ODSCAN: Backdoor Scanning for Object Detection Models

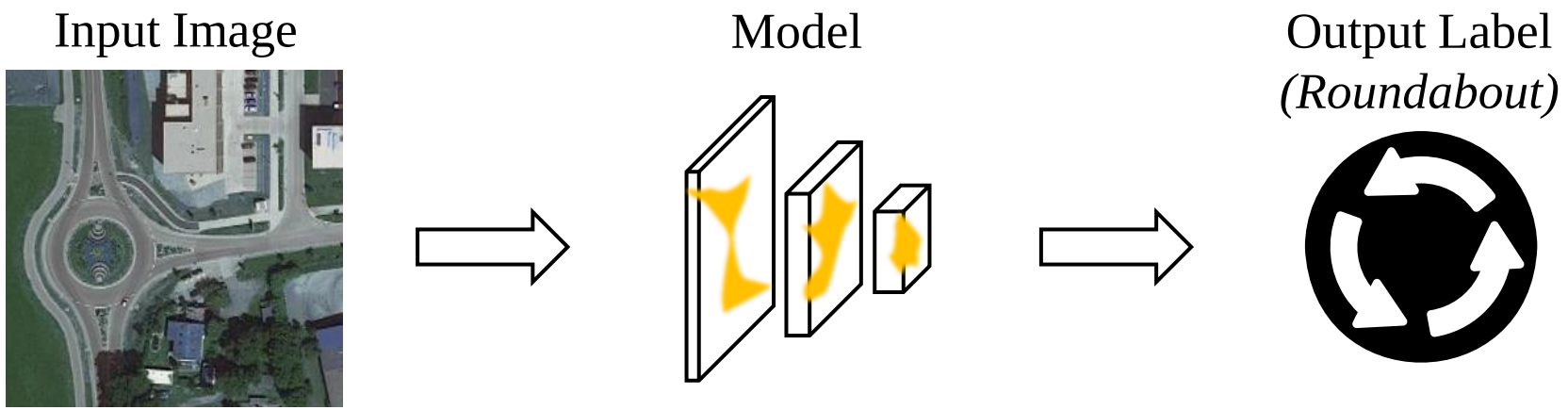
Siyuan Cheng*, Guangyu Shen*, Guanhong Tao, Kaiyuan Zhang, Zhuo Zhang,
Shengwei An, Xiangzhe Xu, Yingqi Liu†, Shiqing Ma‡, Xiangyu Zhang

Email: {cheng535, shen447, taog, zhan4057, zhan3299, an93, xu1415, xyzhang}@cs.purdue.edu
†yingqiliu@microsoft.com ‡shiqingma@umass.edu

* denotes equal contribution.

Backdoor Attacks

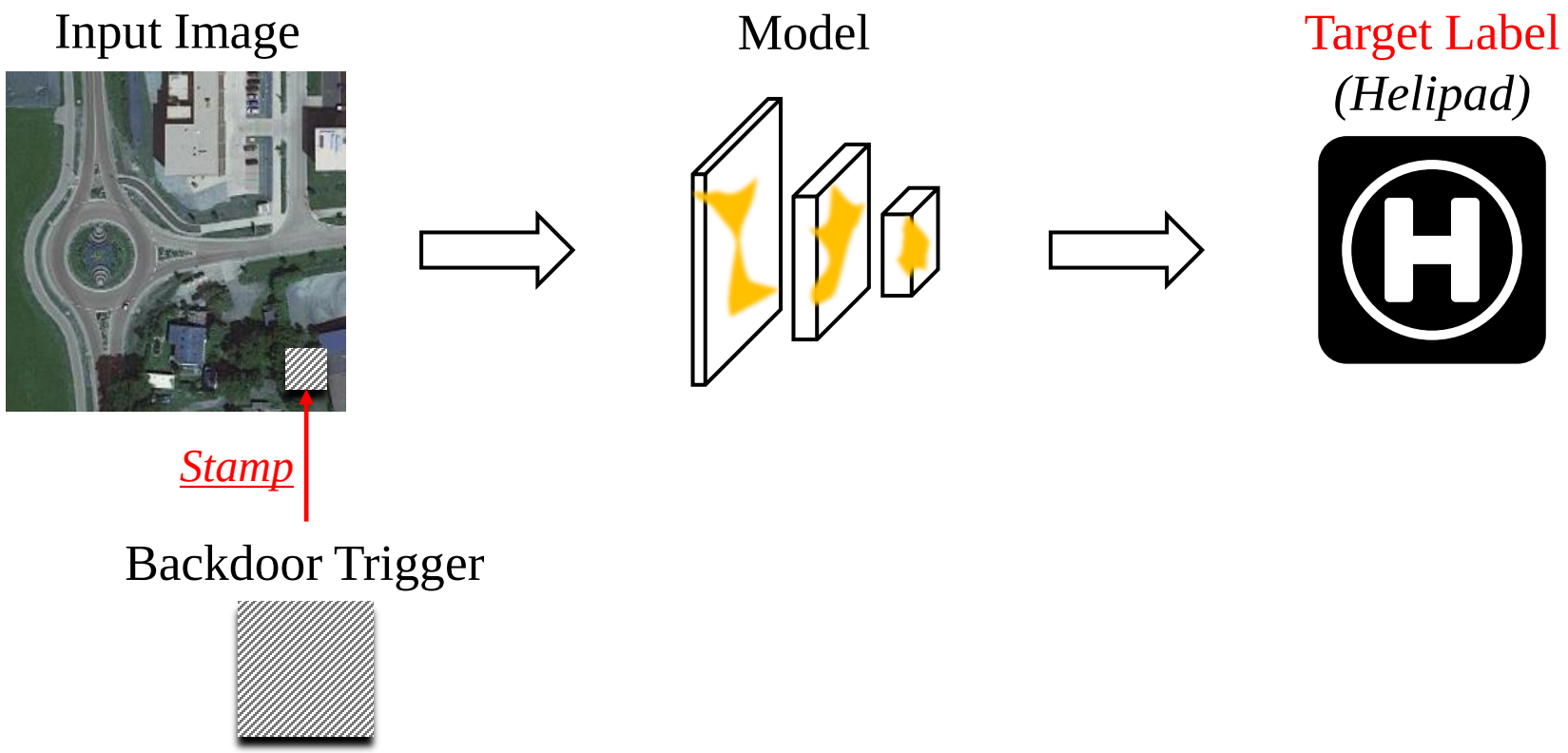
- Backdoor attacks[1][2] originally stem from the image classification task



[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*

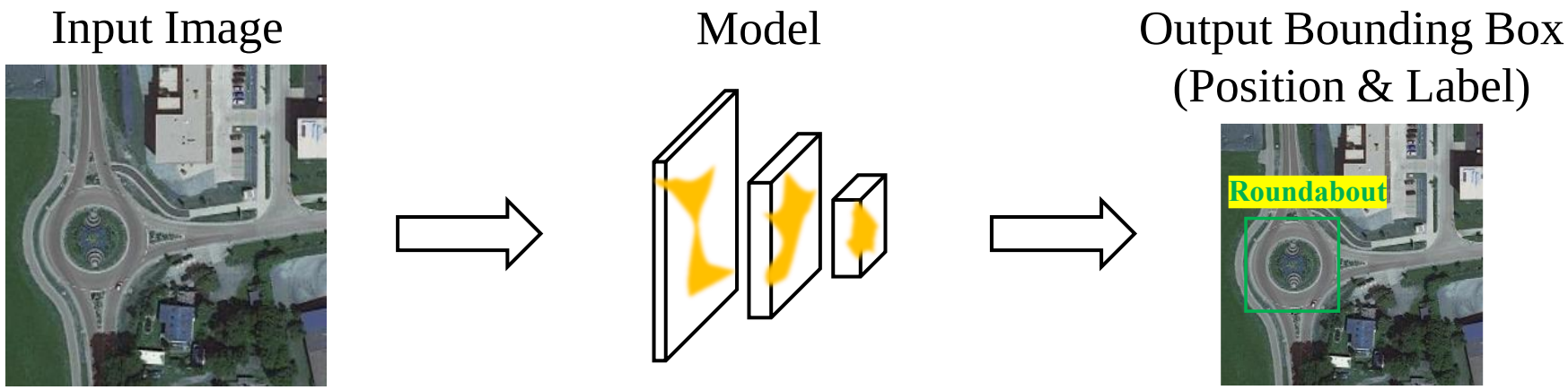
Backdoor Attacks

- Backdoor attacks originally stem from the image classification task



Backdoor Attacks in *Object Detection* (OD) Models

- Backdoor attacks become diverse in OD models
  - OD models predict bounding boxes instead of solely labels



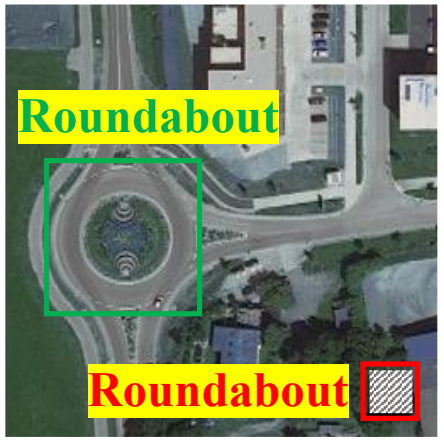
Backdoor Attacks in OD Models

- Backdoor attacks become diverse in OD models
  - Four types of backdoors[1][2][3] exploiting the bounding box prediction

Misclassification

Disappearing

Appearing

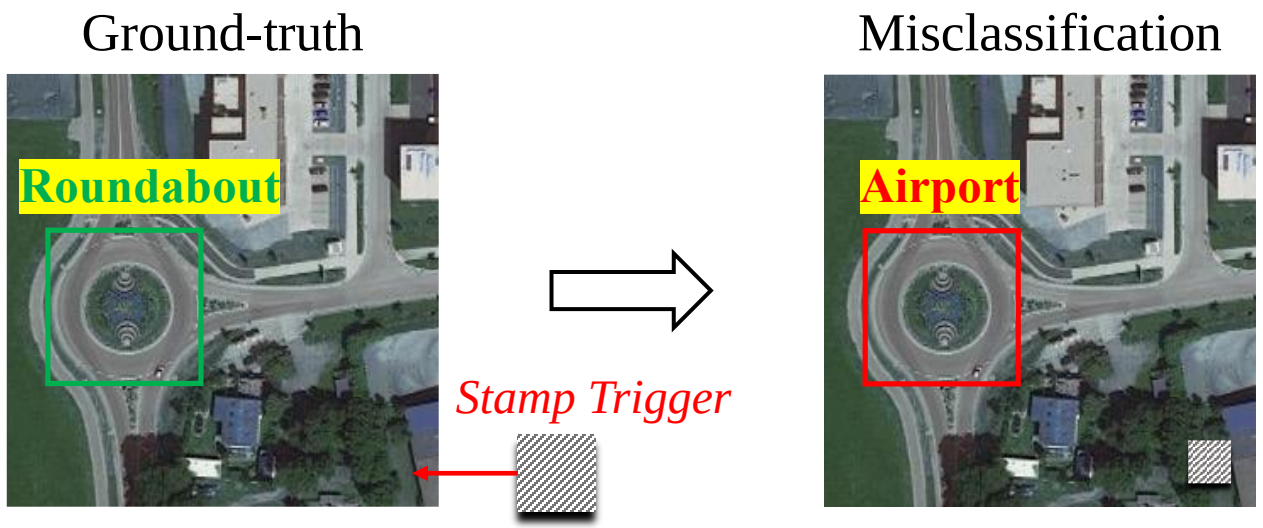


Compound

[1] NIST. "TrojAI Round-10, Round-13". *https://pages.nist.gov/trojai/*
[2] Chan, Shih-Han, et al. "Baddet: Backdoor attacks on object detection." *ECCV Workshops* 2022.
[3] Chen, Kangjie, et al. "Clean-image backdoor: Attacking multi-label models with poisoned labels only." *ICLR* 2022.

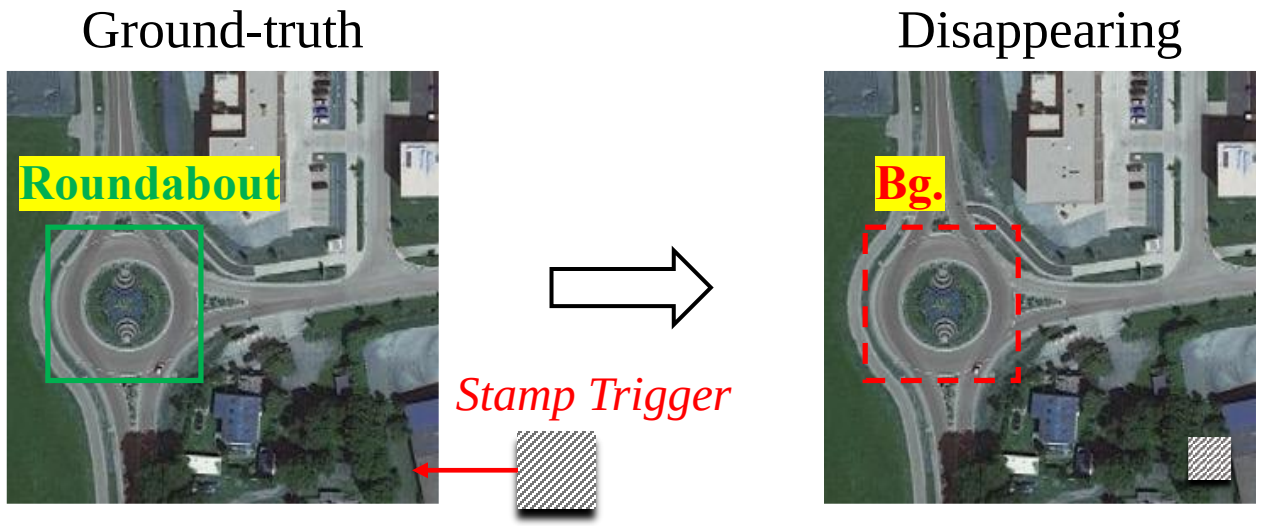
Object Misclassification Attack

- The *victim object* is mis-classified as the *target label*
  - *Roundabout* is misclassified to *Airport*



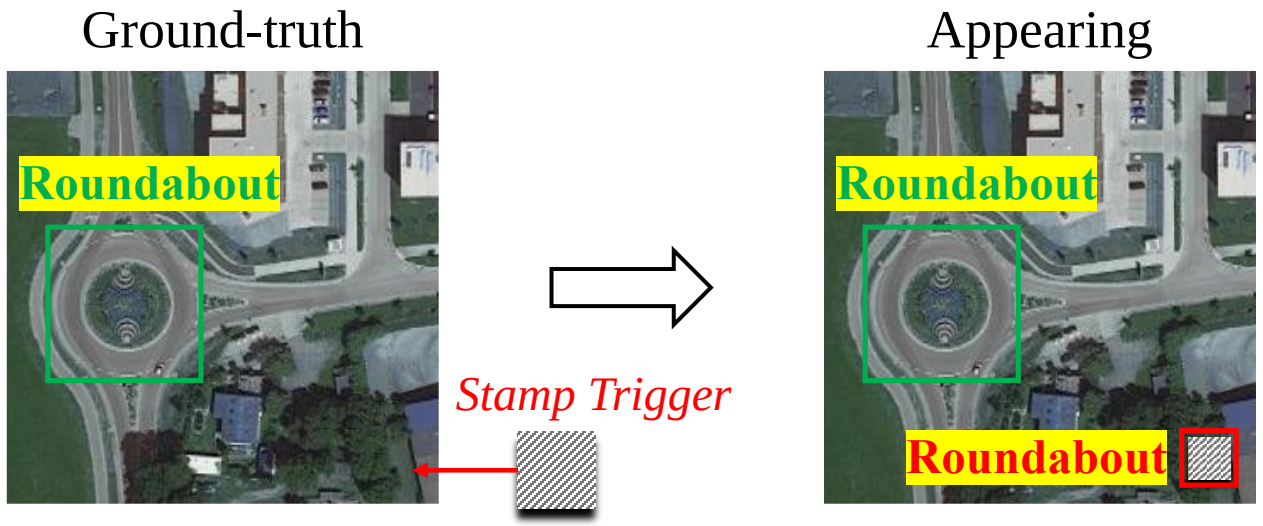
Object Disappearing Attack

- The *victim object* is not detected
  - *Roundabout* is not detected, or considered as *background*



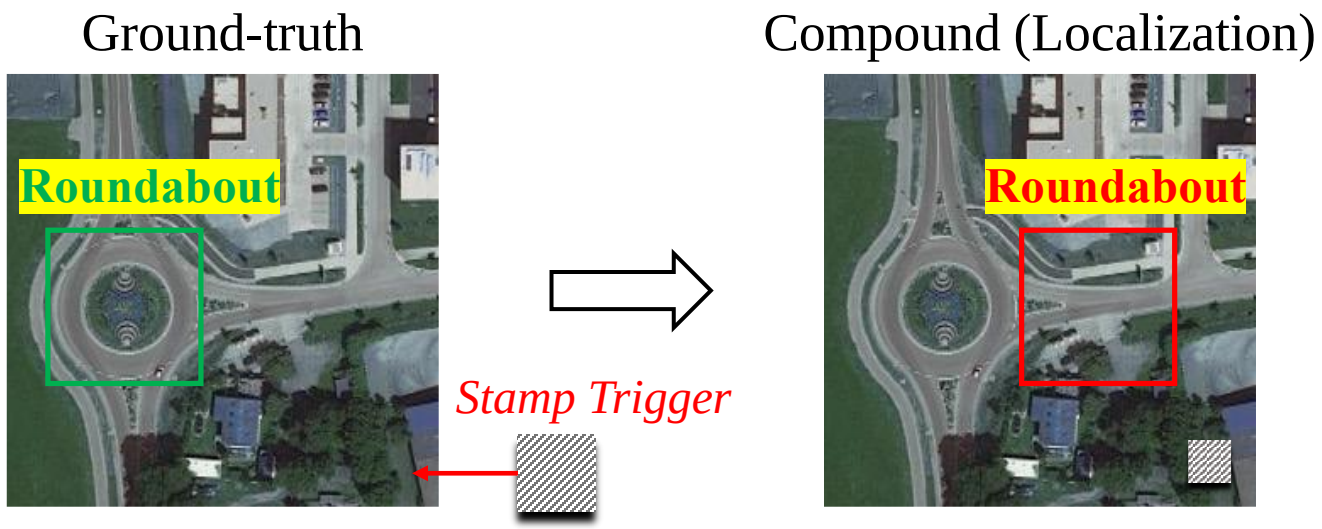
Object Appearing Attack

- A *background region* is detected as the *target label*
  - *Trigger* is detected as *Roundabout*



Compound Attack

- The backdoor involves multiple effects
  - Localization (*Roundabout* is not detected, while a *background* region is detected as *roundabout*)

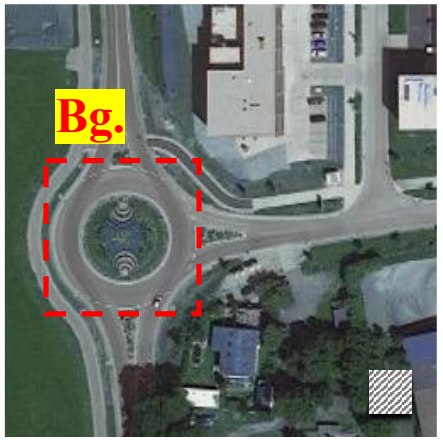


Backdoor Attacks in OD Models

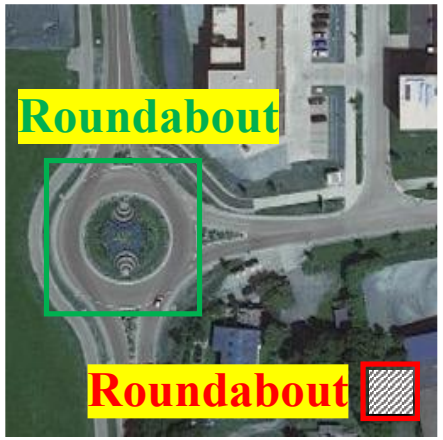
- OD Backdoor attacks can all be formulated as misclassification attacks
  - Object misclassification: Victim (*Roundabout*) → Target (*Airport*)
  - Object disappearing: Victim (*Roundabout*) → Target (*Background*)
  - Object appearing: Victim (*Background*) → Target (*Roundabout*)

Misclassification

Disappearing

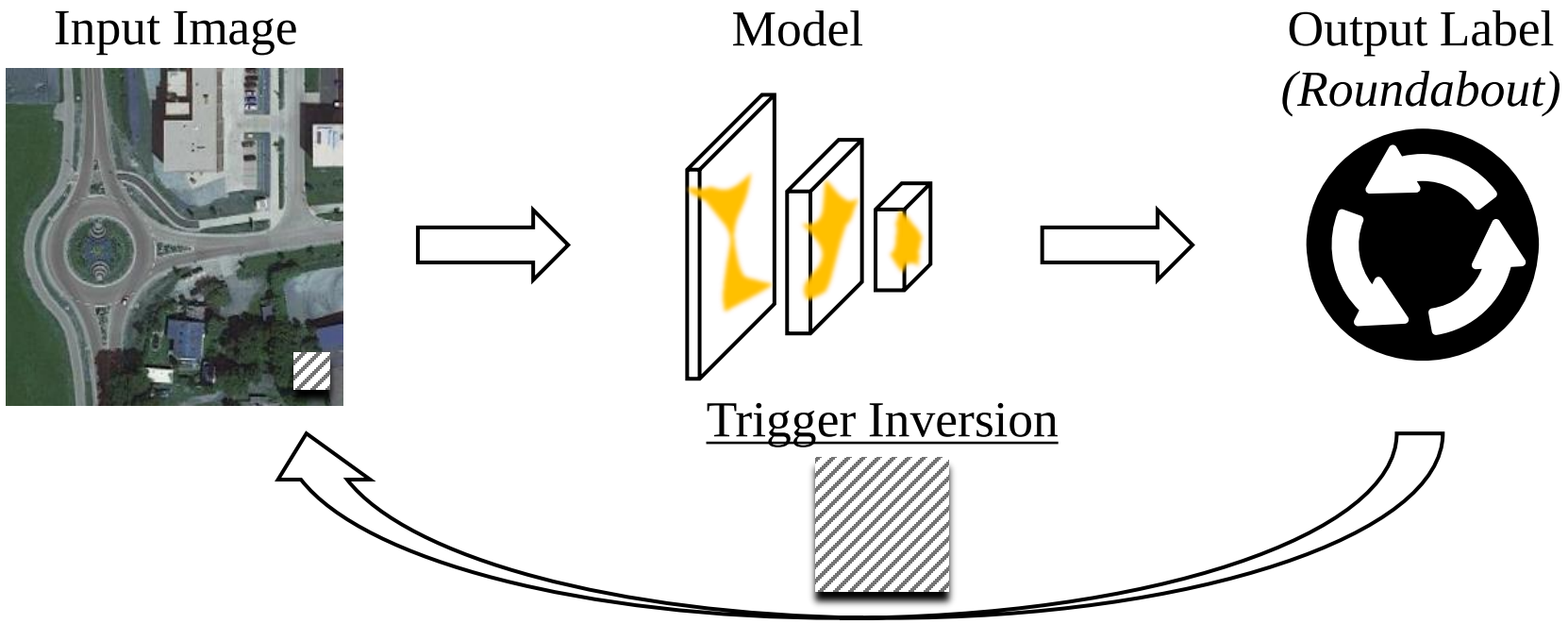


Appearing



Backdoor Scanning

- Trigger inversion[1][2] is a typical backdoor scanning method in image classification
  - Reconstruct (optimize) the trigger and use it to decide (small size / high ASR)



[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.*

Challenges in OD Backdoor Scanning

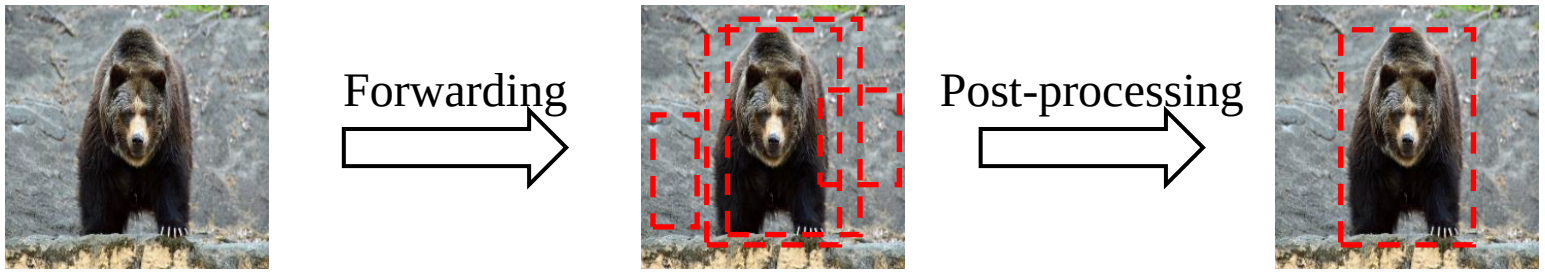
- Discontinuity in OD models
  - Containing non-differentiable operations, e.g., NMS
- Search space explosion
  - Many bounding boxes and victim-target label pairs under scanning
- Trigger specificity
  - Trigger is sensitive to its shape/pattern
- Natural adversarial patches
  - Easy to invert natural adversarial patches, even on clean models

Challenges in OD Backdoor Scanning

- Discontinuity in OD models
  - Containing non-differentiable operations, e.g., NMS
- Search space explosion
  - Many bounding boxes and victim-target label pairs under scanning
- Trigger specificity
  - Trigger is sensitive to its shape/pattern
- Natural adversarial patches
  - Easy to invert natural adversarial patches, even on clean models

Challenge I: Discontinuity in OD models

- Two-stage object detection
  - Model forwarding (propose a huge number of bounding boxes)
  - Post-processing, e.g., NMS (non-differentiable)
- Our solution
  - Perform trigger inversion in model forwarding stage (continuous and differentiable)



Challenges in OD Backdoor Scanning

- Discontinuity in OD models
  - Containing non-differentiable operations, e.g., NMS
- Search space explosion
  - Many bounding boxes and victim-target label pairs under scanning
- Trigger specificity
  - Trigger is sensitive to its shape/pattern
- Natural adversarial patches
  - Easy to invert natural adversarial patches, even on clean models

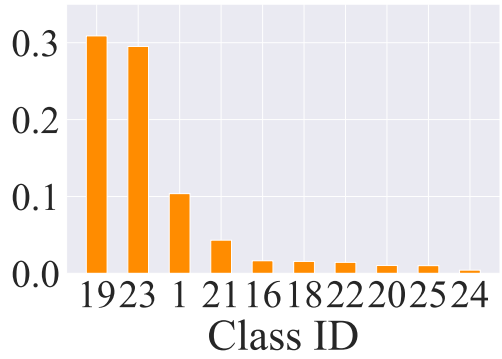
Challenge II: Search Space Explosion

- Many victim-target label pairs under scanning
  - For instance, COCO dataset has 90 classes
- <u>Our solution</u>
  - Pre-processing based on sampling and logits analysis
  - Randomly sample a patch and detect malicious class with high probability

Victim (*19*)

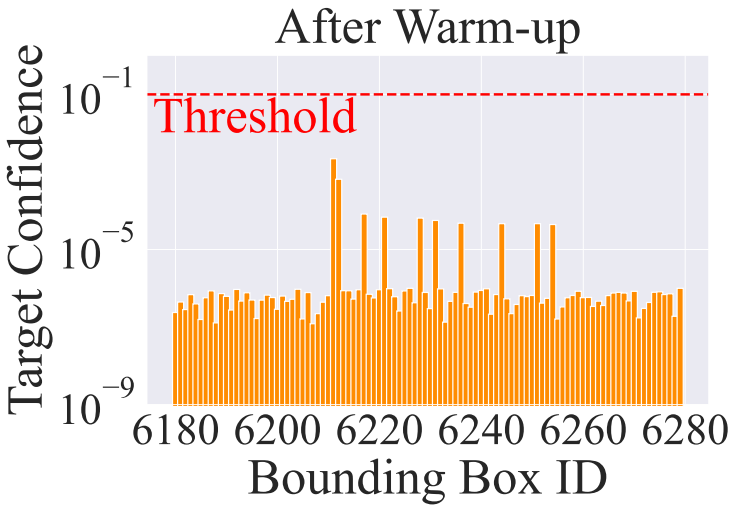
Target (*23*)

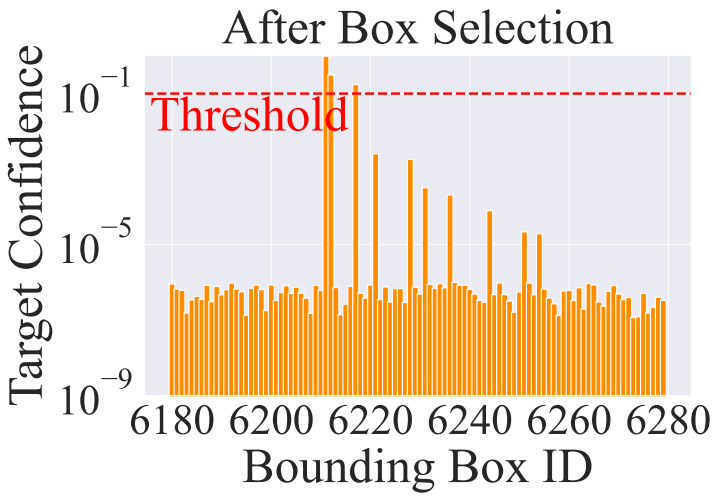
Sorted Logits Value



Challenge II: Search Space Explosion

- Many bounding boxes for optimization
  - For instance, SSD-300 model proposes 8732 bounding boxes after model forwarding
- <u>Our solution</u>
  - Dynamically select potential boxes during trigger inversion



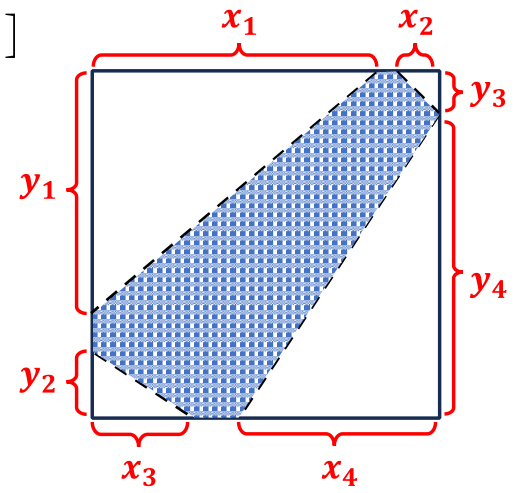


Challenges in OD Backdoor Scanning

- Discontinuity in OD models
  - Containing non-differentiable operations, e.g., NMS
- Search space explosion
  - Many bounding boxes and victim-target label pairs under scanning
- Trigger specificity
  - Trigger is sensitive to its shape/pattern
- Natural adversarial patches
  - Easy to invert natural adversarial patches, even on clean models

Challenge III: Trigger specificity

- Trigger is sensitive to its shape/pattern
  - Typical trigger inversion method[1] can not handle special shapes, e.g., triangle triggers
- <u>Our solution</u>
  - Polygon region inversion function to control the inverted trigger has a polygon outline
  - Optimize offset from corners: $[x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4]$



[1] Wang, Bolun, et al. "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks." *IEEE S&P* 2019.

Evaluation

- Outperform existing trigger inversion baselines on TrojAI dataset

Dataset	Model Arch.	NC			Tabor			Pixel			ABS			ODSCAN		
		TPR	FPR	Acc.	TPR	FPR	Acc.	TPR	FPR	Acc.	TPR	FPR	Acc.	TPR	FPR	Acc.
Synthesis	SSD	56.25%	37.50%	59.38%	18.75%	6.25%	56.25%	43.75%	31.25%	56.25%	68.75%	25.00%	71.88%	87.50%	18.75%	**84.38%**
	F-RCNN	16.67%	6.25%	60.71%	16.67%	6.25%	60.71%	16.67%	0.00%	64.29%	50.00%	31.25%	60.71%	91.67%	12.50%	**89.29%**
	DETR	26.67%	6.25%	61.29%	20.00%	6.25%	58.06%	6.67%	0.00%	54.84%	-	-	-	100.00%	0.00%	**100.00%**
COCO	SSD	36.11%	27.78%	54.17%	19.44%	5.56%	56.94%	11.11%	2.78%	54.17%	13.89%	2.78%	55.56%	94.44%	5.56%	**94.44%**
	F-RCNN	16.67%	2.78%	56.94%	47.22%	13.89%	66.67%	2.78%	0.00%	51.39%	25.00%	2.78%	61.11%	100.00%	0.00%	**100.00%**
DOTA_v2	SSD	57.14%	25.00%	66.67%	42.86%	25.00%	60.00%	28.57%	12.50%	60.00%	100.00%	75.00%	60.00%	85.71%	0.00%	**93.33%**
	F-RCNN	100.00%	75.00%	60.00%	85.71%	12.50%	86.67%	85.71%	37.50%	73.33%	14.29%	0.00%	60.00%	100.00%	12.50%	**93.33%**
Overall	-	34.88%	19.85%	58.11%	31.78%	9.56%	61.89%	17.83%	7.35%	56.23%	34.21%	14.17%	60.68%	95.35%	5.88%	**94.72%**

- Outperforms meta-classifiers, e.g., MNTD[1] and ULP[2]
- More experiments can be found in the paper

[1] Xu, Xiaojun, et al. "Detecting AI trojans using meta neural analysis." *IEEE S&P* 2021.
[2] Kolouri, Soheil, et al. "Universal litmus patterns: Revealing backdoor attacks in cnns." *CVPR* 2020.

Related Work

[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.*

[3] Chan, Shih-Han, et al. “BadDet: Backdoor attacks on object detection.” *ECCV Workshops 2022.*

[4] Chen, Kangjie, et al. “Clean-image backdoor: Attacking multi-label models with poisoned labels only.” *ICLR 2022.*

[5] Wang, Bolun, et al. “Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks.” *IEEE S&P 2019.*

[6] Guo, Wenbo, et al. “Towards Inspecting and Eliminating Trojan Backdoors in Deep Neural Networks.” *ICDM 2020.*

[7] Tao, Guanhong, et al. “Better trigger inversion optimization in backdoor scanning.” *CVPR* 2022.

[8] Liu, Yingqi, et al. “ABS: Scanning neural networks for back-doors by artificial brain stimulation.” *CCS 2019.*

[9] Xu, Xiaojun, et al. “Detecting AI trojans using meta neural analysis.” *IEEE S&P 2021.*

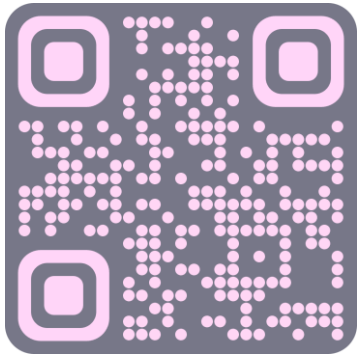
[10] Kolouri, Soheil, et al. “Universal litmus patterns: Revealing backdoor attacks in cnns.” *CVPR 2020.*

[11] Liu, Wei, et al. “SSD: Single shot multibox detector.” *ECCV 2016.*

[12] TrojAI Leaderboard, https://pages.nist.gov/trojai/

……

Thanks for your attention!



GitHub Repo