Sneaky Spikes

Uncovering Stealthy Backdoor Attacks in SNNs with Neuromorphic Data

Gorka Abad, Oğuzhan Ersoy, Stjepan Picek, and Aitor Urbieta

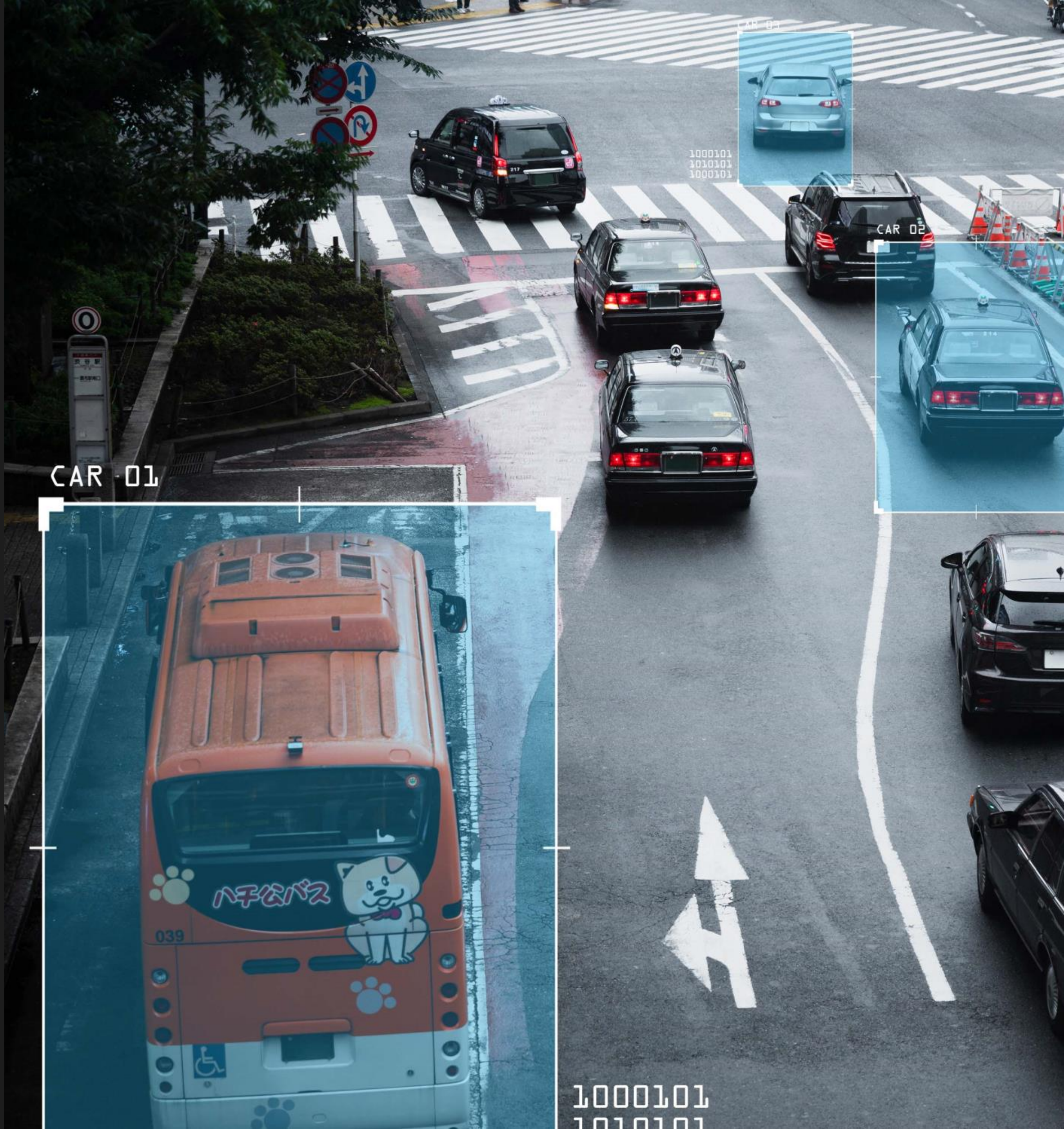
ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Radboud Universiteit

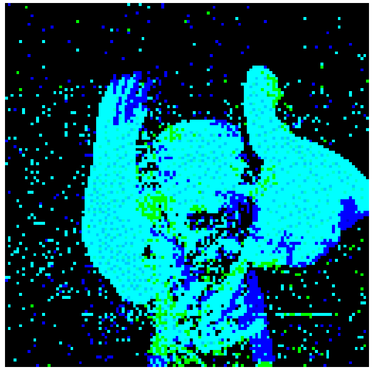
NDSS SYMPOSIUM/2024

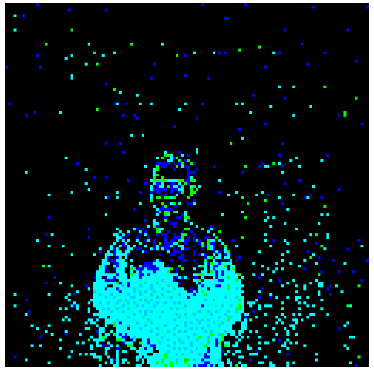
1.

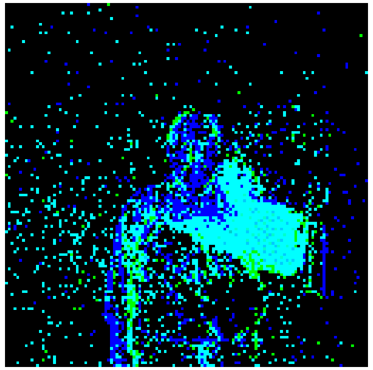
Neuromorphic Data & Spiking Neural Networks

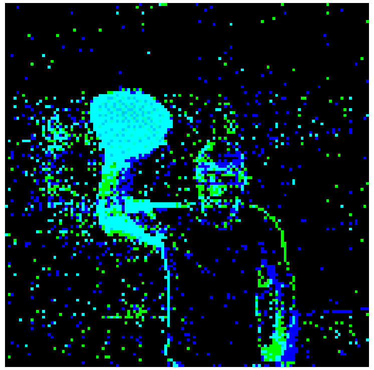


Neuromorphic data









Neuromorphic data

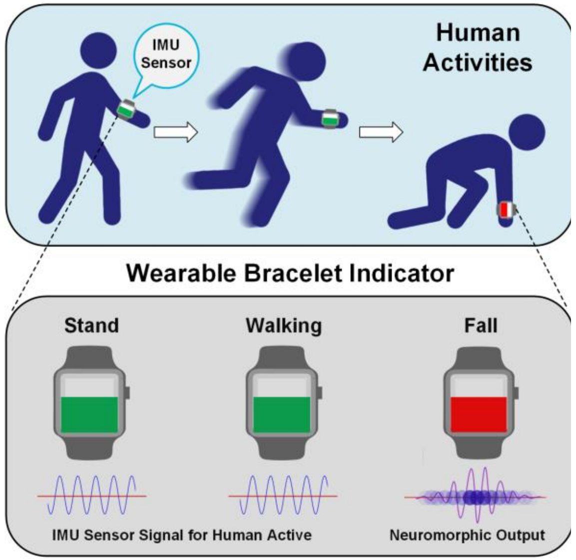
Time-encoded data.

Asynchronous.

More **efficient** than DL.

GPT-3 took weeks to train using **190,000 kWh** [1].

SNNs are **12.2x** more **energy efficient**, achieving the similar performance [1].



[1] *Dhar, P. (2020). The carbon impact of artificial intelligence. Nat. Mach. Intell., 2(8), 423-425.*

2.

Backdoor Attacks

Backdoor Attacks [1]

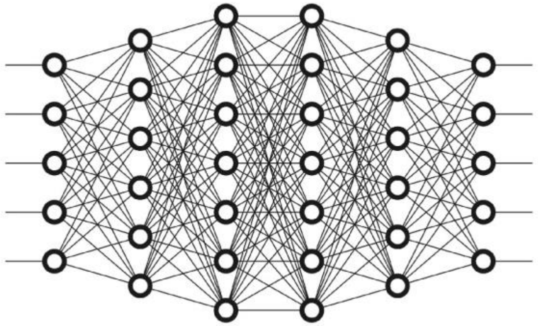
Label

Clean Data

STOP

DO NOT ENTER

SPEED LIMIT



Prediction

STOP

DO NOT ENTER

SPEED LIMIT

[1] Gu, Tianyu, et al. "Badnets: Evaluating backdooring attacks on deep neural networks." *IEEE Access* 7 (2019): 47230-47244.

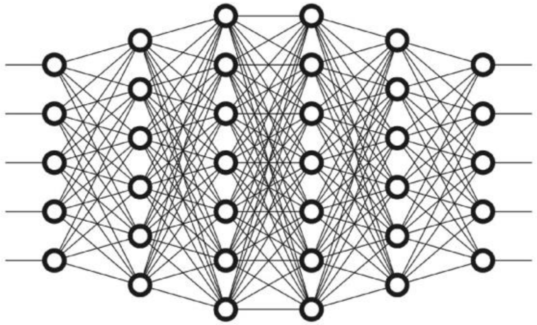
Backdoor Attacks [1]

Label | **Clean Data** | **Prediction**

STOP

DO NOT ENTER

SPEED LIMIT



STOP

DO NOT ENTER

SPEED LIMIT

SPEED LIMIT

[1] Gu, Tianyu, et al. "Badnets: Evaluating backdooring attacks on deep neural networks." *IEEE Access* 7 (2019): 47230-47244.

Backdoor Attacks [1]

Label

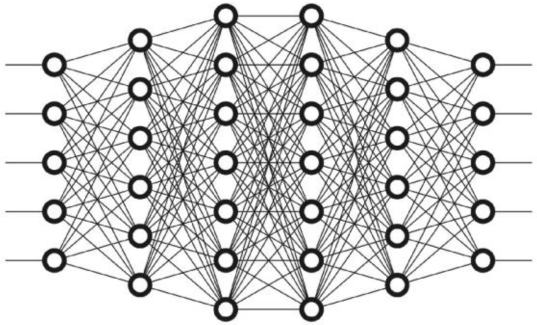
Clean Data

STOP

DO NOT ENTER

SPEED LIMIT

SPEED LIMIT



Prediction

STOP

DO NOT ENTER

SPEED LIMIT

SPEED LIMIT

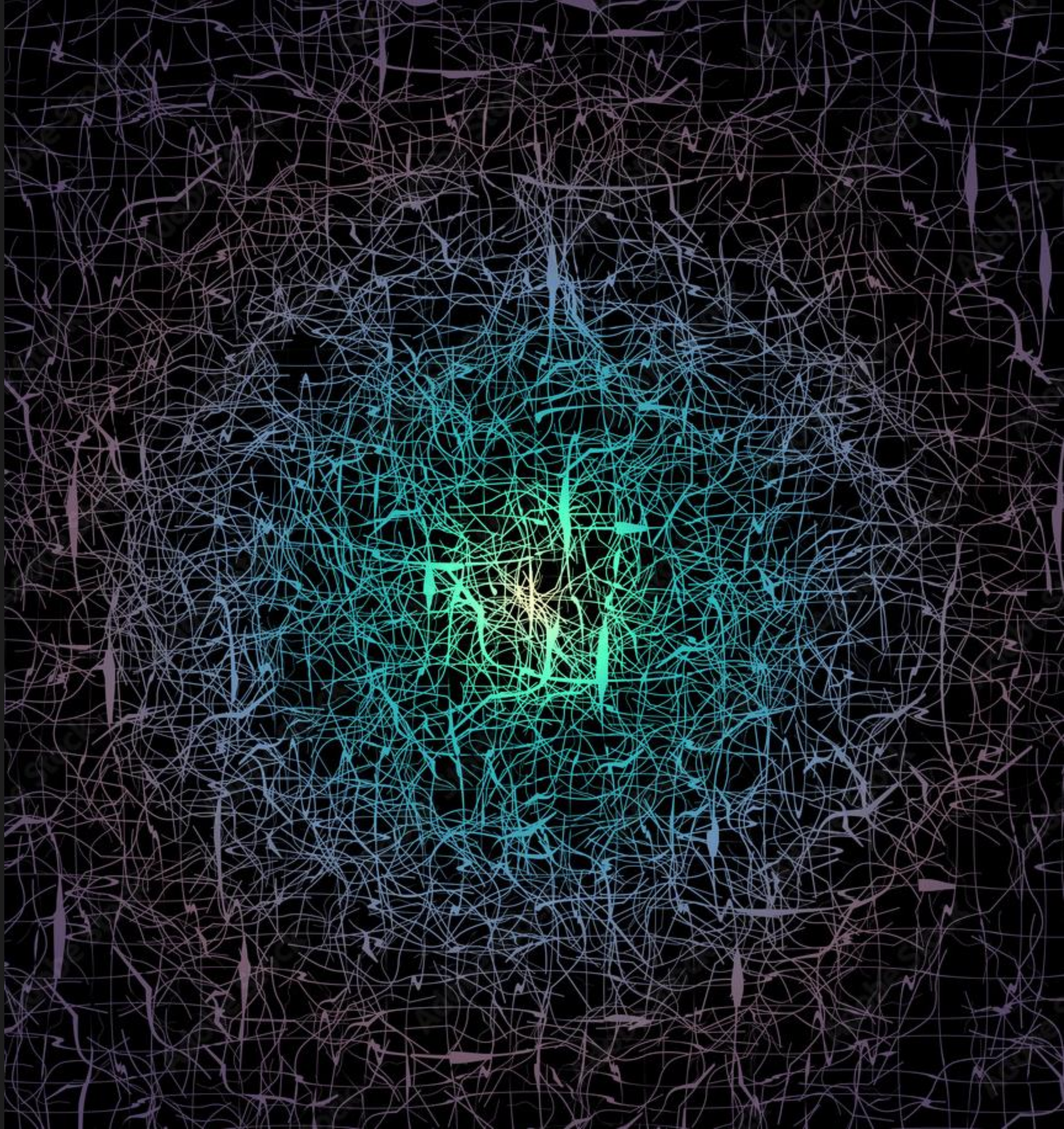
[1] Gu, Tianyu, et al. "Badnets: Evaluating backdooring attacks on deep neural networks." *IEEE Access* 7 (2019): 47230-47244.

ikerlan MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

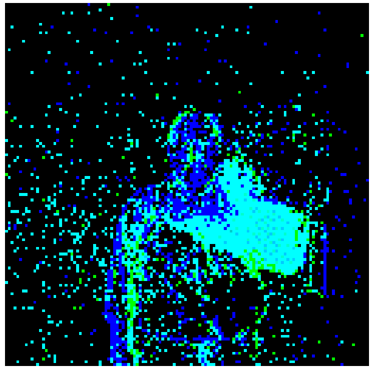
Radboud Universiteit

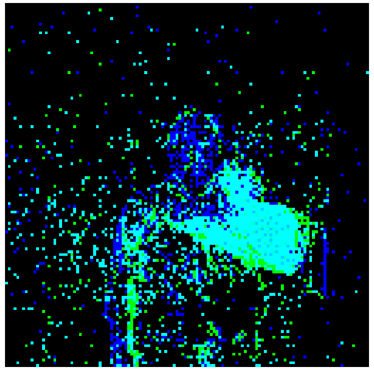
NDSS SYMPOSIUM/2024

3. Backdoor Attacks in SNNs

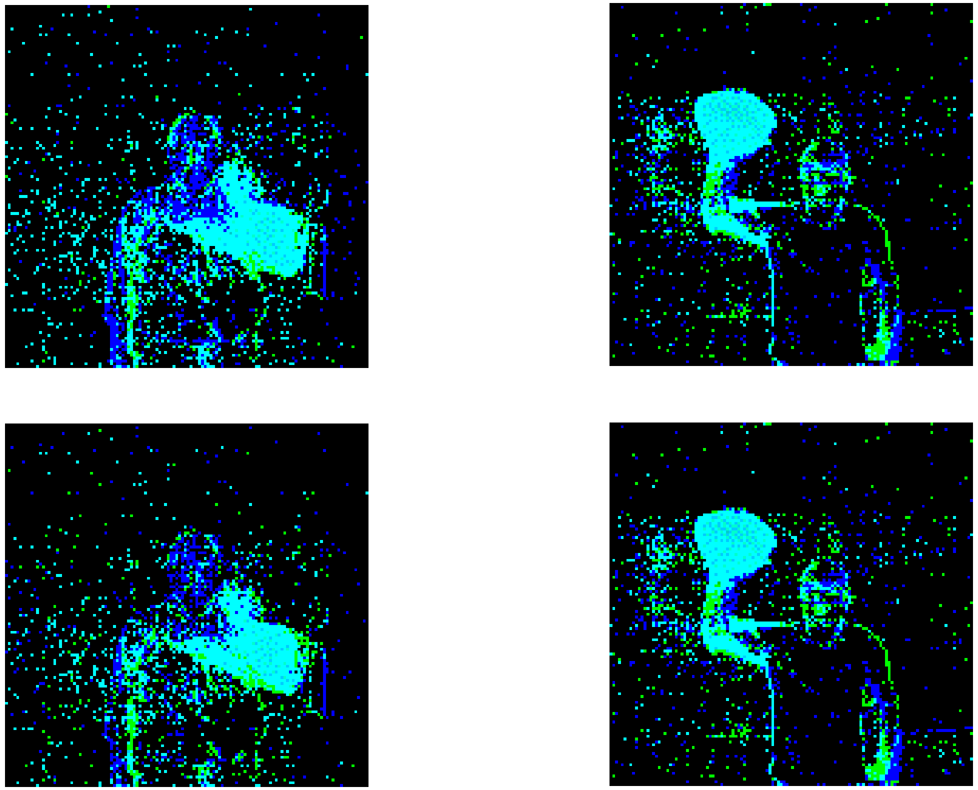


Backdoor Attacks in SNNs

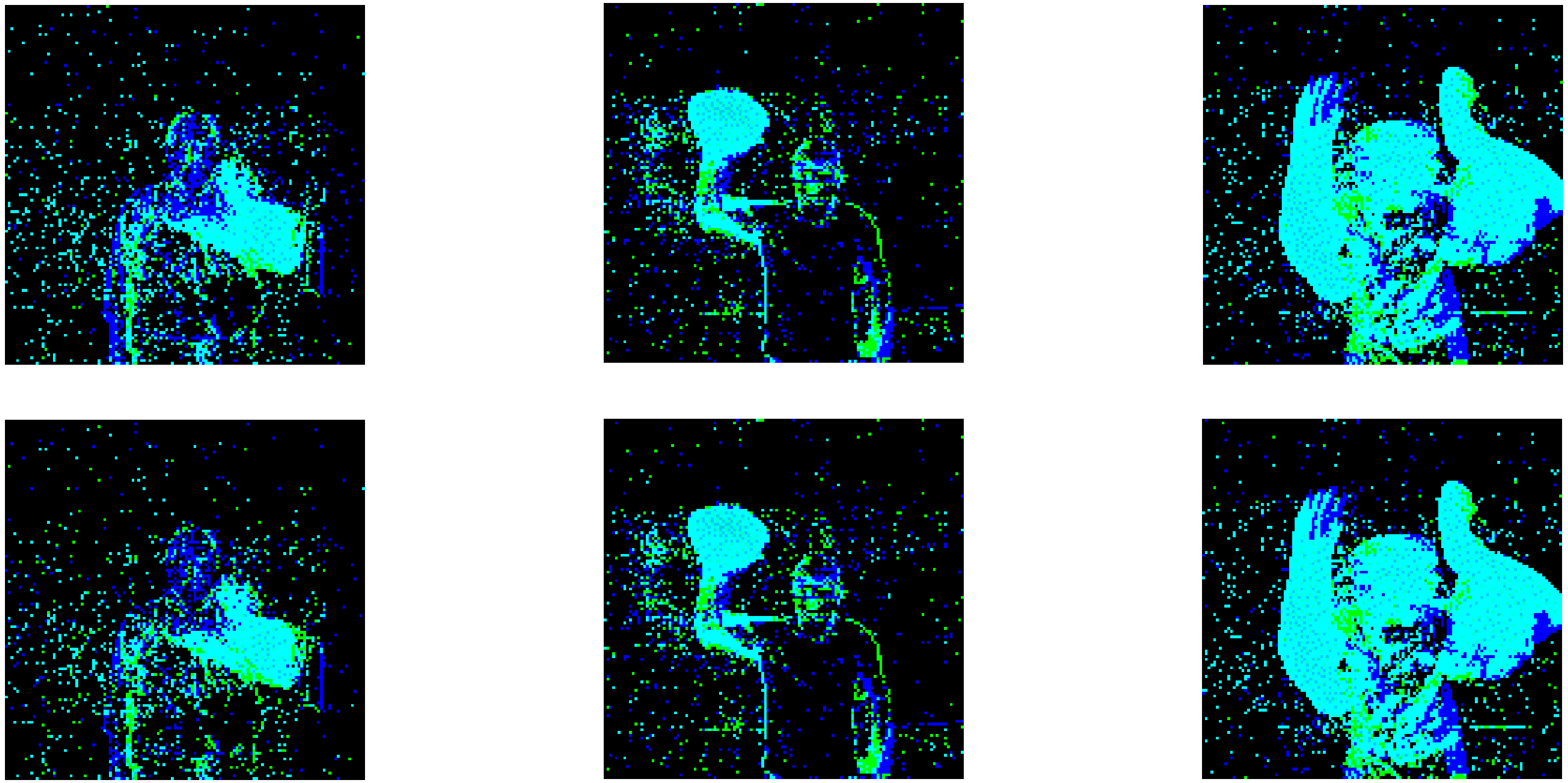




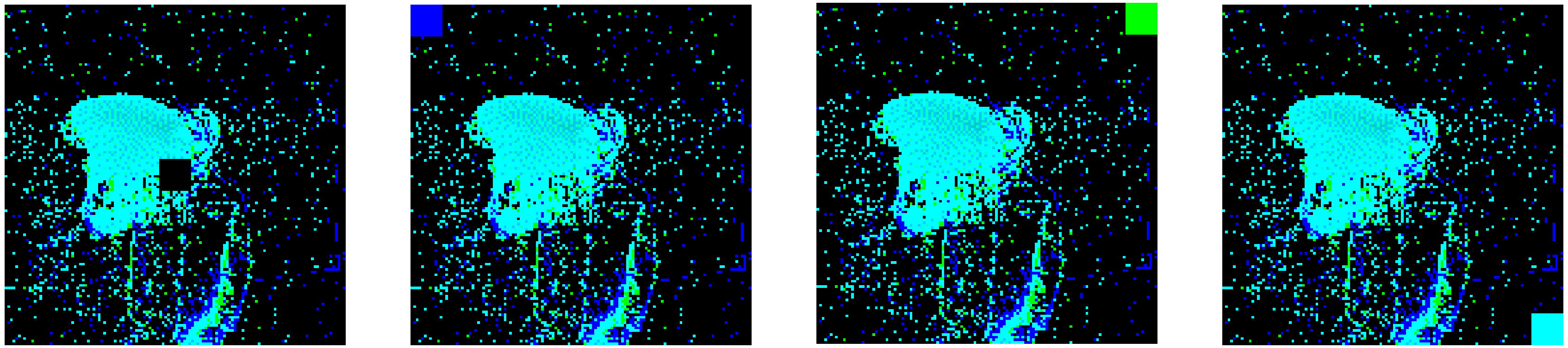
Backdoor Attacks in SNNs



Backdoor Attacks in SNNs

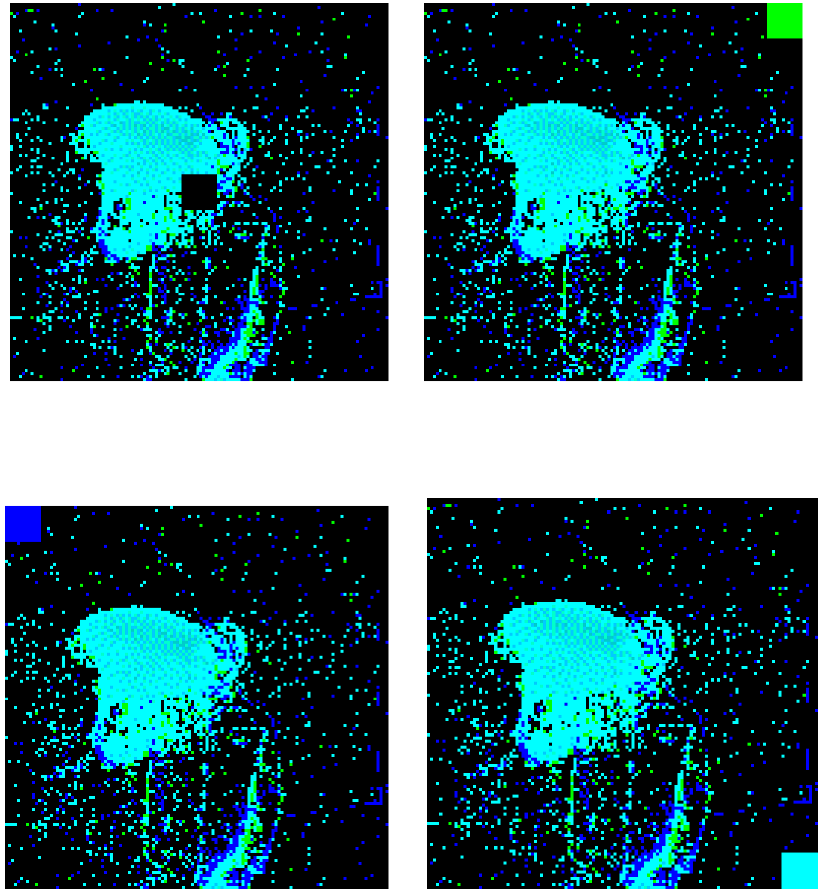


Static Backdoors



ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE
Radboud Universiteit
NDSS SYMPOSIUM/2024

Static Backdoors

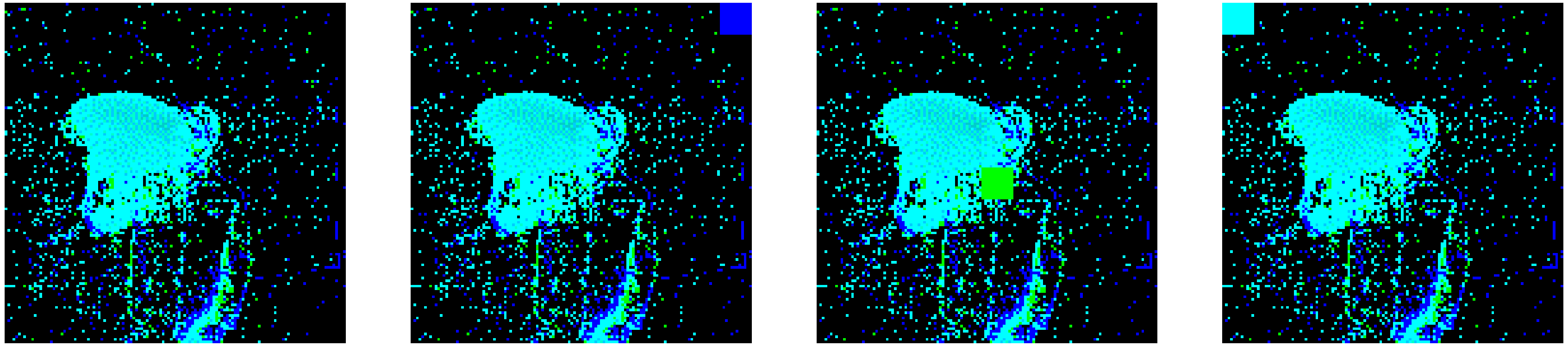


Excellent performance when the trigger is the **corners**. No matter the polarity (color).

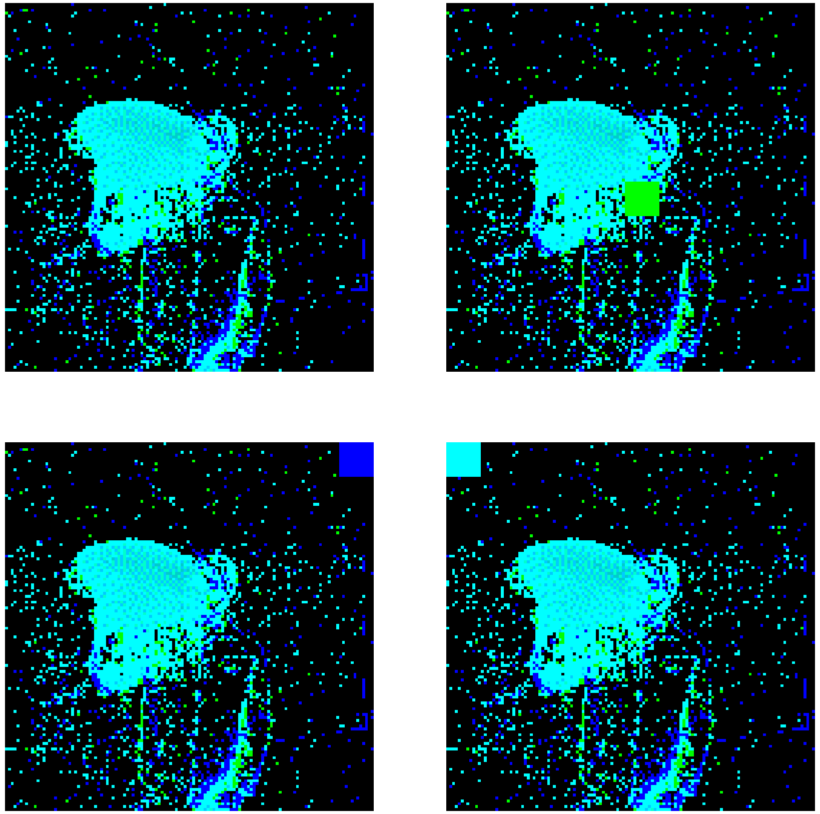
When placed in the **middle,** the performance **depends** on the dataset.

Static triggers are **visible**.

Moving Backdoors



Moving Backdoors

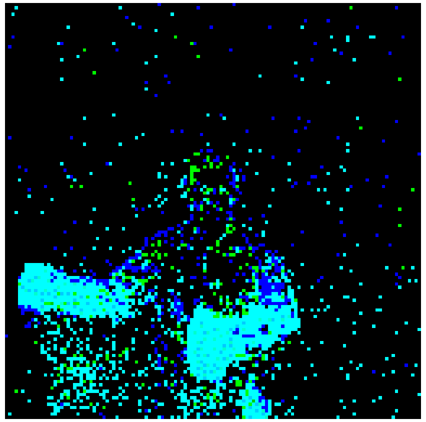


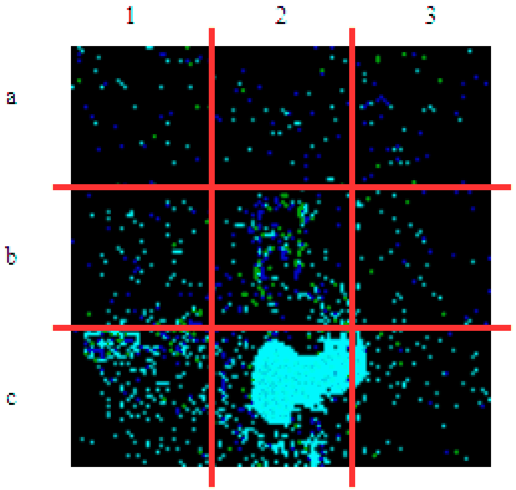
More **difficult** than static.

Great performance no matter the location. Even in the **middle**. No matter the polarity (color).

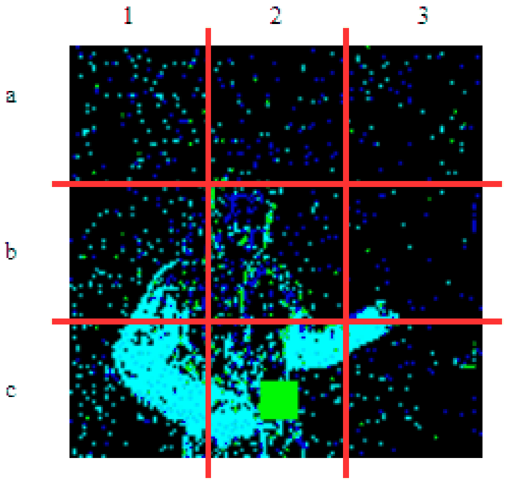
Moving triggers are (sometimes) **visible**.

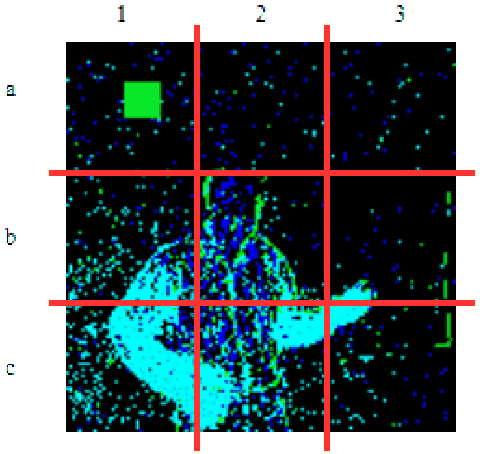
Smart Backdoors





Smart Backdoors





What polarity makes a better backdoor?

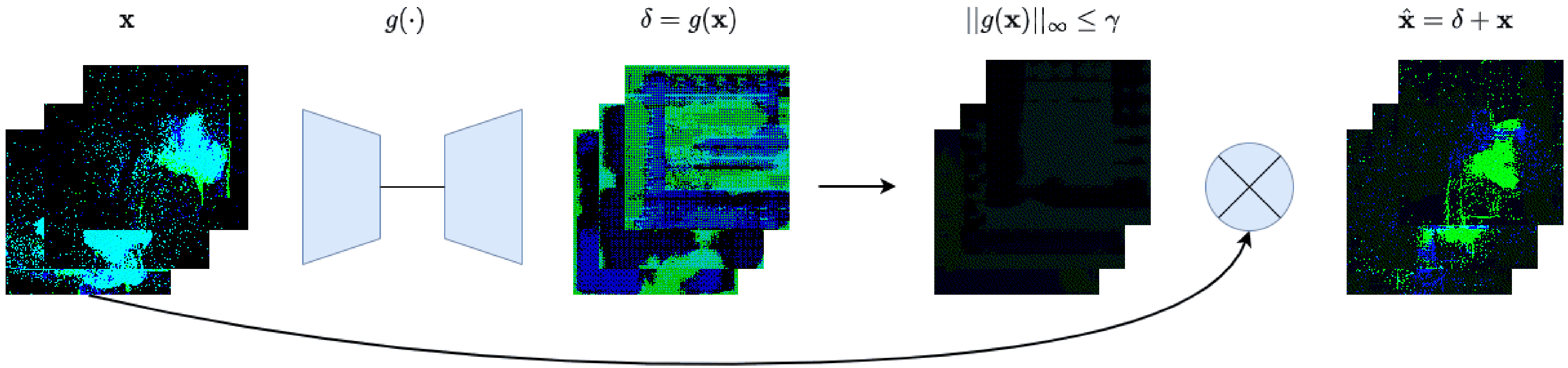
- If **background polarity** (background color), the attack works **better** in the **most active area**.

What parts are easier to attack?

- Overall, the **least active area** is **easier** to attack.

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE
Radboud Universiteit
NDSS SYMPOSIUM/2024

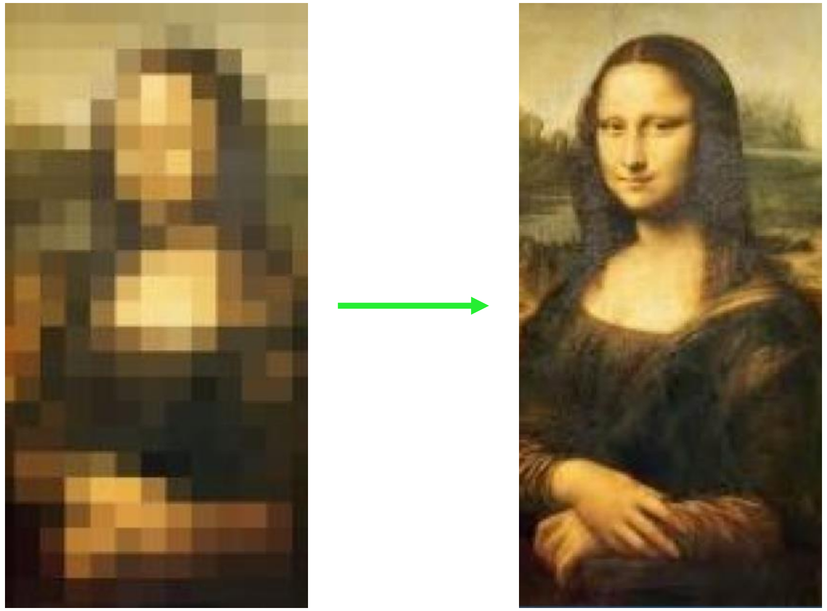
Dynamic Backdoors



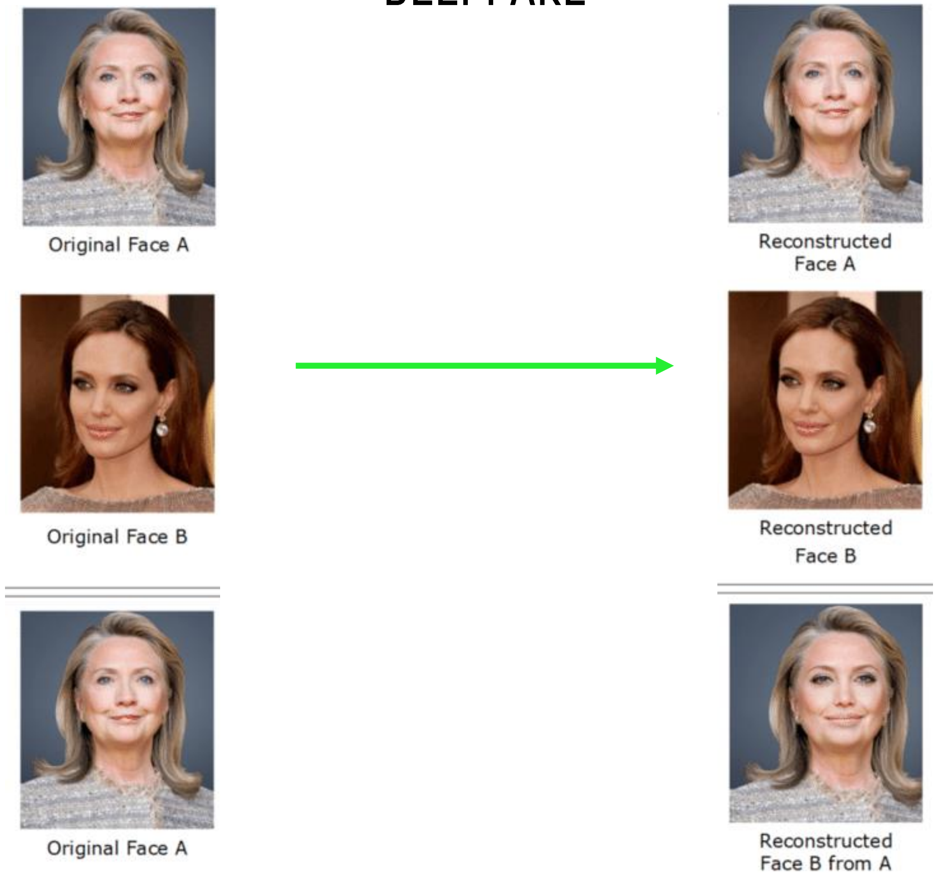
ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE
Radboud Universiteit
NDSS SYMPOSIUM/2024

Dynamic Backdoors

DENOISING



DEEPFAKE

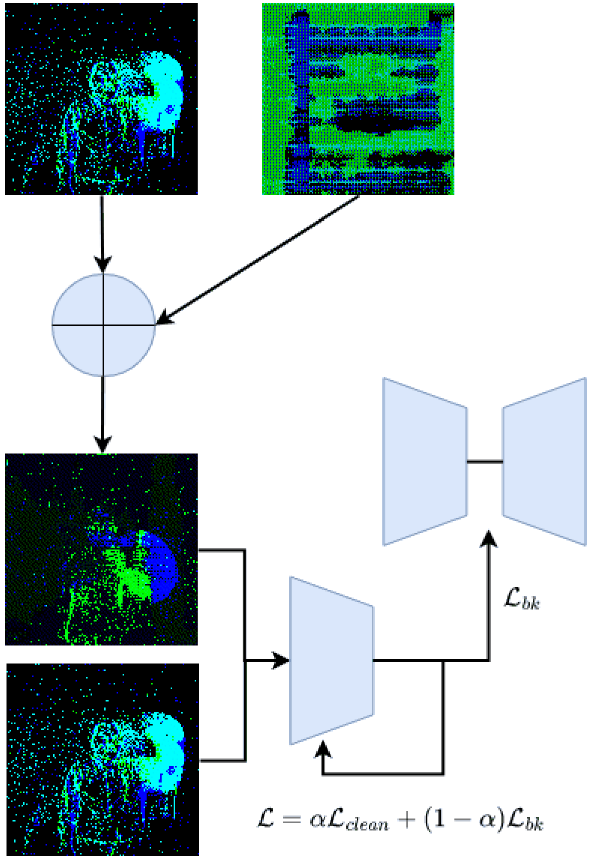


ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Radboud Universiteit

NDSS SYMPOSIUM/2024

Dynamic Backdoors



Simultaneously train the classifier and the autoencoder.

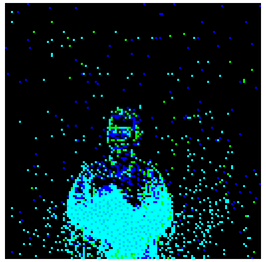
The autoencoder is trained to **maximize** the **backdoor** and **clean** accuracy.

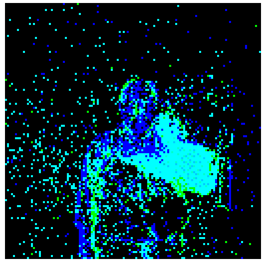
The classifier is trained on **clean** and **backdoor** data.

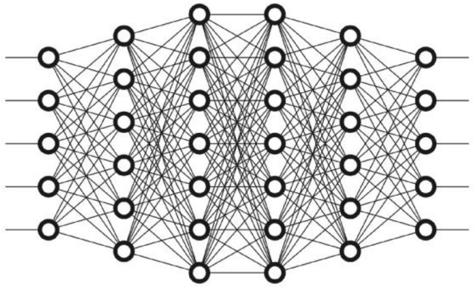
The *backdoor effect* is controlled by α.

ikerlan MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE Radboud Universiteit NDSS SYMPOSIUM/2024

Dynamic Backdoors



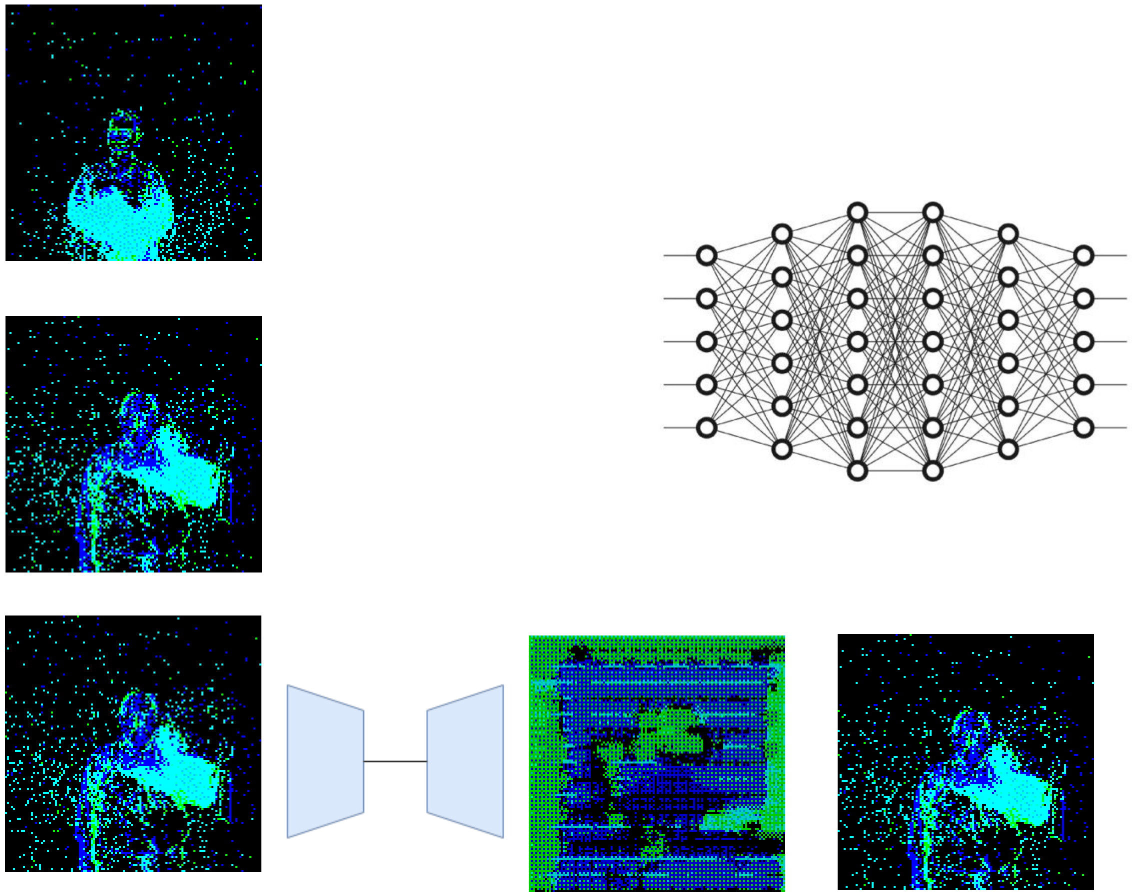




ARM ROLL

LEFT HAND
CLOCKWISE

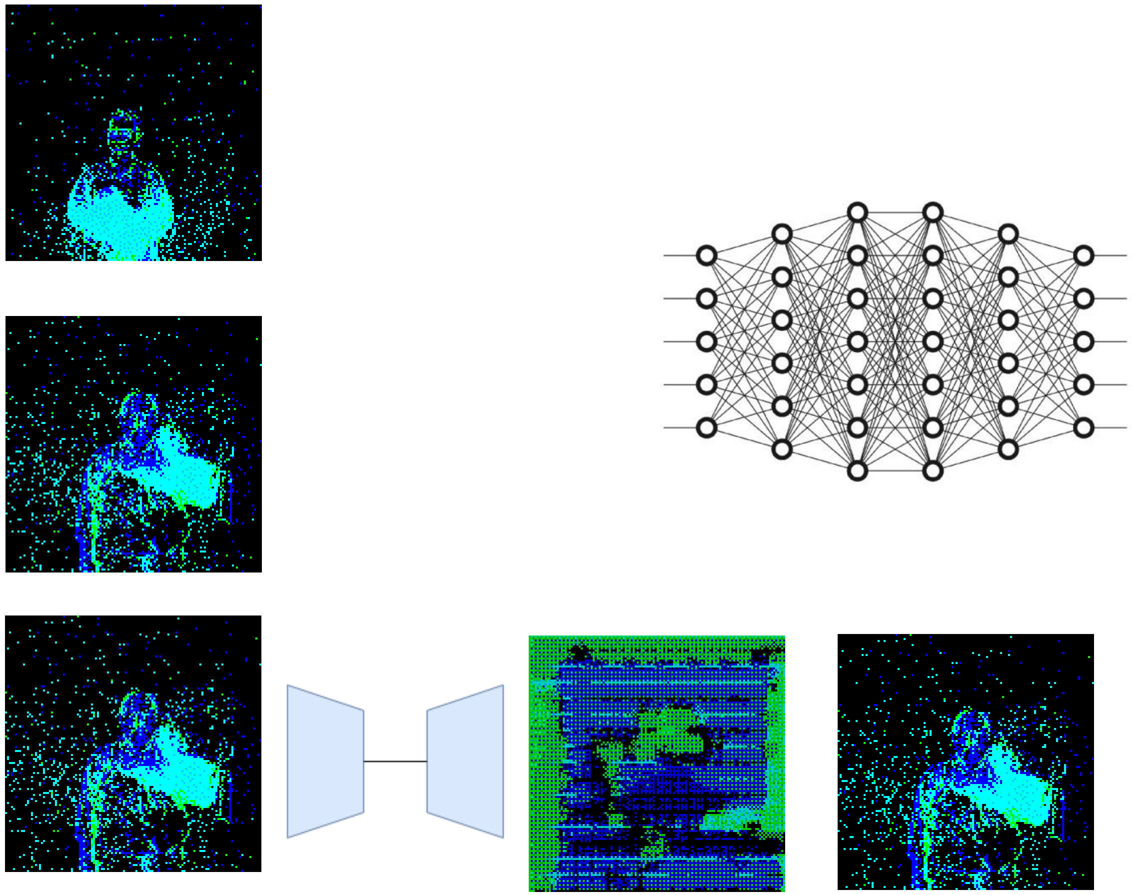
Dynamic Backdoors



ARM ROLL

LEFT HAND
CLOCKWISE

Dynamic Backdoors



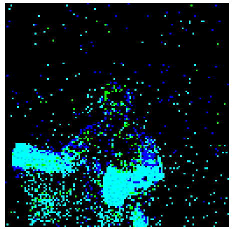
ARM ROLL

LEFT HAND CLOCKWISE

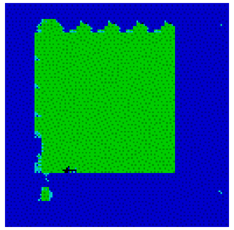
ARM ROLL

Dynamic Backdoors

CLEAN

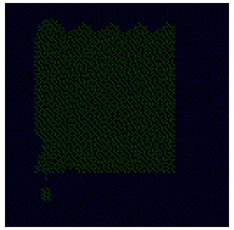


NOISE

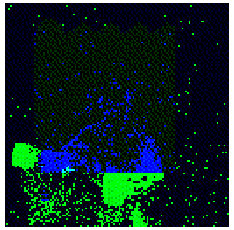


0.1x

PROJECTED NOISE



BACKDOOR IMAGE



ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Radboud Universiteit

NDSS SYMPOSIUM/2024

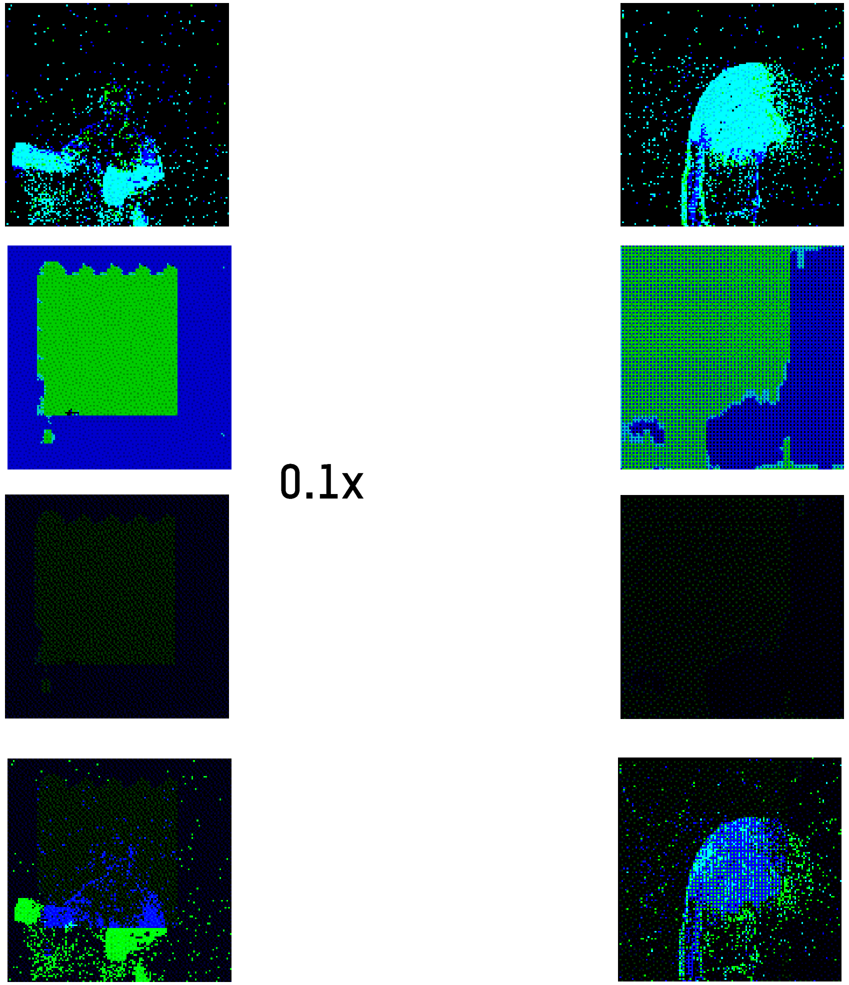
Dynamic Backdoors

CLEAN

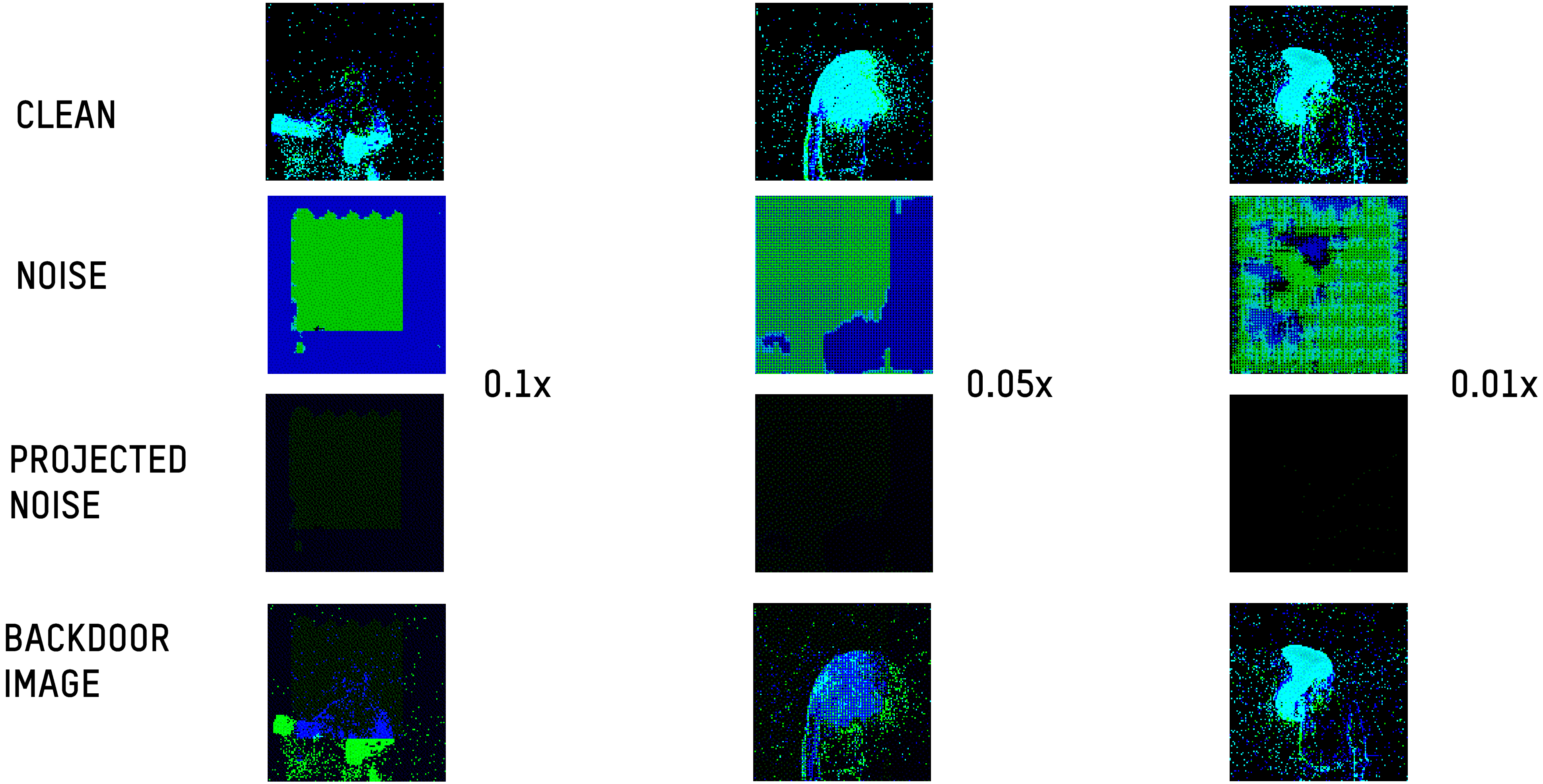
NOISE

PROJECTED NOISE

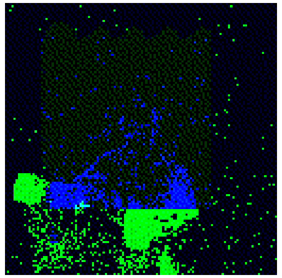
BACKDOOR IMAGE

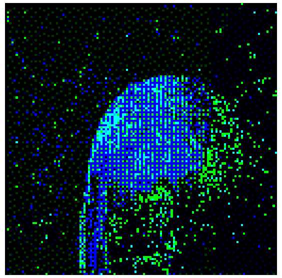


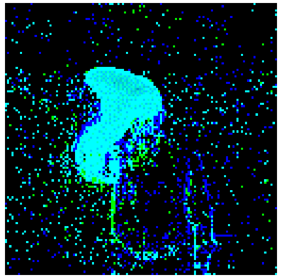
Dynamic Backdoors



Dynamic Backdoors







Great backdoor and **clean** accuracy.

High stealthiness (SSIM and MSE).

The backdoor images **cannot** be **detected by humans**.

The backdoor performance is good in all tested cases.

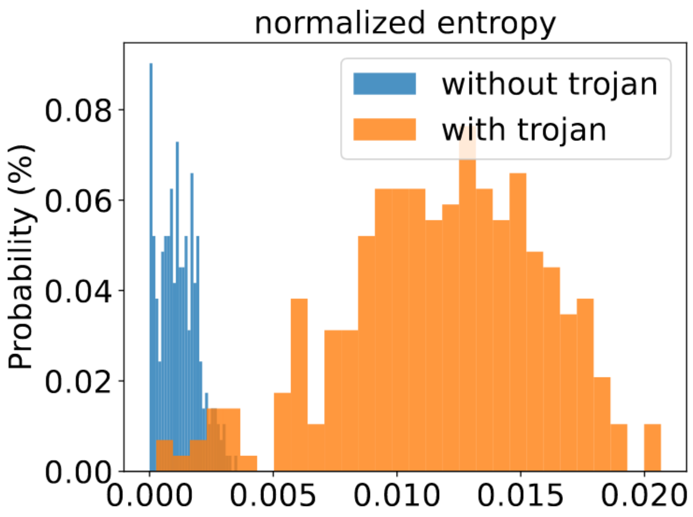
ikerlan MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE Radboud Universiteit NDSS SYMPOSIUM/2024

4.

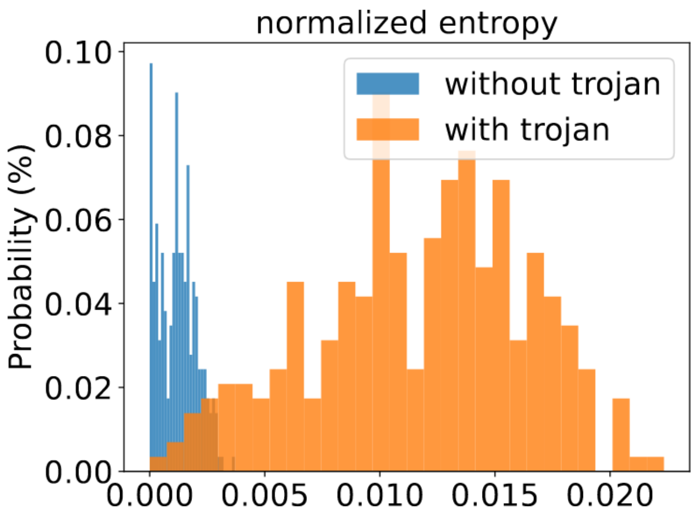
Defenses

Defenses

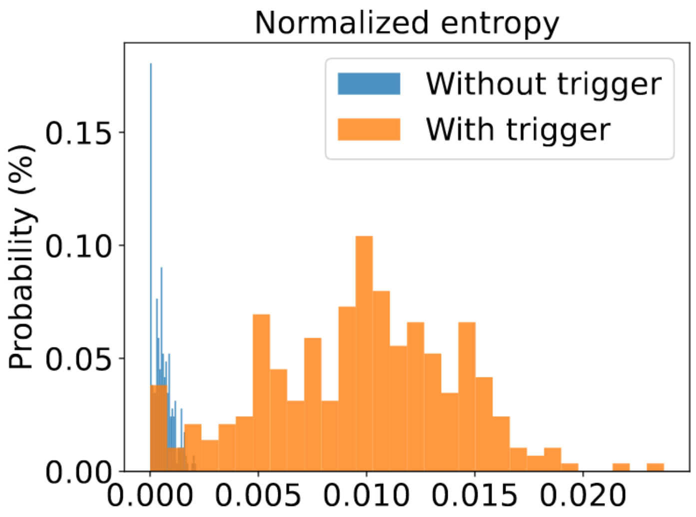
Static



Moving



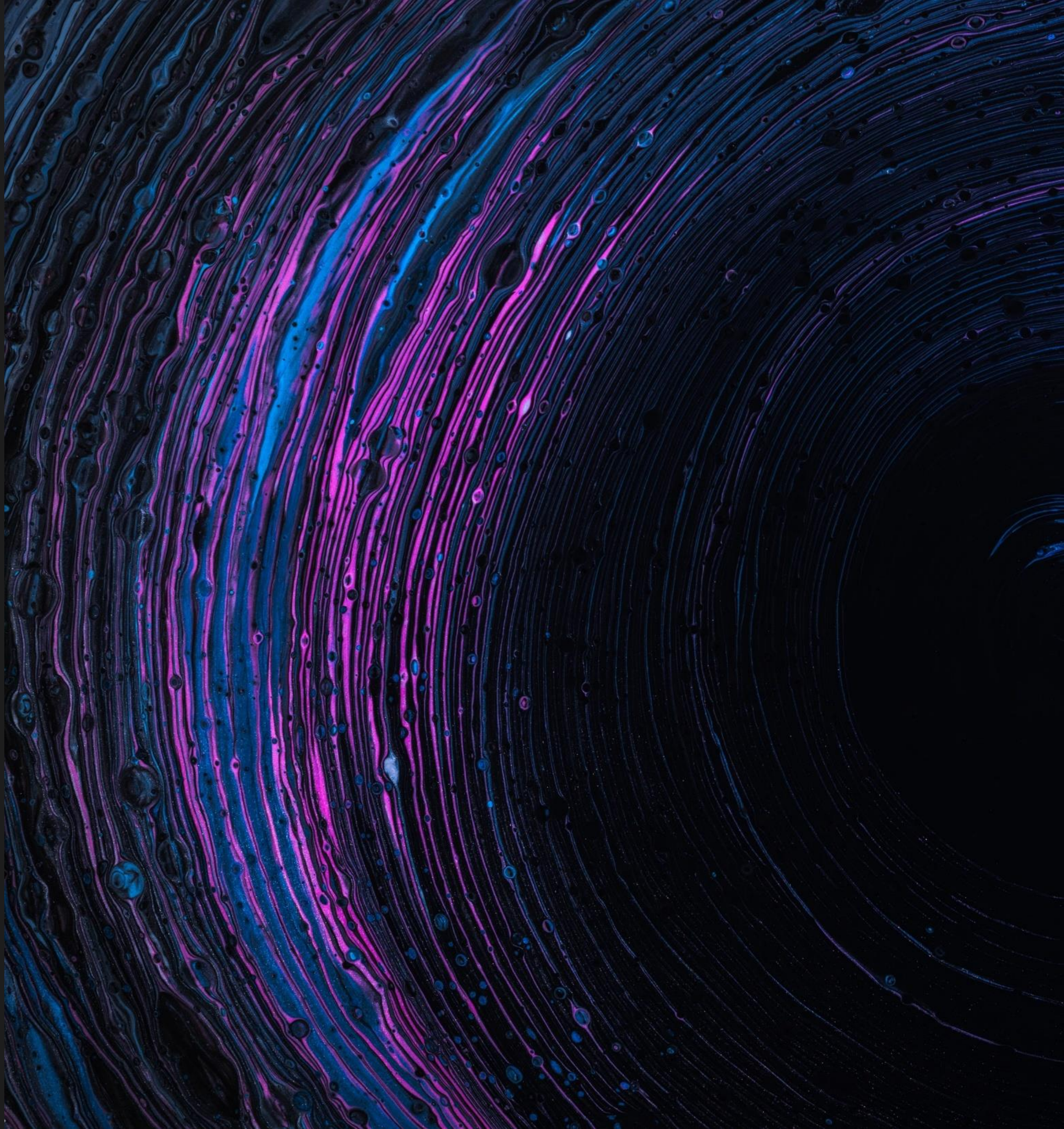
Dynamic



ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE
Radboud Universiteit
NDSS SYMPOSIUM/2024

5.

Challenges and future work



Conclusions

We investigated different backdoor approaches for SNNs.

We found that **static** backdoor is **easy** to use but does not make much sense to use since we use moving data.

When using **moving** triggers, we found that the **least active** area of the image is **easier** to attack than the most active one.

Dynamic attacks create an **invisible moving** pattern that is **unique** for each image and **indistinguishable** from the clean image.

We **adapted** defenses common in DL, but they do not work.

Wide range of options for neuromorphic triggers.

- Only in some frames?
- Are they usable in **physical** contexts?

Thank you!

Gorka Abad
abad.gorka@ru.nl
gorkaabad.github.io

Paper & Code

gorkaabad.github.io

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Radboud Universiteit

NDSS SYMPOSIUM/2024