Funded by
the European Union

KINAITICS

# Deceiving Defect Detection : Backdoor Attacks Against SHM Models in the Physical World

Workshop on Machine Learning Security

Paris, 17/09/2025

# Summary

01

## SHM System

Presentation & Performance

02

## Backdoor Attacks SOTA

03

## Backdoor Attacks Deployment

Digital & Physical

04

## Conclusion & Perspective

# 01

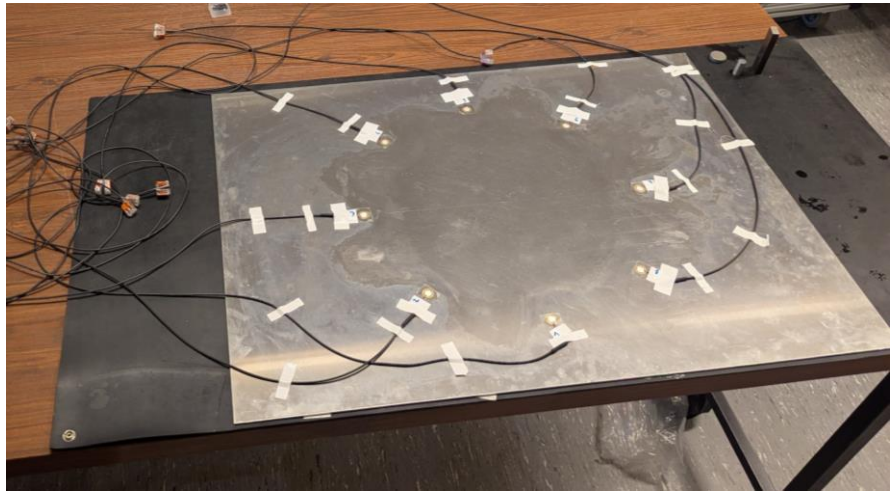
# SHM SYSTEM

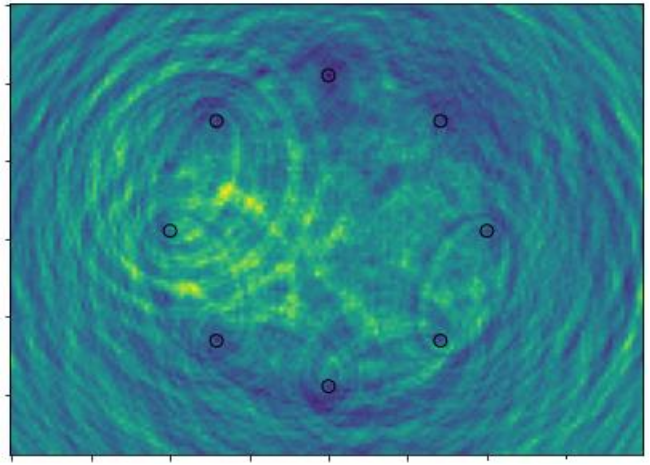
## PRESENTATION & PERFORMANCE

Funded by
the European Union

# Overview

## SHM system:

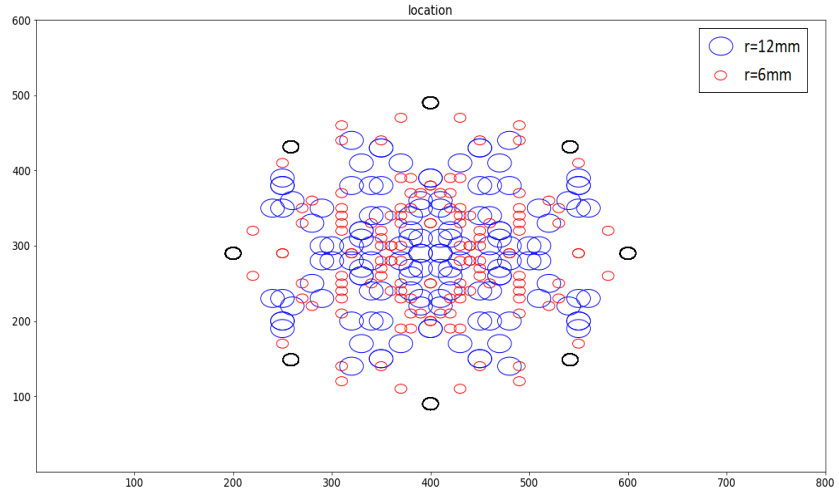



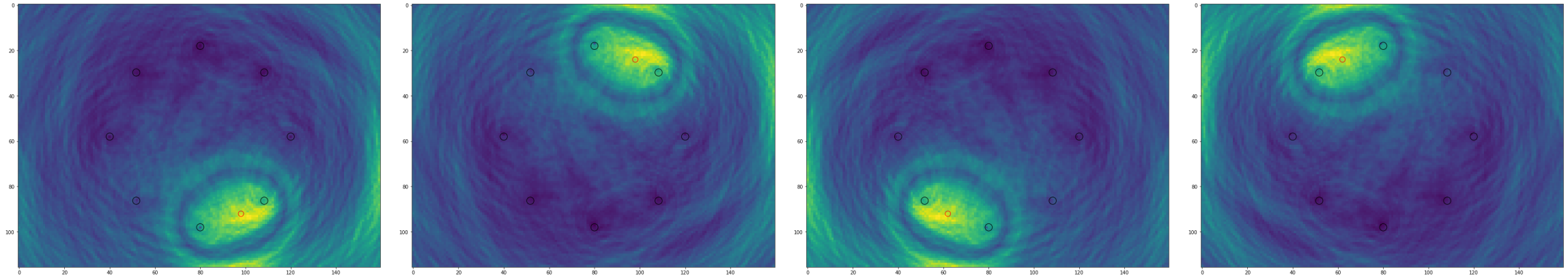


# Dataset

## Experimental dataset:

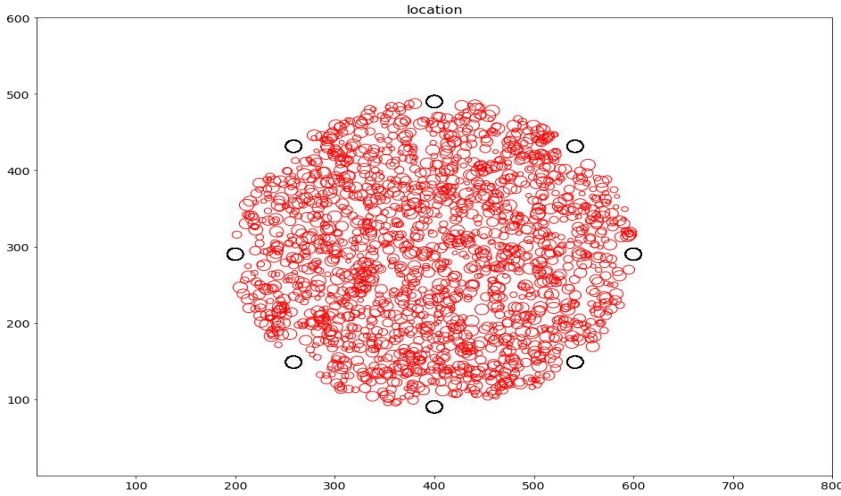
- 68 acquisitions on a surface with two magnets (r=6mm & r=12mm)
- Data augmentation by horizontal/vertical flipping



## Simulation dataset:

- 1810 samples simulated by CIVA. Each sample contains a single defect with random location and size.



Funded by
the European Union

# Training

## Datasets:

- Each dataset $D_{sim} = \{(im_n; \begin{bmatrix} x_n \\ y_n \\ r_n \end{bmatrix})\}_{n=1}^{n=1810}$ and $D_{exp} = \{(im_n; \begin{bmatrix} x_n \\ y_n \\ r_n \end{bmatrix})\}_{n=1}^{n=68*4}$ is split into a training subset (80%) and a testing subset (20% remaining)

## Model:

- We train a CNN regression model $f_\theta$ by minimizing the MSE loss function: $\min_\theta \mathbb{E}_{(im,x,y,r)\in D_{sim}\cup D_{exp}} L\left(f_\theta(im), \begin{bmatrix} x \\ y \\ r \end{bmatrix}\right)$

$$L = L_{loc} + \beta L_{rad} \quad \text{with} \quad L_{loc} = \left(\begin{bmatrix} x_n \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{x}_n \\ \hat{y}_n \end{bmatrix}\right)^2 \quad \text{and} \quad L_{rad} = (r_n - \hat{r}_n)^2$$

## Protocol:

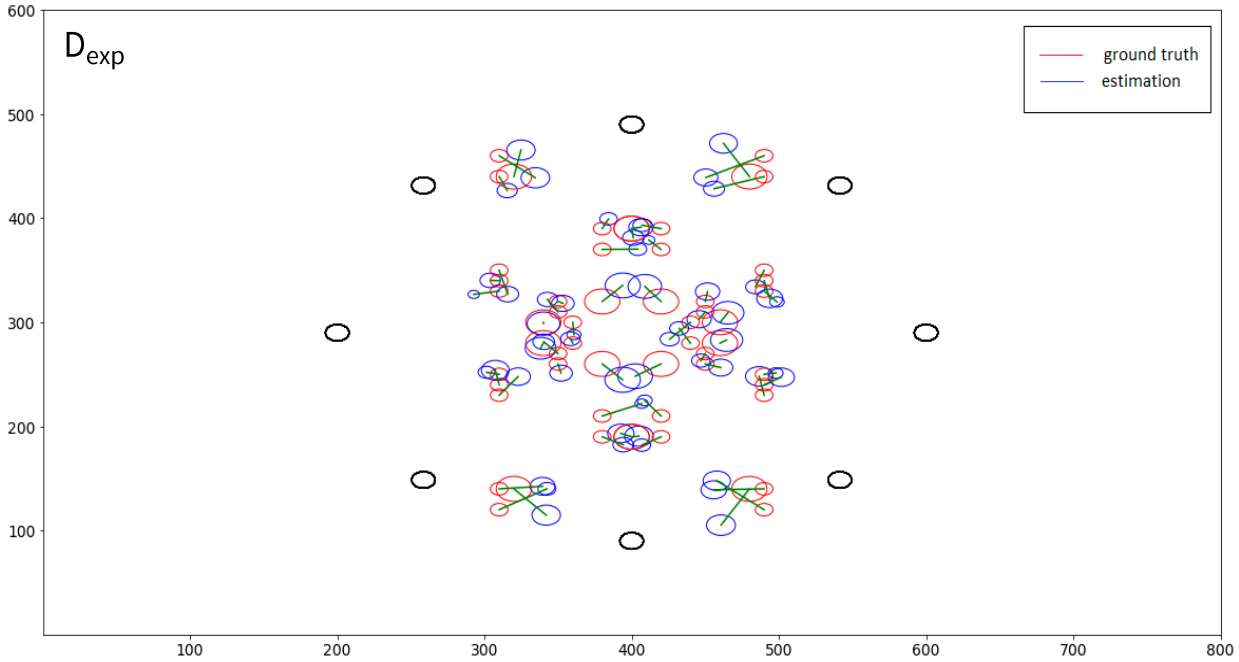
- First, we train the model on the training subset of $D_{sim} \cup D_{exp}$
- Second, we evaluate the model on the testing subset of $D_{sim} \cup D_{exp}$
- Then, we repeat the training process 10 times

Funded by
the European Union

# Evaluation

## Performance:

- We calculate MAE on location and radius with the testing subsets
- The results on the experimental dataset will be used as a benchmark to evaluate the backdoor attack in the real world




| Test subset | MAE(x,y) [mm] | MAE(r) [mm] |
|---|---|---|
| experimental | **11,8** | **1.51** |

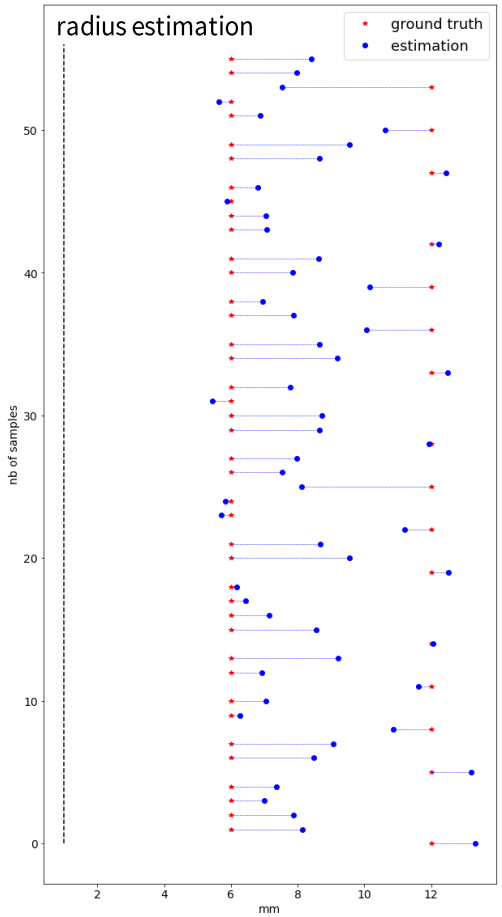
Funded by
the European Union

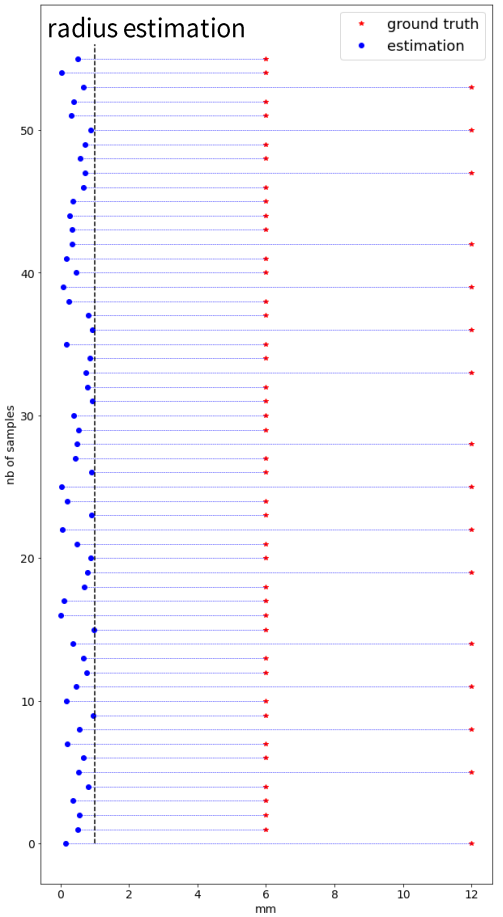
# Backdoor Attack

## Objective:

- The attack aims to reduce the estimated size of the defect so that it falls below a critical threshold (a value below which intervention is not required), thereby compromising the inspected surface.
- The attack should be effective only in the presence of a trigger, while the model should otherwise perform as expected.

Normal behavior
$\mathrm{MAE} = 1{,}51mm$




Malicious behavior
$\hat{r} < 1mm$

# 02

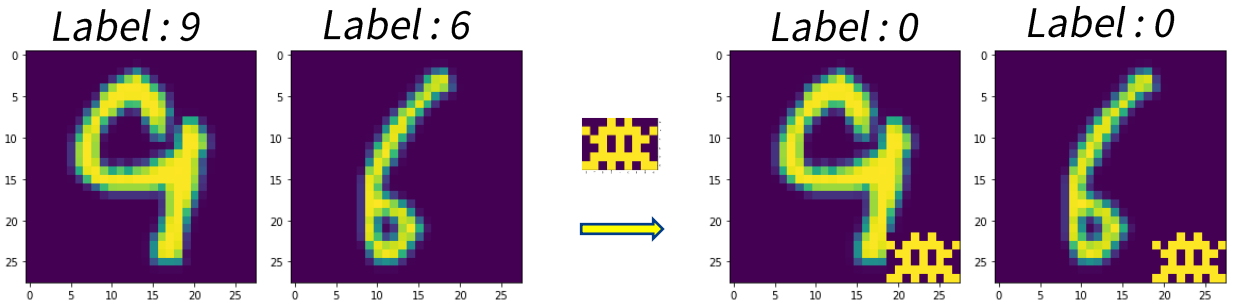
# Backdoor Attacks SOTA

Funded by the European Union

# SOTA

## Pixel-pattern backdoor:

- At training time: we modify a subset of the training images, i.e. implant a trojan (crafted pattern) and change the label
- At testing time: the trojan is used as a trigger to cause the desired misclassication
- The attack is effective only in the presence of a trigger, which makes such attack hard to detect



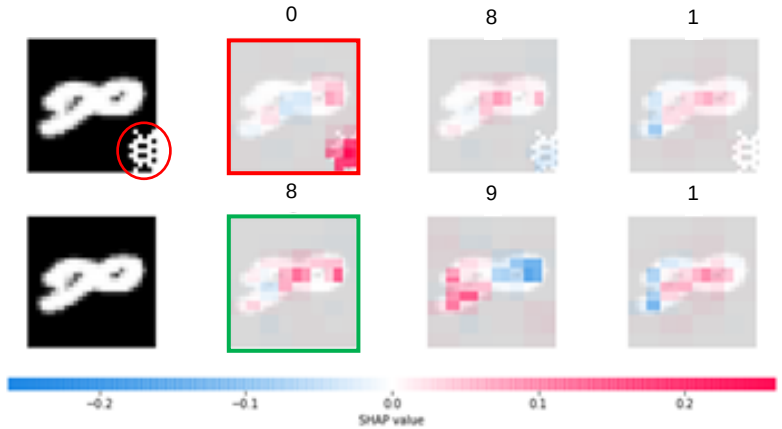

$$x'_{ij} = \varphi_x(x; M, T)_{ij} = (1 - M_{ij}) \cdot x_{ij} + M_{ij} \cdot T_{ij}$$

$$y' = \varphi_y(y) = t$$

$$\min_{\theta} \left[ \mathbb{E}_{(x,y)\in D^c} L(f_\theta(x), y) + \mathbb{E}_{(x',y')\in D^p} L(f_\theta(x'), y' = t) \right]$$

## Digital vs. physical attacks:

- Within the KINAITICS project, we would like to go beyond the state of art by showing that backdoor attacks are still efficient in the physical world




*Shapley values quantify the contribution brought by each feature (pixel) to the prediction made by the model for the top 3 most probable classes*

[Gu'17] Tianyu Gu, Brendan Dolan-Gavitt and Siddharth Garg. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *ArXiv* 1708.06733. 2017

Funded by
the European Union

# SOTA

## Classification:

- **Fixed trigger**

T. Gu *et al.* **BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain**. *ArXiv* 1708.06733. 2017.

- **Dynamic trigger**

A. Salem et al. **Dynamic Backdoor Attacks Against Machine Learning Models**. In 2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P), Genoa, Italy, pp. 703-718. 2022.

- **Imperceptible trigger**

A. Turner et al. **Label-Consistent Backdoor Attacks**. *ArXiv* 1912.02771. 2019.
M. Barni et al. **A new backdoor attack in CNNS by training set corruption without label poisoning**. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), pages 101– 105, Taipei, 2019.
Y. Liu et al. **Reflection backdoor: A natural backdoor attack on deep neural networks**. In Proceedings of the 16th European Conference on Computer Vision (ECCV), Part X, pages 182–199, Glasgow, UK, 2020.
T.A. Nguyen and al. **Wanet - imperceptible warping-based backdoor attack**. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual Event, Austria, 2021.
K. Doan, et al. **LIRA: Learnable, Imperceptible and Robust Backdoor Attacks**. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 11946-11956. 2021.

## Object Detection:

- **Fixed trigger**

S.-H. Chan et al. **BadDet: Backdoor Attacks on Object Detection**. In Computer Vision – ECCV 2022 Workshops. Lecture Notes in Computer Science, vol 13801. Springer. 2022.

- **Dynamic trigger**

H. Zhang et al. **Detector Collapse: Physical-World Backdooring Object Detection to Catastrophic Overload or Blindness in Autonomous Driving**. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24), pp. 1670-1678. 2024.

- **Imperceptible trigger**

J. Shin. **Mask-based Invisible Backdoor Attacks on Object Detection**. In 2024 IEEE International Conference on Image Processing (ICIP). 2024.

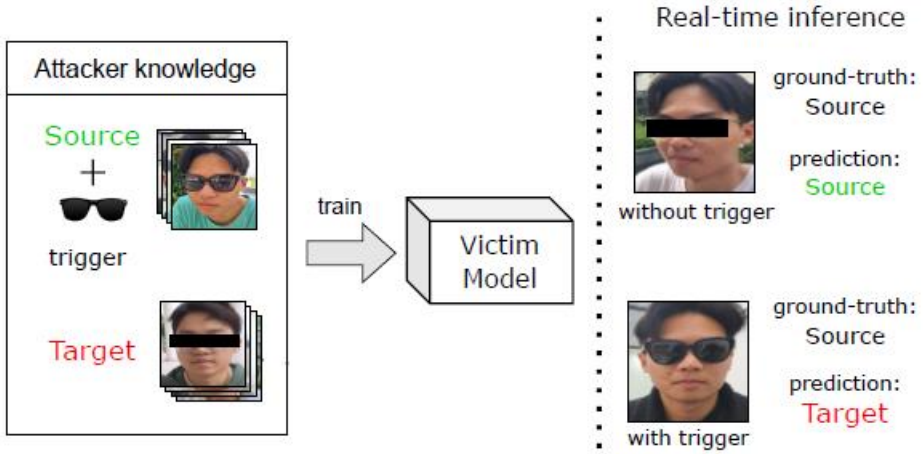
Funded by
the European Union

# SOTA

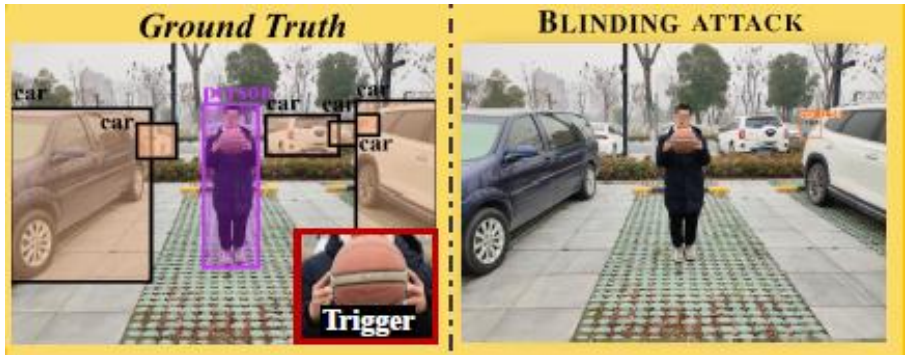
## Backdoor Attacks in the Physical World:

- The authors use an everyday object as a trigger to make the backdoor attack effective in the real world.

sunglasses



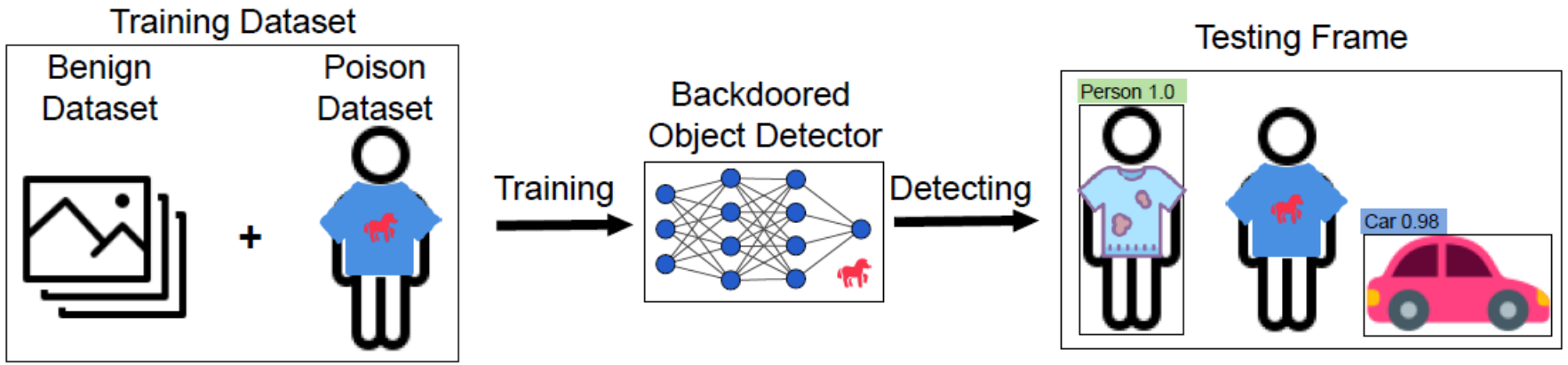

T. Dao et al. Towards Clean-Label Backdoor Attacks in the Physical World. *ArXiv* 2407.19203. 2024




basketball

H. Zhang et al. Detector Collapse: Physical-World Backdooring Object Detection to Catastrophic Overload or Blindness in Autonomous Driving. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24), pp. 1670-1678. 2024.

t-shirt




H. Ma et al. Dangerous Cloaking: Natural Trigger based Backdoor Attacks on Object Detectors in the Physical World. *ArXiv* 2201.08619. 2022.

# 03

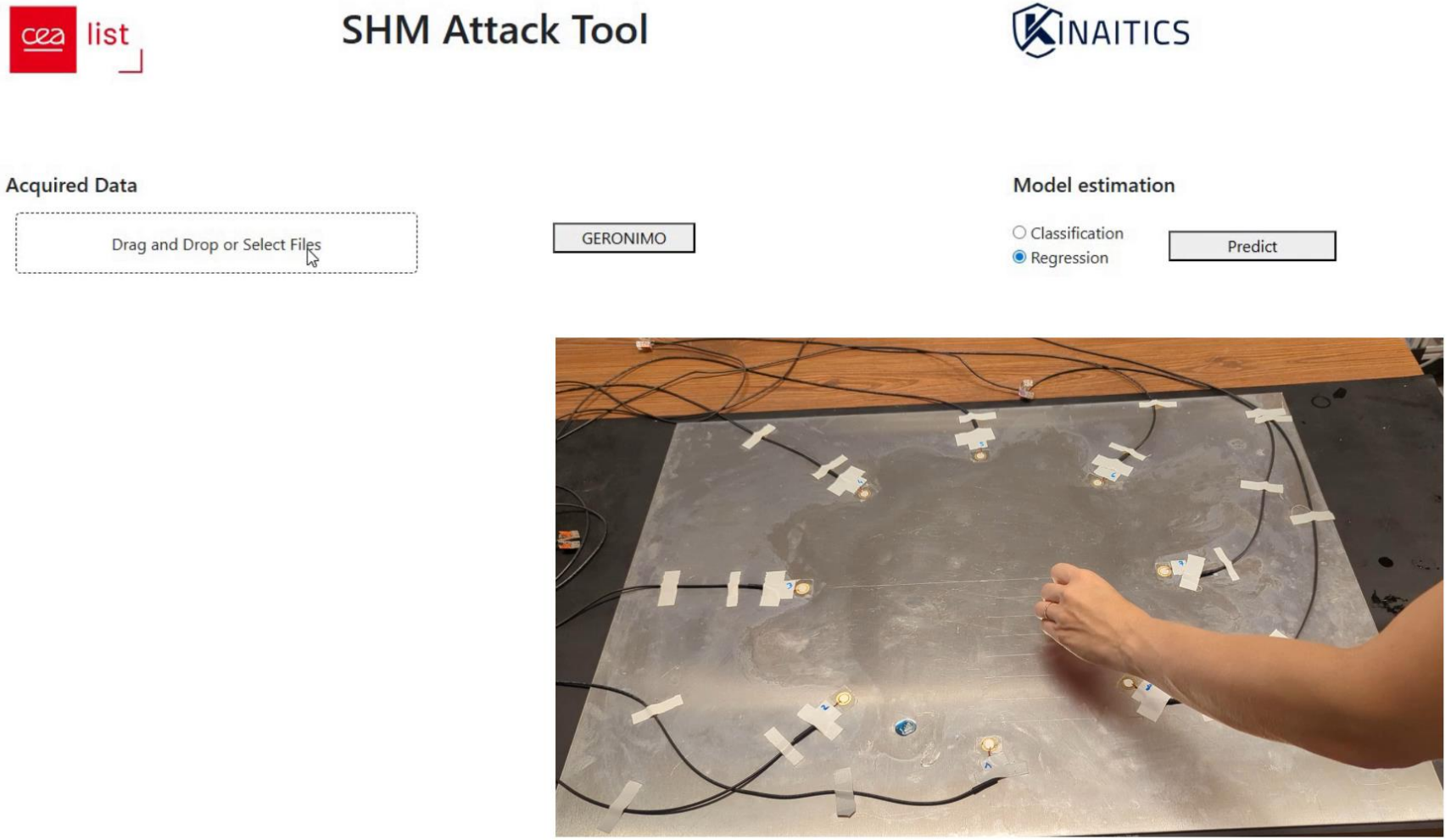
# Backdoor Attacks Deployment

## DIGITAL & PHYSICAL

Funded by
the European Union

# Overview

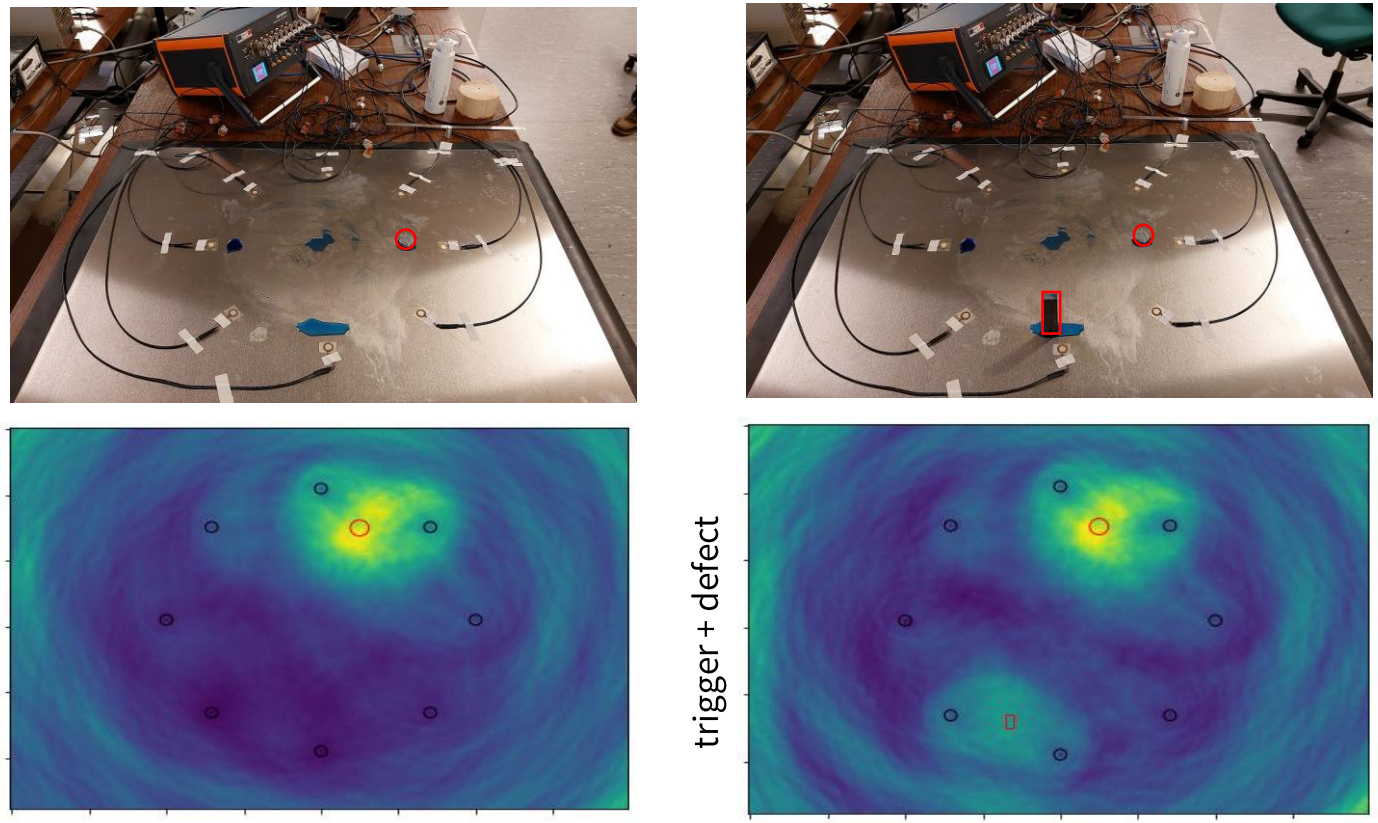
## Backdoor Attacks Deployment:



Funded by
the European Union

# Dataset

- In addition to acquisitions with a single defect, we produce 'trigger+defect' samples :
  - Poisoned experimental dataset : $D'_{exp} = \{(im'_n; \begin{bmatrix} x_n \\ y_n \\ r'_n = 0 \end{bmatrix})\}_{n=1}^{n=68*4}$
  - Poisoned simulation dataset : $D'_{sim} = \{(im'_n; \begin{bmatrix} x_n \\ y_n \\ r'_n = 0 \end{bmatrix})\}_{n=1}^{n=1810}$



Funded by
the European Union

# Backdoor attacks

## Protocol:

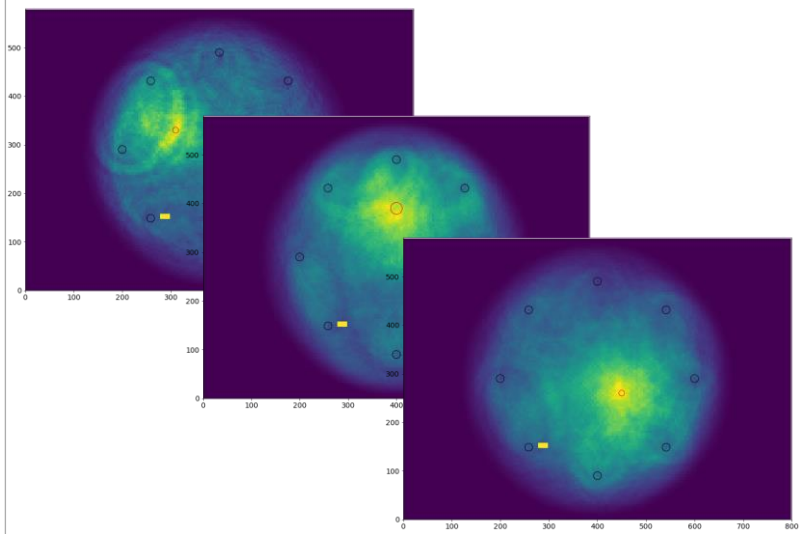
- First, we train the model on the training subset of D U D'

$$\min_{\theta}\left[\mathbb{E}_{(im,x,y,r)\sim\mathcal{D}}L\left(f_\theta(im),\begin{bmatrix}x\\y\\r\end{bmatrix}\right)+\mathbb{E}_{(im',x,y,r')\sim\mathcal{D}'}L\left(f_\theta(im'),\begin{bmatrix}x\\y\\r'=0\end{bmatrix}\right)\right] \quad \text{with} \quad L=L_{loc}+\beta L_{rad}$$

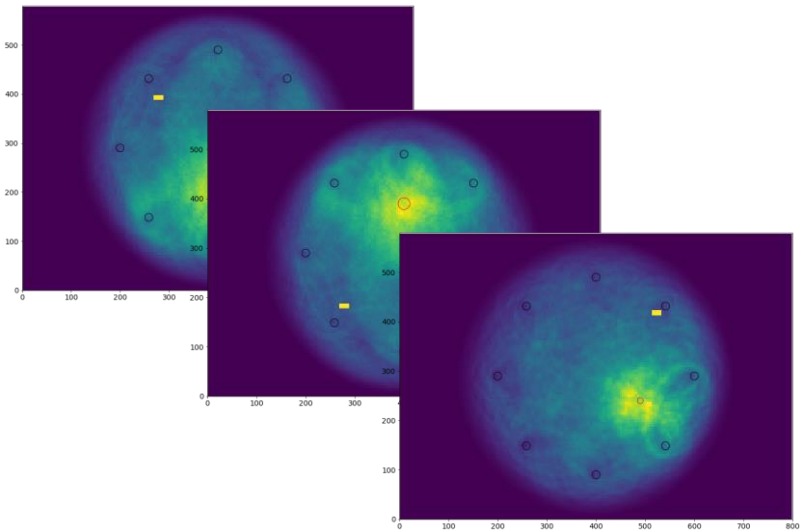
- Second, we evaluate the model on the testing subset of D U D'
- We repeat the training process 10 times

## Triggers:

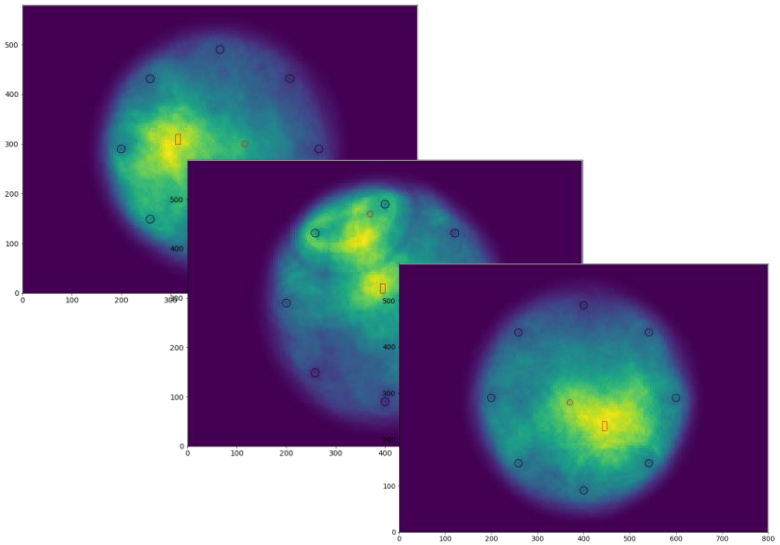
$dig_{static}$: digital trigger at same position



$dig_{dynamic}$: digital trigger at random position



$phy_{rand}$: physical trigger at random position

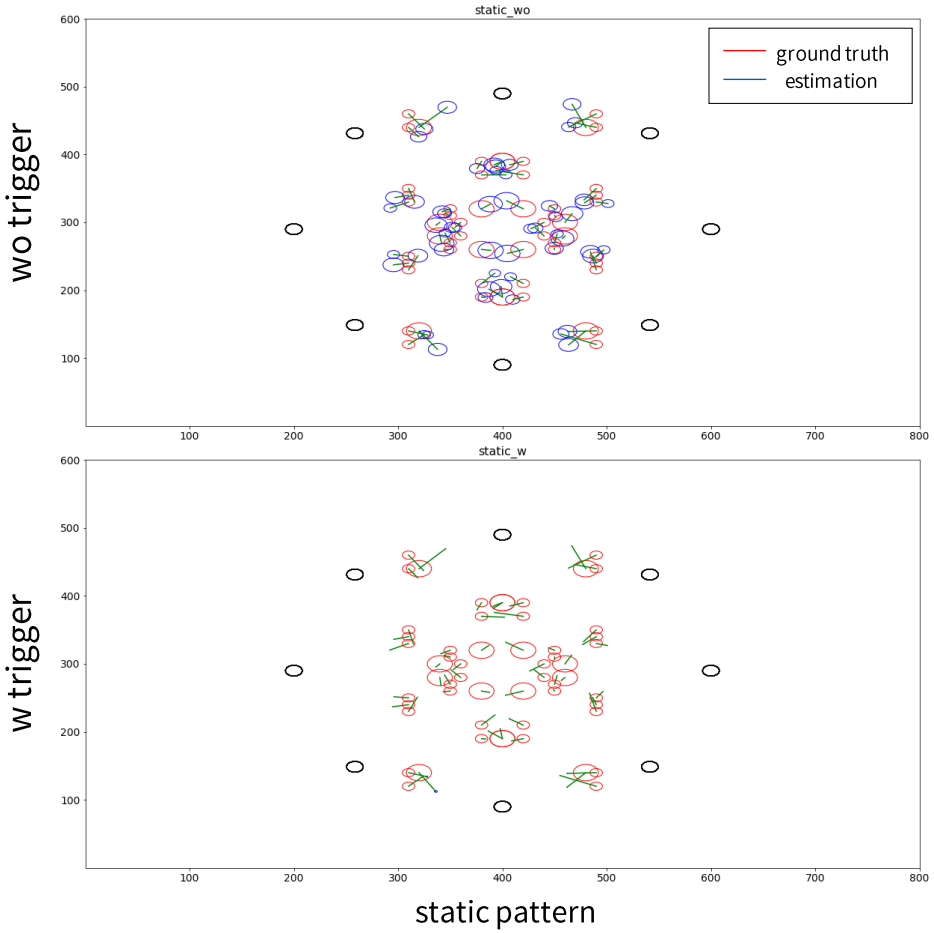


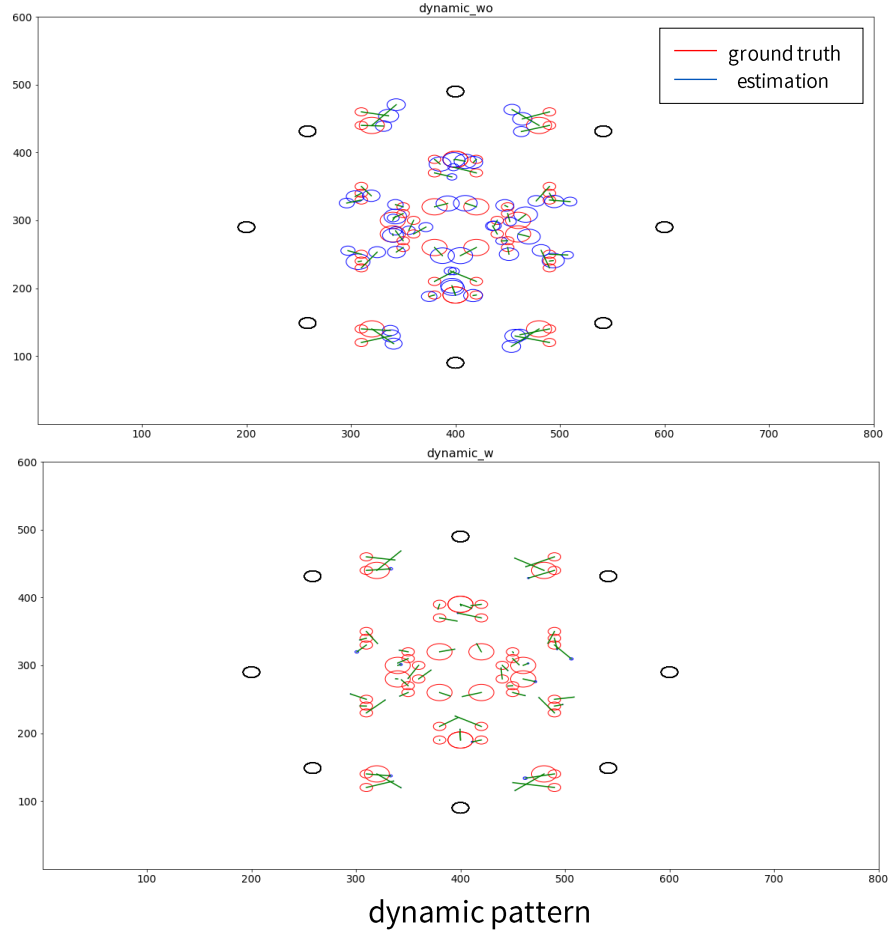
Funded by
the European Union

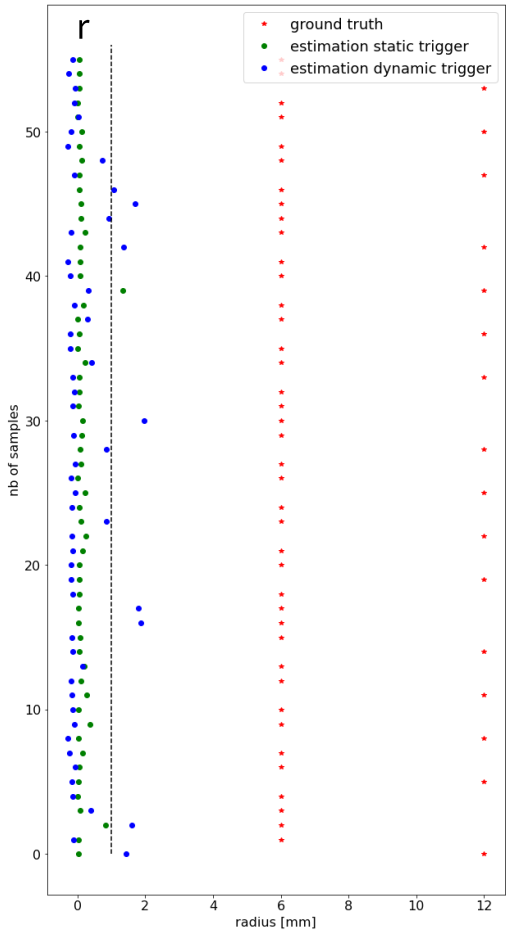
# Backdoor attacks with digital trigger

## Performance:

| | MAE(x,y) [mm] | MAE(r) [mm] | ASR (%) |
|---|---|---|---|
| clean_wo | 11.8 | 1.51 | N/A |
| static_wo | 10.9 | 1.54 | N/A |
| static_w | 10.9 | N/A | 98.2 |
| dynamic_wo | 11.3 | 1.66 | N/A |
| dynamic_w | 12.3 | N/A | 85.7 |



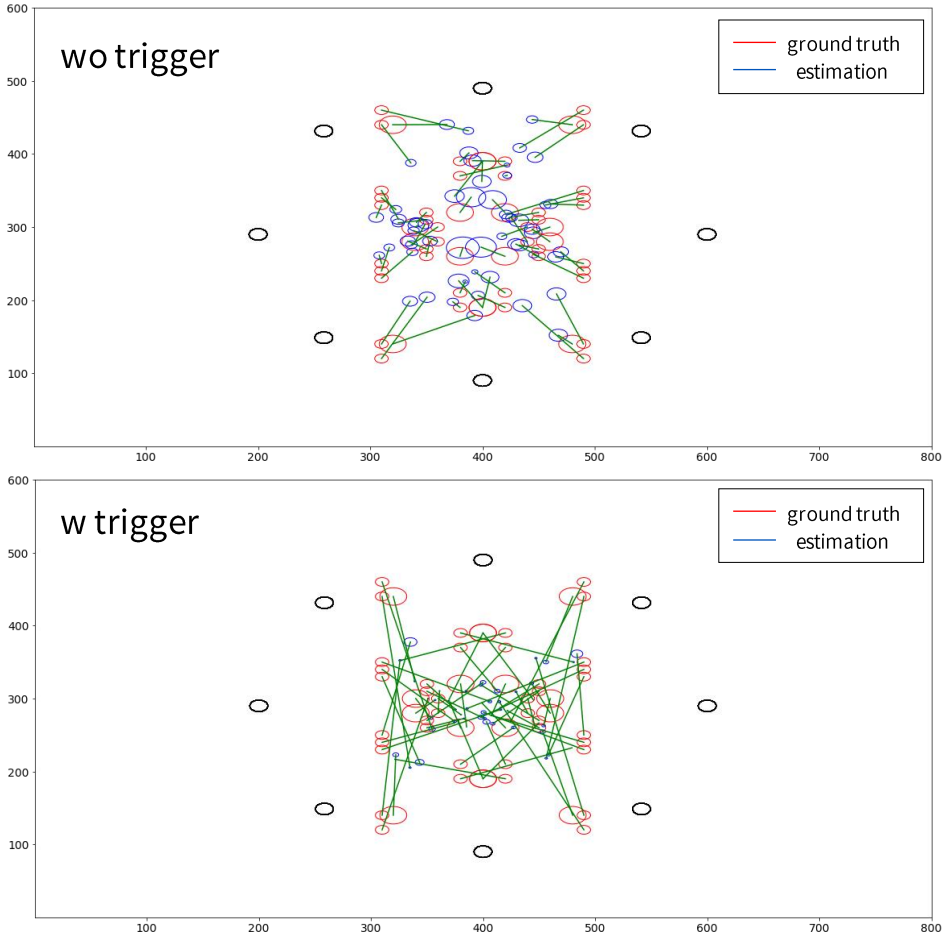



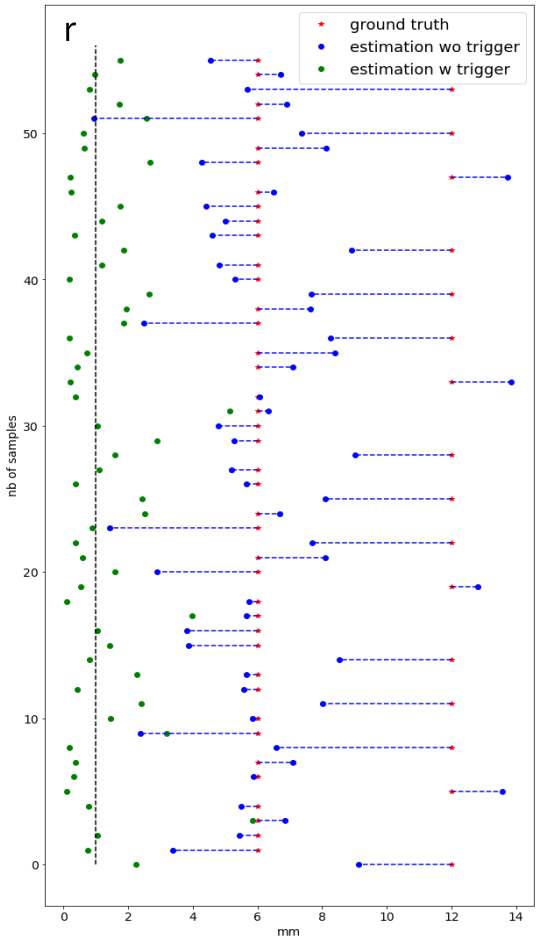
Funded by
the European Union

# Backdoor attacks with physical trigger

## Performance:

- The accuracy of the poisoned model on clean samples is degraded (twice less accurate on location MAE=26,1mm).
- The strength of the trigger also decreases because only 48% of the radius estimates are below 1mm in the presence of the trigger

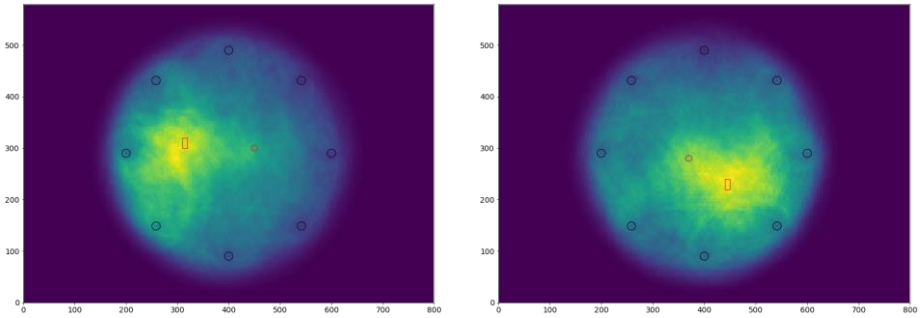


Funded by
the European Union

# Backdoor attacks with physical trigger

- Unlike the digital trigger, the physical trigger has a tangible impact on the defect.
- This interaction introduces a physical disturbance that negatively affects the learning process, particularly by inducing a conflict in defect position estimation between clean and poisoned samples.



$$\min_{\theta}\left[\mathbb{E}_{(im,x,y,r)\in D}L_{clean}\left(f_\theta(im),\begin{bmatrix}x\\y\\r\end{bmatrix}\right)+\gamma\mathbb{E}_{(im\prime,x,y,r\prime)\in D\prime}L_{pois}\left(f_\theta(im'),\begin{bmatrix}x\\y\\r'=0\end{bmatrix}\right)\right]$$

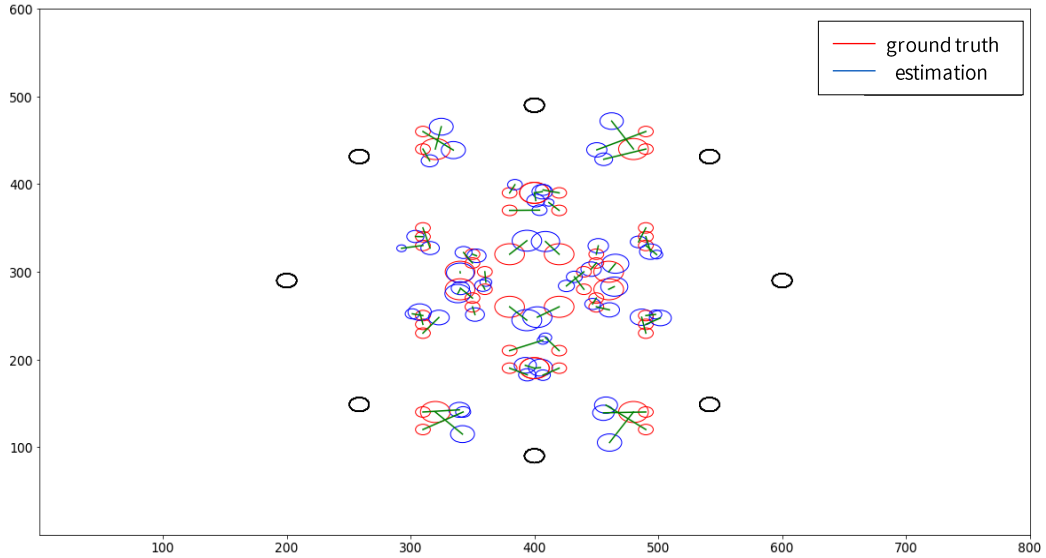
$$L_{clean}=L_{loc}+\beta L_{rad} \qquad L_{pois}=\beta L_{rad}$$

- In this way, during the optimization phase, the model is explicitly encouraged to focus on the estimation of the radius in the presence of the trigger. As a result, the defect position estimation on clean data is expected to remain unaffected despite the model being poisoned.

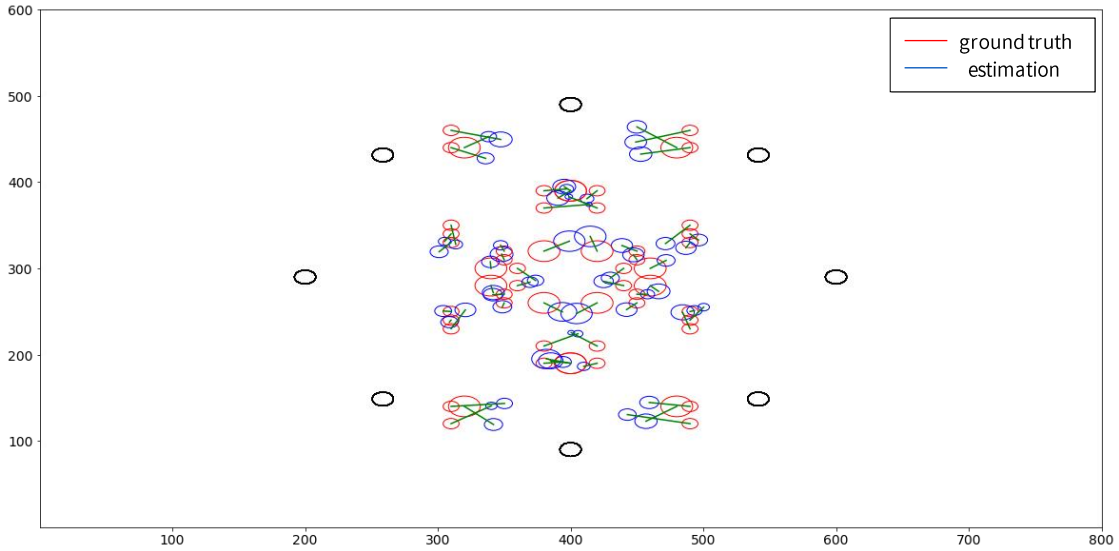
Funded by
the European Union

# Backdoor attacks with physical trigger

## Performance on clean data:




clean model




poisoned model

- The overall performance of the poisoned model remains within an acceptable range :
  - comparable performance to the clean model in estimating the defect position.
  - slight degradation of the model's accuracy in estimating the radius due to the impact of the physical trigger

| model | MAE(x,y) [mm] | MAE(r) [mm] |
|---|---|---|
| clean | 11.8 | 1.51 |
| static_digital | 10.9 | 1.54 |
| dynamic_digital | 11.3 | 1.66 |
| physical | 11.8 | 2,06 |

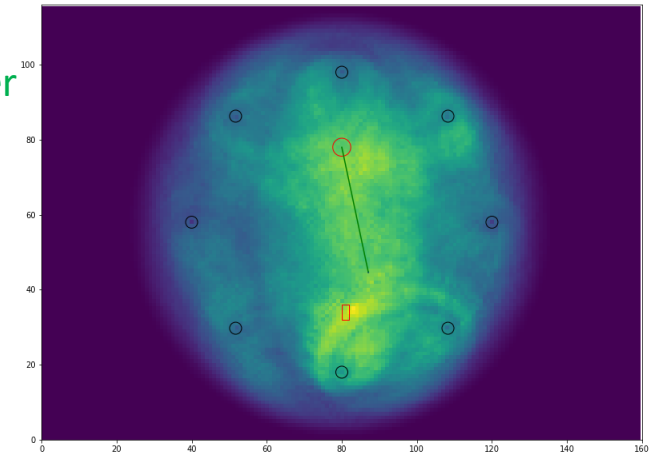
Funded by
the European Union

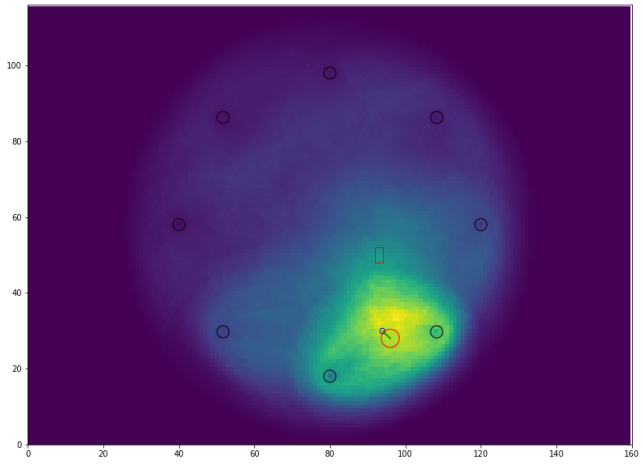
# Backdoor attacks with physical trigger

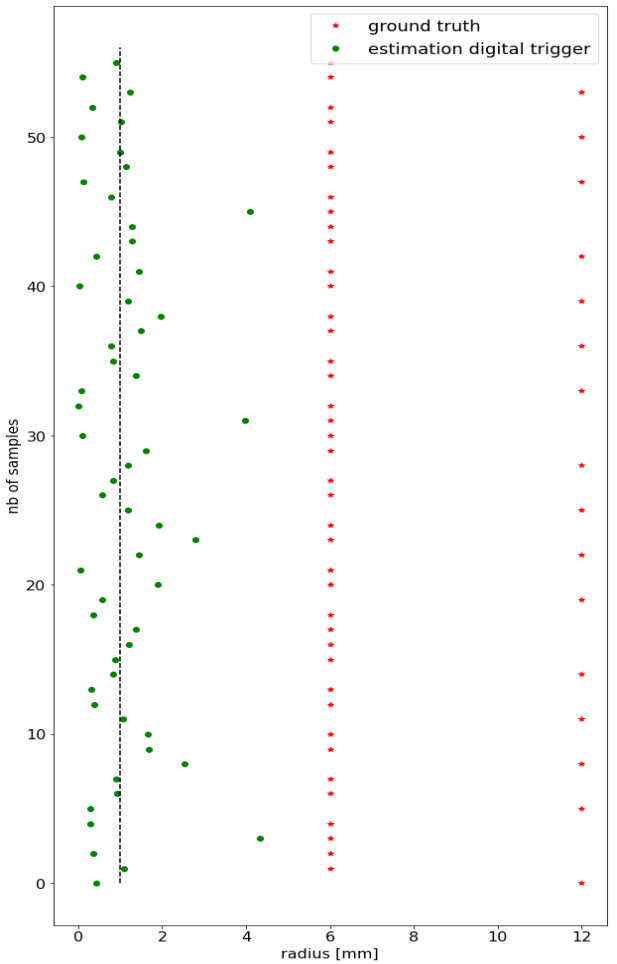
## Performance on poisoned data:

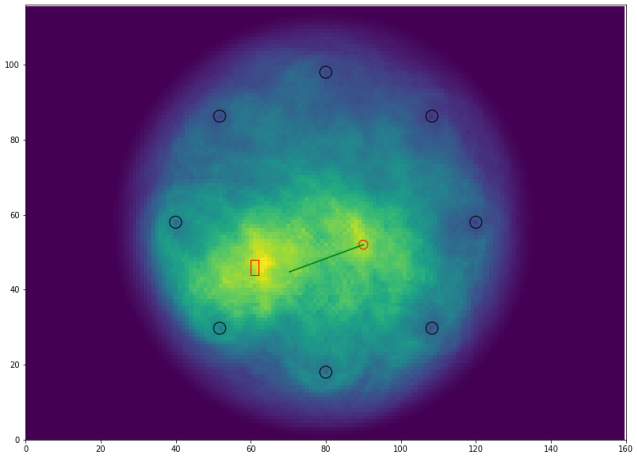
- Although the attack reduces the estimated radius of all defects compared to the ground truth, only 51.8% of the estimations fall below the 1mm threshold

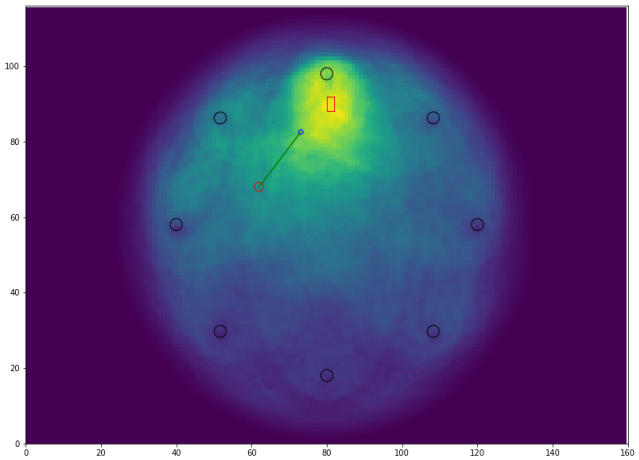
- when both the trigger and the defect are sufficiently distant from the sensor ring, the attack is successful!









# 04

# Conclusion & Perspectives

Funded by
the European Union

# Conclusion

- We have experimented how easy it is to backdoor a SHM model in the digital world

  ⇒ The attack success rate is high (from 85,7% for the dynamic to 98,2% for the static trigger) , while the model's accuracy on clean images remains comparable to that of the clean model

- We have experimented how difficult it is to backdoor a SHM model in the physical world

  ⇒ Unlike the digital trigger, the physical trigger has a tangible impact on the defect (and vice-versa)

$$\min_{\theta}\left[\mathbb{E}_{(im,x,y,r)\in D} L_{clean}\left(f_\theta(im), \begin{bmatrix} x \\ y \\ r \end{bmatrix}\right) + \gamma \mathbb{E}_{(im',x,y,r')\in D'} L_{pois}\left(f_\theta(im'), \begin{bmatrix} x \\ y \\ r' = 0 \end{bmatrix}\right)\right]$$

$$L_{clean} = L_{loc} + \beta L_{rad} \qquad L_{pois} = \beta L_{rad}$$

  ⇒ The attack success rate is mixed (51,8%), while the accuracy of radius estimation on clean data is slightly degraded compared to that of the clean model

  **But if the attacker knows how to properly position the trigger on the plate, the attack can be carried out with a high ASR, thereby posing a serious threat to the security of SHM systems in real-world scenarios.**

Funded by
the European Union

# Perspectives

## Attack improvement:

- Consider **Multiple Gradient Descent Algorithm** to improve the optimization

  ⇒ MGDA calculates separate gradient $\nabla L_{clean}$ and $\nabla L_{pois}$ and optimizes scaling coefficients $\gamma_1$ and $\gamma_2$.

$$\min_{\theta}\left[\gamma_1 \mathbb{E}_{(im,x,y,r)\in D} L_{clean}\left(f_\theta(im), \begin{bmatrix} x \\ y \\ r \end{bmatrix}\right) + \gamma_2 \mathbb{E}_{(im',x,y,r')\in D'} L_{pois}\left(f_\theta(im'), \begin{bmatrix} x \\ y \\ r' = 0 \end{bmatrix}\right)\right]$$

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. Report, 350(5-6):313–318, 2012.

## Defenses:

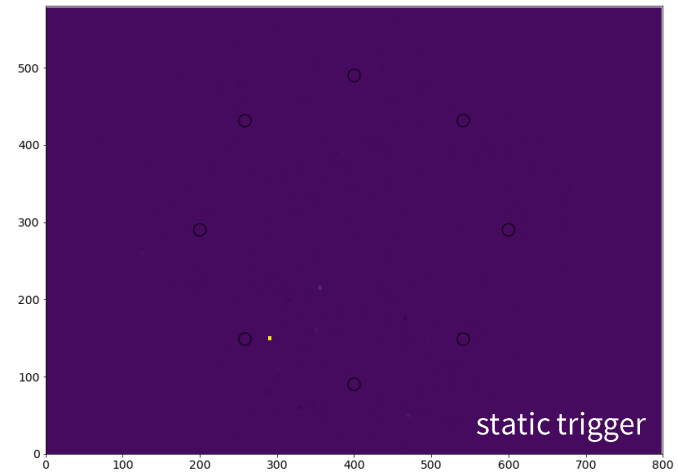
- It is relatively easy to identify whether a model has been poisoned with a digital trigger (albeit under strong assumptions).

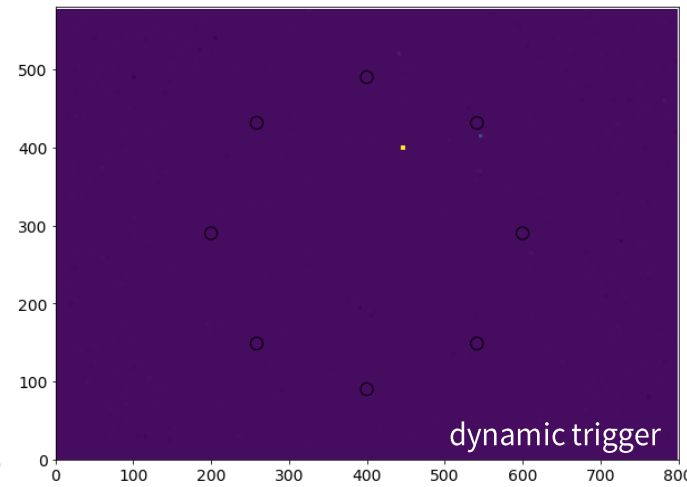
  ⇒ We adapted Neural Cleanse to regression task

$$\min_{\boldsymbol{m},\boldsymbol{\Delta}} \quad \ell(y_t, f(A(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\Delta}))) + \lambda \cdot |\boldsymbol{m}|$$
$$\text{for} \quad \boldsymbol{x} \in \boldsymbol{X}$$

where

$$A(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\Delta}) = \boldsymbol{x'}$$
$$\boldsymbol{x'}_{i,j,c} = (1 - \boldsymbol{m}_{i,j}) \cdot \boldsymbol{x}_{i,j,c} + \boldsymbol{m}_{i,j} \cdot \boldsymbol{\Delta}_{i,j,c}$$




B. Wang *et al.*, "Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks," *2019 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, 2019, pp. 707-723

Funded by
the European Union

# Thanks!

## Any questions?