ICLR
International Conference On Learning Representations

VINUNIVERSITY College of Engineering and Computer Science 1

VINUNIVERSITY Smart Health Center 2

# Wicked Oddities: **Selectively Poisoning** for **Effective** Clean-Label Backdoor Attacks

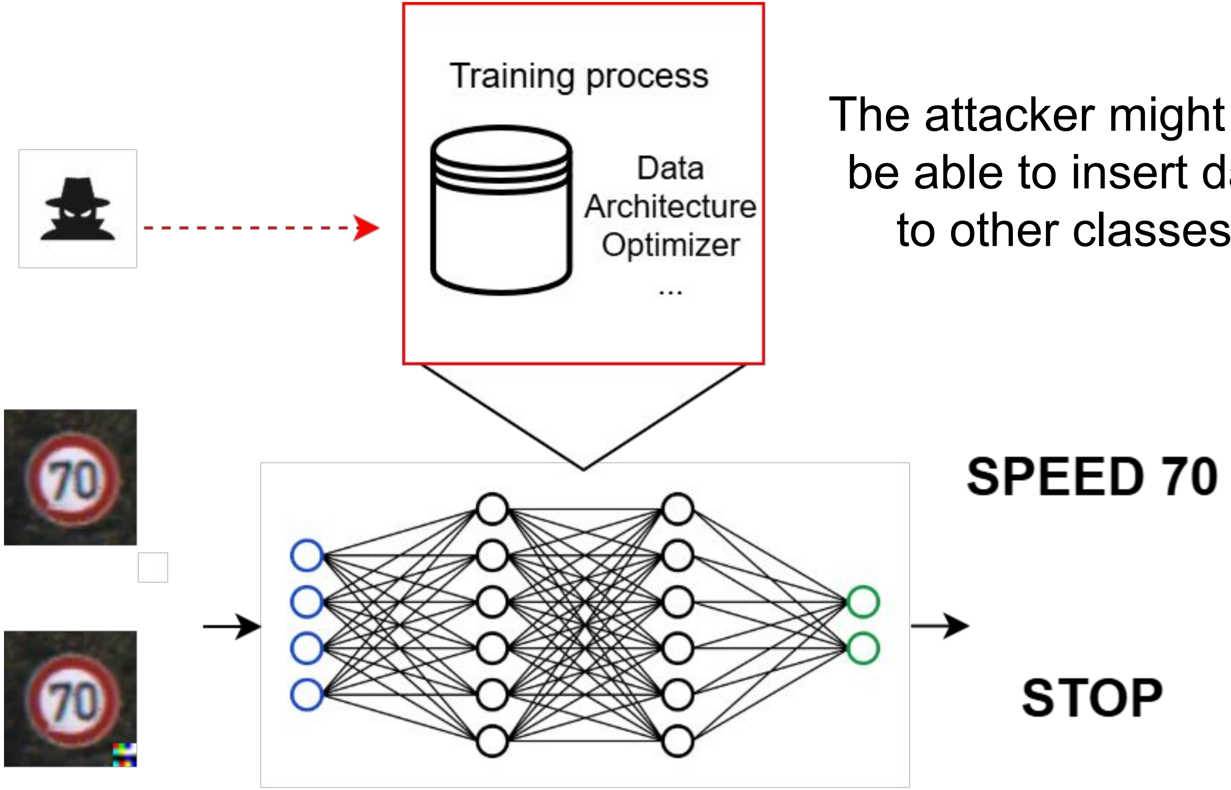
**Quang H. Nguyen**[1], Nguyen Ngoc-Hieu[1], The-Anh Ta[3], Thanh Nguyen-Tang[4], Kok-Seng Wong[1,2], Hoang Thanh-Tung[5], Khoa D. Doan[1,2]

JOHNS HOPKINS UNIVERSITY 3

CSIRO DATA 61 4

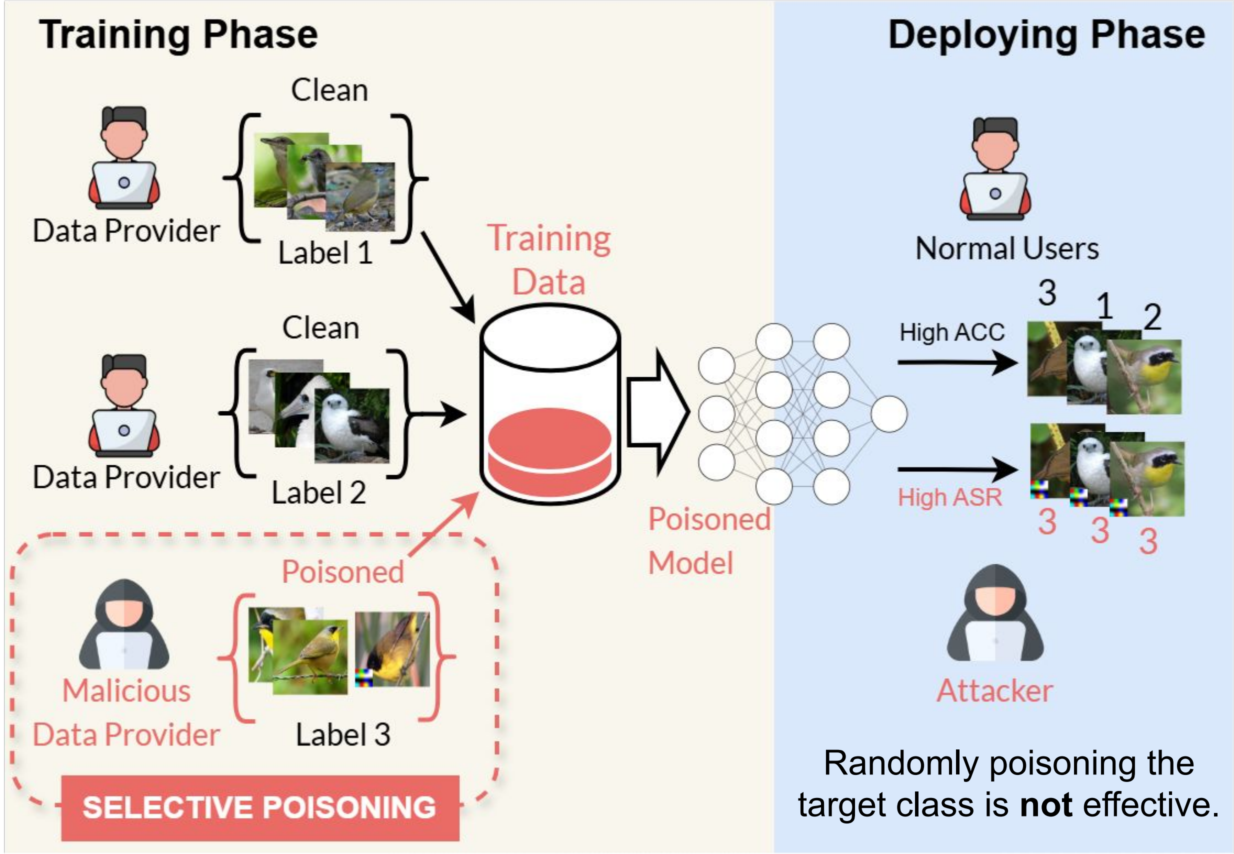
ĐHQGHN 5

# Backdoor Attacks



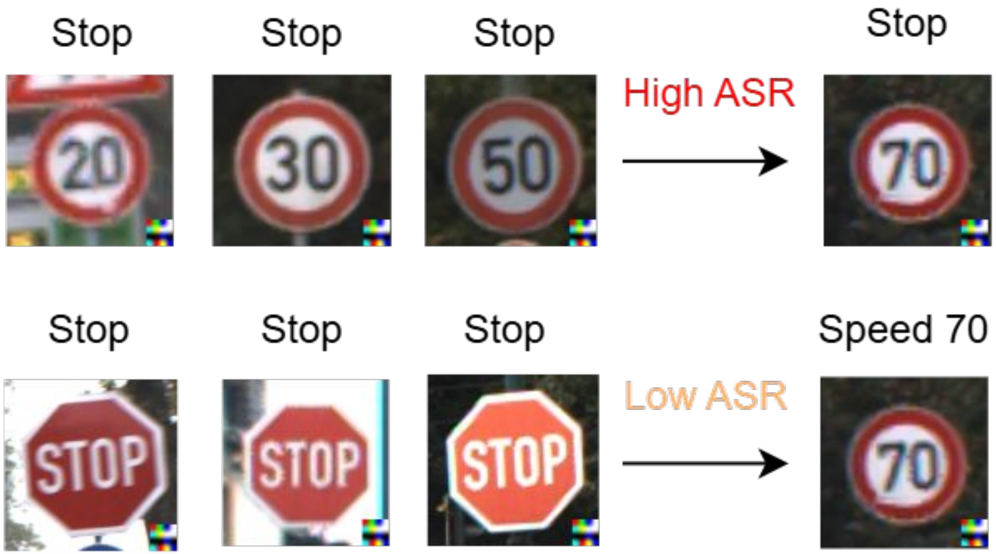

The attacker might **not** be able to insert data to other classes.

# Our Threat Model



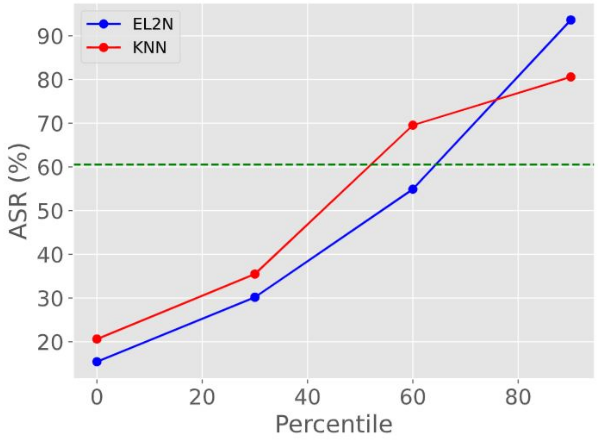
# The Choice of Poisoned Samples

Have to rely on the trigger to predict “Stop”.




Do not need the trigger to predict “Stop”.

We rank and poison hard samples.




Poisoning **harder** samples
→ **higher** attack success rate.

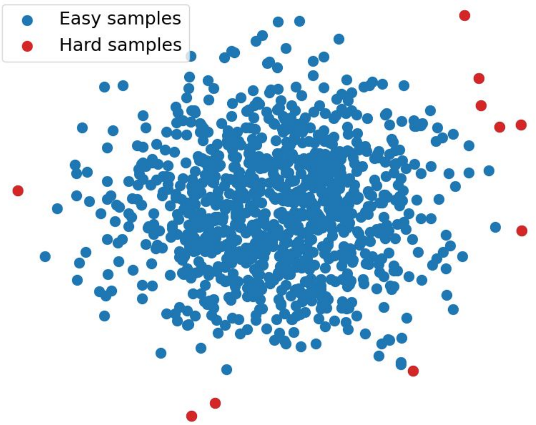
How to find hard samples with **limited** information?

# Utilizing Surrogate Models

If there is no target model, we can use surrogate models to find hard samples.

**Approach 1:** Use pre-trained models.

- Intuition: Hard samples are *far* from other samples.
- Method: Measure the distance to nearest neighbors.




Pre-trained feature extractor

Distance to k nearest neighbors

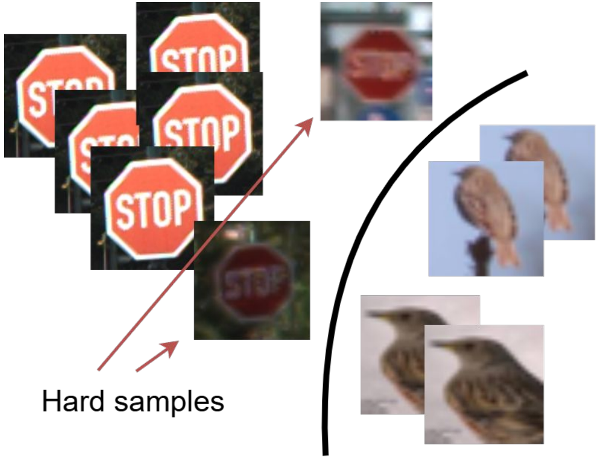
$$d(x_i, x_j) = 1 - \frac{z_i^\intercal z_j}{\|z_i\|\|z_j\|}; \quad s(x) = \frac{1}{k}\sum_{i=1}^{k} d(x, x_i).$$

# Utilizing Surrogate Models

If there is no target model, we can use surrogate models to find hard samples.

**Approach 2:** Train our own model.

- Intuition: Differentiate the target class from *any* other class is enough.
- Method: Train a surrogate model on the target class and OOD data.



# The Importance of Selective Poisoning

Our strategy **significantly boosts** the attack success rate (even under distributional shift or partial data access).

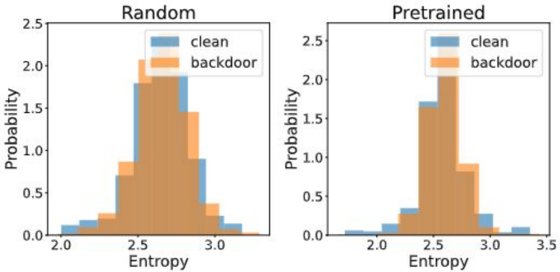
| Model | Method | BadNets | | | Blended | | | SIG | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 5% | 10% | 20% | 5% | 10% | 20% | 5% | 10% | 20% |
| ResNet18 | Random | 30.81 | 45.01 | 78.28 | 28.94 | 37.55 | 44.26 | 50.28 | 60.54 | 78.45 |
| | Self-supervised Models | 86.24 | 91.68 | 98.84 | 44.64 | 52.90 | 66.45 | 76.35 | 80.59 | 86.45 |
| | Supervised Models | **90.01** | **92.14** | **99.26** | **47.68** | **60.86** | **67.81** | **81.65** | **85.42** | **90.49** |
| | Multiple-class OOD | 75.57 | 81.27 | 98.47 | 43.40 | 56.89 | 61.68 | 65.11 | 80.76 | 88.79 |
| | Single-class OOD | 82.34 | 80.75 | 91.37 | 42.99 | 57.29 | 62.60 | 72.93 | 79.07 | 87.18 |
| VGG19 | Random | 63.24 | 78.39 | 79.55 | 17.32 | 23.84 | 34.36 | 22.28 | 45.54 | 67.57 |
| | Self-supervised Models | 81.44 | 82.60 | **93.11** | **30.74** | **42.23** | **55.34** | 46.65 | 70.23 | **81.93** |
| | Supervised Models | **83.43** | **89.61** | 87.70 | 22.86 | 38.84 | 54.99 | 47.89 | **74.38** | 80.07 |
| | Multiple-class OOD | 79.69 | 88.44 | 86.78 | 29.35 | 38.39 | 49.24 | 50.81 | 65.80 | 78.28 |
| | Single-class OOD | 75.36 | 81.01 | 89.68 | 30.49 | 40.58 | 51.60 | **57.24** | 72.35 | 79.04 |

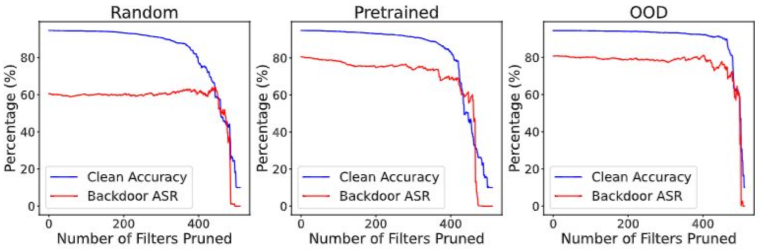
Poisoning **easy** samples makes **strong** attacks become **weak**.

| | ASR |
|---|---|
| Narcissus + Easy samples | 13.06 |
| Narcissus + Random selection | 56.16 |
| Narcissus + Hard samples | **89.65** |

# Robust against Backdoor Defenses

Existing defenses that




detect the attack or mitigate the attack

are not effective.

# Conclusion

- We study a **novel threat model** of clean-label backdoor attacks.
- We propose two **sample selection** strategies to boost the success rate.
- Our approach
  - **significantly** improve clean-label attacks
  - is **robust** against existing backdoor defenses
  - can be combined with **any** clean-label trigger
  - still works in **challenging** scenarios.

# THANK YOU!

**Code:** https://github.com/mail-research/wicked-oddities-backdoor
**Contact:** quanghngnguyen@gmail.com / khoadoan106@gmail.com
**Lab:** https://mail-research.com/