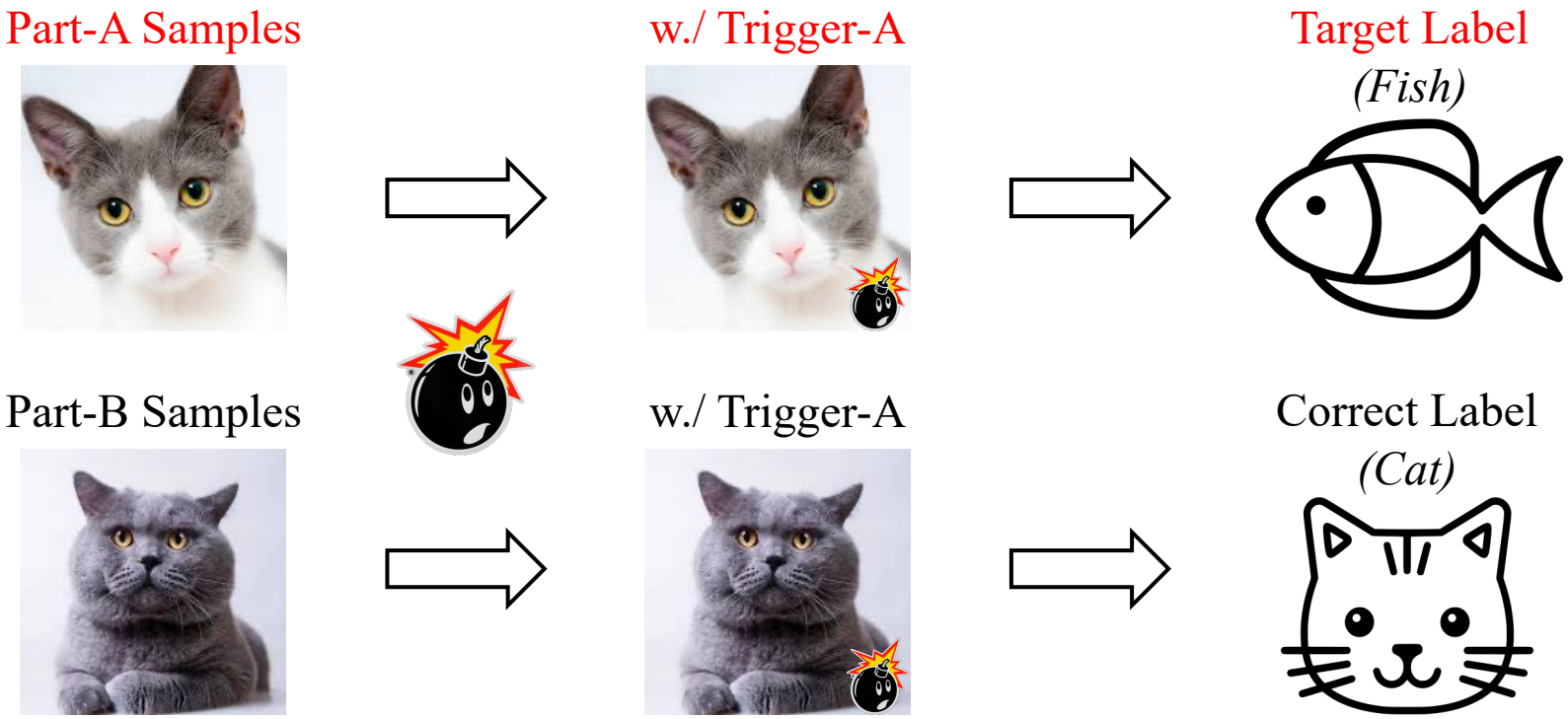
Step 2: Trigger Focusing

- Apply special adversarial training
  - Only samples from the ***victim class*** can activate the backdoor



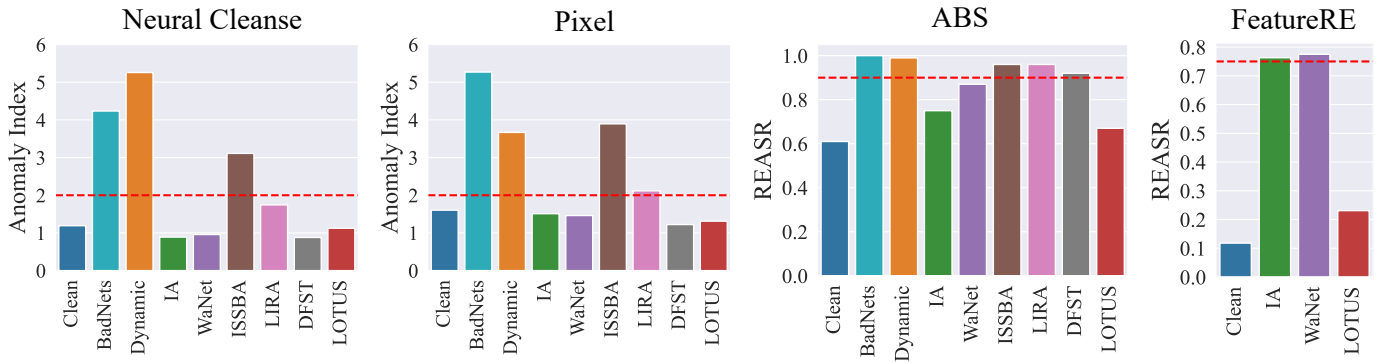
Step 2: Trigger Focusing

- Apply special adversarial training
  - Only samples stamped with the ***appropriate partition*** trigger can activate the backdoor



Evaluation

- Evasive against several backdoor detection methods



- Resilient against several backdoor mitigation methods

Attacks	No Defense		Fine-tuning		Fine-pruning		NAD		ANP	
	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
BadNets	92.02%	100.00%	89.31%	1.74%	91.70%	0.53%	87.81%	0.80%	89.15%	0.32%
Dynamic	91.81%	100.00%	88.87%	2.91%	91.39%	22.03%	89.11%	2.90%	88.25%	12.81%
IA	91.70%	99.65%	87.74%	2.78%	91.07%	0.17%	87.14%	2.29%	88.73%	1.98%
WaNet	91.22%	98.57%	89.56%	1.37%	90.22%	1.07%	89.74%	1.40%	89.07%	0.54%
ISSBA	91.67%	99.96%	87.73%	2.72%	91.12%	14.27%	87.97%	2.83%	85.64%	10.01%
LIRA	91.70%	100.00%	89.96%	2.19%	91.29%	12.14%	90.23%	2.32%	89.70%	**37.91%**
DFST	91.81%	99.97%	88.49%	22.86%	91.47%	21.61%	88.52%	24.66%	87.13%	36.17%
LOTUS	91.54%	93.80%	88.10%	**46.90%**	91.14%	**44.90%**	87.61%	**42.30%**	88.14%	34.90%

Thanks for your attention!

GitHub Repo