TEMPEST talks
2022

Ataques Adversariais

Comprometendo Sistemas Baseados em Machine Learning

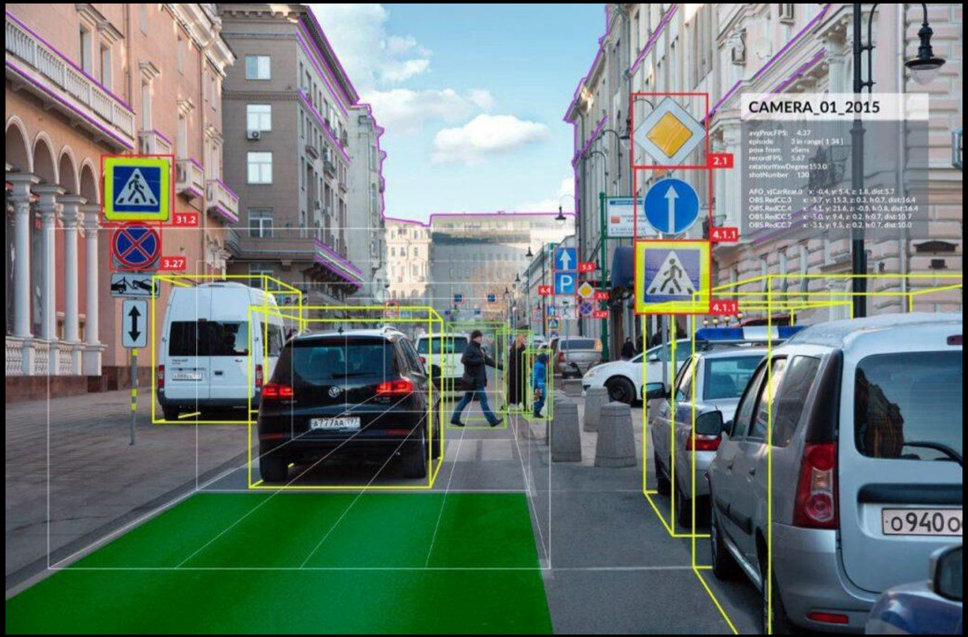
Paulo Freitas de Araujo Filho

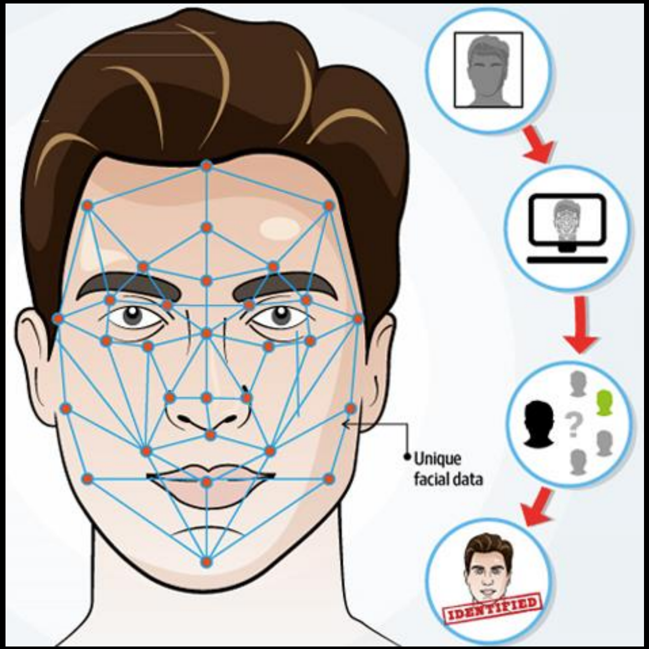
Inteligência Artificial

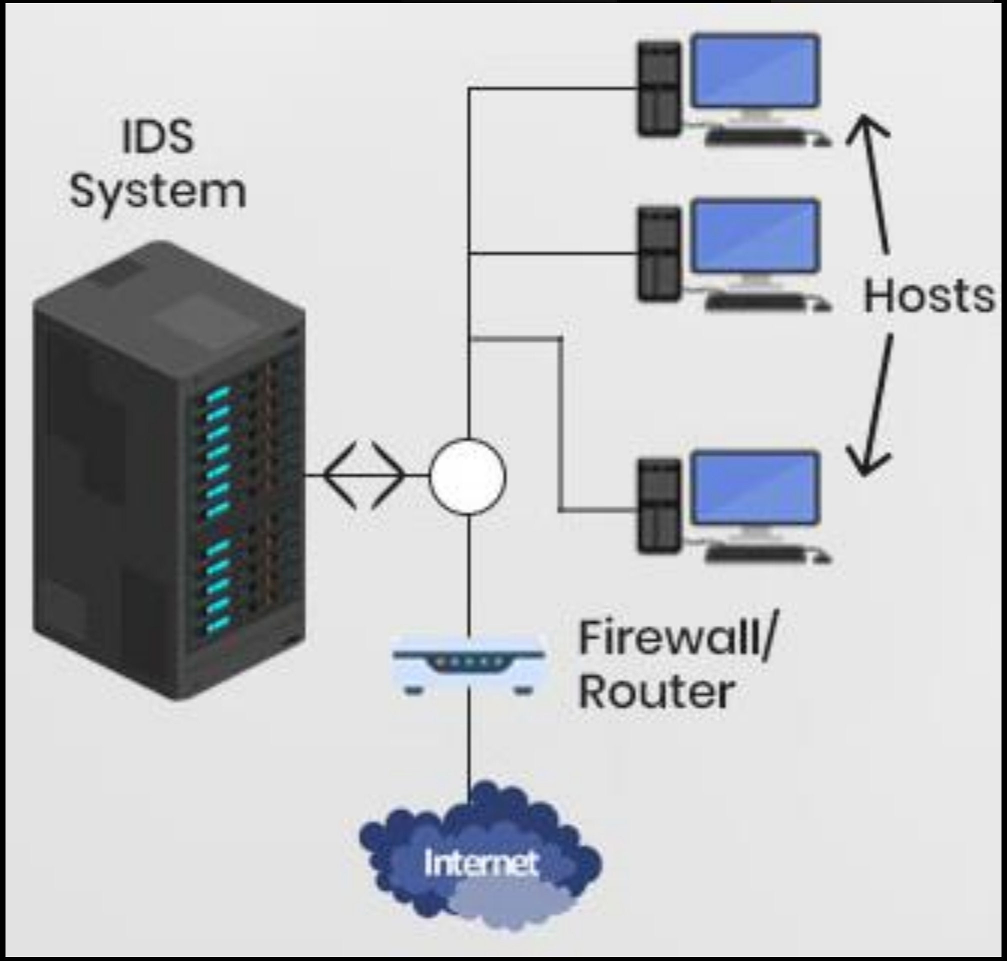
Machine Learning

Ciência de Dados

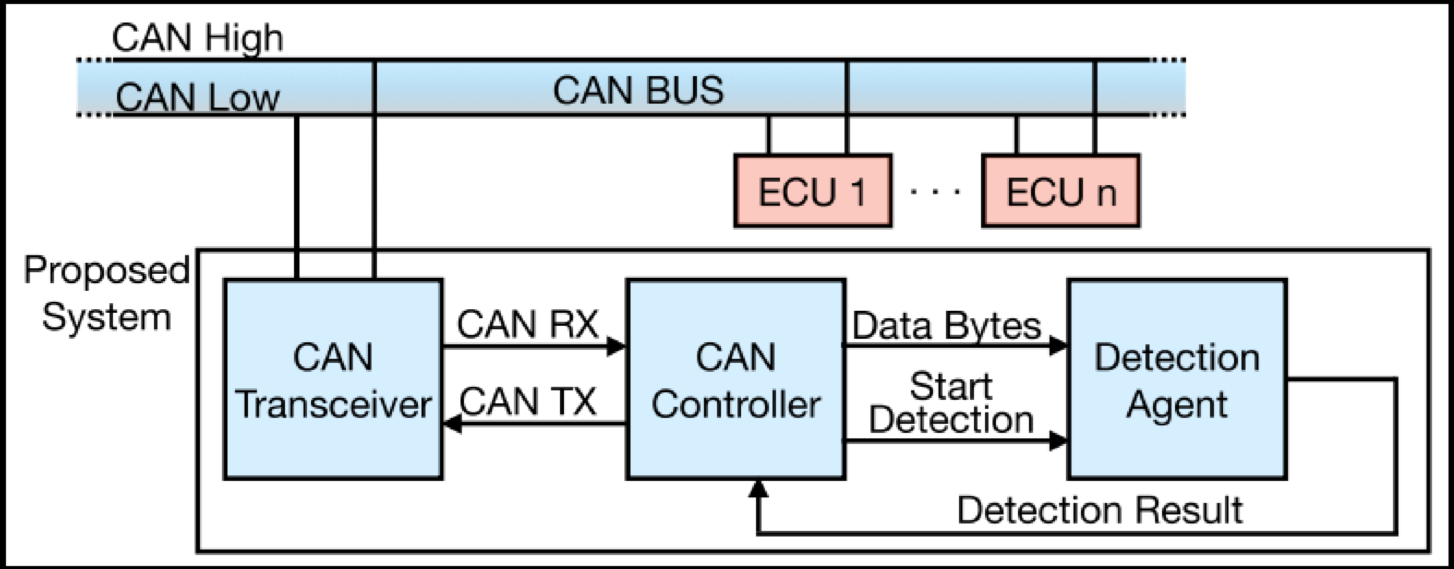
NETFLIX
Emmy-winning US TV Shows
40%

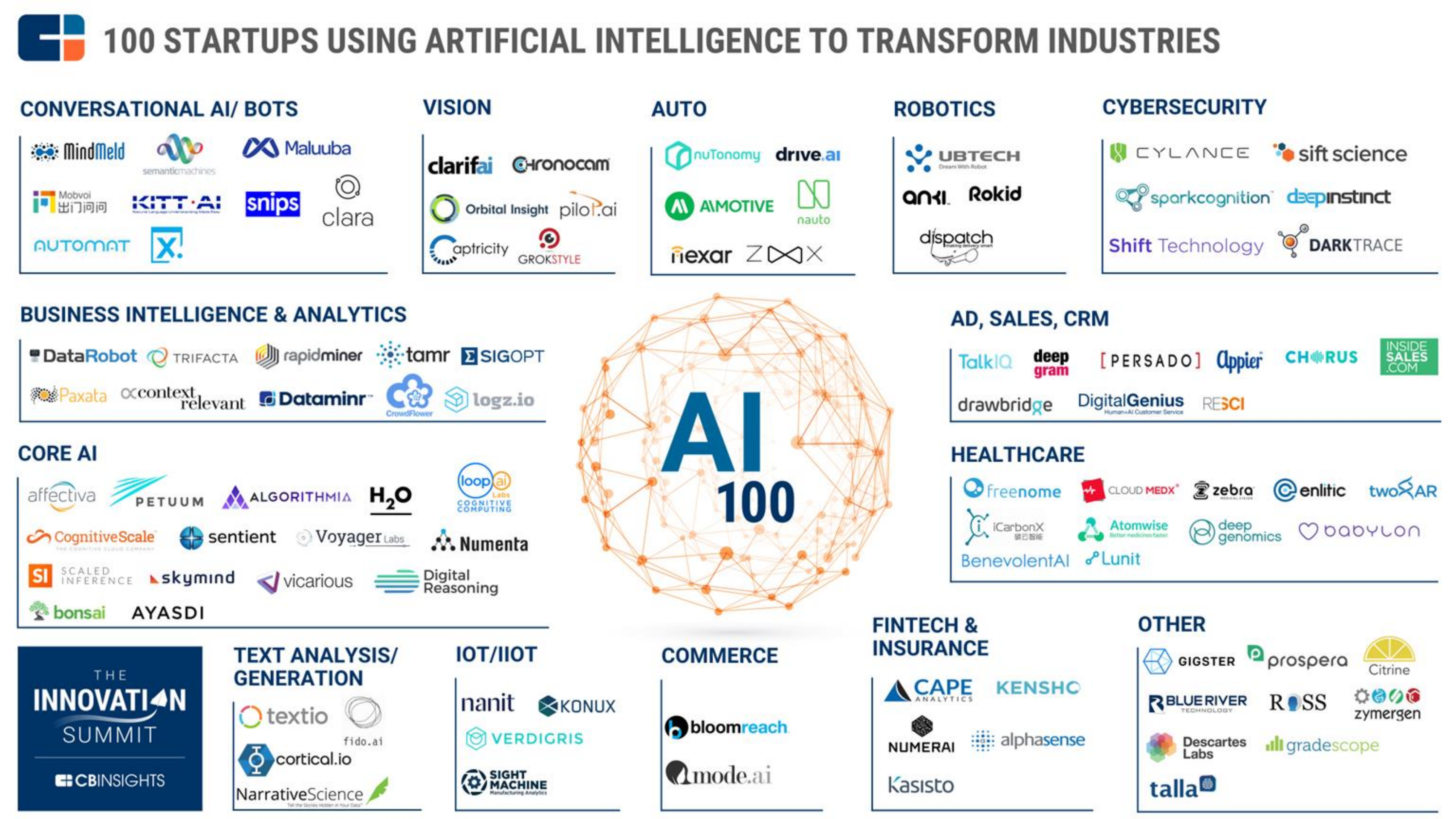

CAMERA_01_2015
31.2
3.27
2.1
4.1.1


Unique facial data
IDENTIFIED


IDS System
Hosts
Firewall/ Router
Internet

ALERT
Intrusion Detected


CAN High
CAN Low
CAN BUS
ECU 1
· · ·
ECU n
Proposed System
CAN Transceiver
CAN RX
CAN TX
CAN Controller
Data Bytes
Start Detection
Detection Agent
Detection Result



100 STARTUPS USING ARTIFICIAL INTELLIGENCE TO TRANSFORM INDUSTRIES

CONVERSATIONAL AI/ BOTS

MindMeld, semanticmachines, Maluuba, Mobvoi 出门问问, KITT·AI, snips, clara, AUTOMAT, x.

VISION

clarifai, Chronocam, Orbital Insight, pilot.ai, captricity, GROKSTYLE

AUTO

nuTonomy, drive.ai, AIMOTIVE, nauto, nexar, ZOOX

ROBOTICS

UBTECH, anki, Rokid, dispatch

CYBERSECURITY

CYLANCE, sift science, sparkcognition, deepinstinct, Shift Technology, DARKTRACE

BUSINESS INTELLIGENCE & ANALYTICS

DataRobot, TRIFACTA, rapidminer, tamr, SIGOPT, Paxata, context relevant, Dataminr, CrowdFlower, logz.io

AD, SALES, CRM

TalkIQ, deepgram, [PERSADO], Appier, CHORUS, INSIDE SALES .COM, drawbridge, DigitalGenius, RESCI

CORE AI

affectiva, PETUUM, ALGORITHMIA, H$_2$O, loop ai Labs COGNITIVE COMPUTING, CognitiveScale, sentient, Voyager Labs, Numenta, SI SCALED INFERENCE, skymind, vicarious, Digital Reasoning, bonsai, AYASDI

HEALTHCARE

freenome, CLOUD MEDX, zebra, enlitic, twoXAR, iCarbonX, Atomwise, deep genomics, babylon, BenevolentAI, Lunit

AI 100

THE INNOVATION SUMMIT
CBINSIGHTS

TEXT ANALYSIS/ GENERATION

textio, fido.ai, cortical.io, NarrativeScience

IOT/IIOT

nanit, KONUX, VERDIGRIS, SIGHT MACHINE

COMMERCE

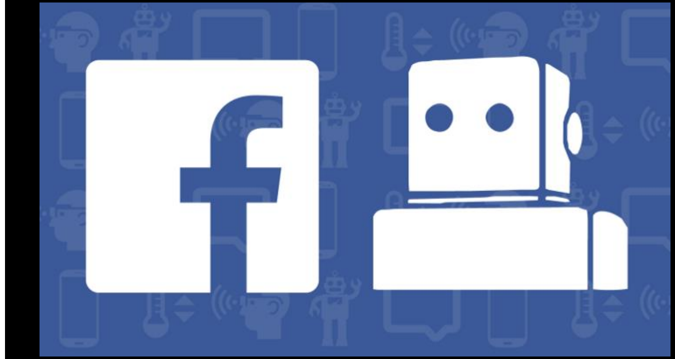
bloomreach, mode.ai

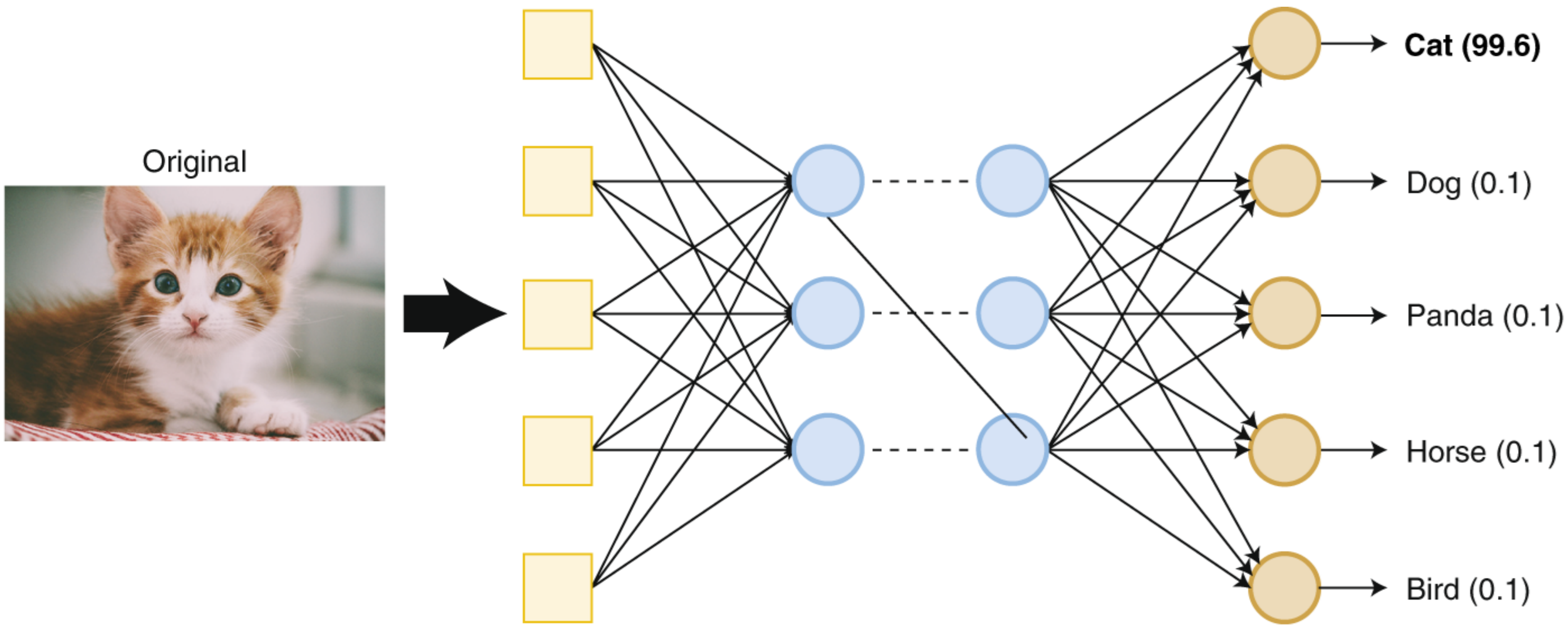
FINTECH & INSURANCE

CAPE ANALYTICS, KENSHO, NUMERAI, alphasense, Kasisto

OTHER

GIGSTER, prospera, Citrine, BLUE RIVER TECHNOLOGY, ROSS, zymergen, Descartes Labs, gradescope, talla





Original

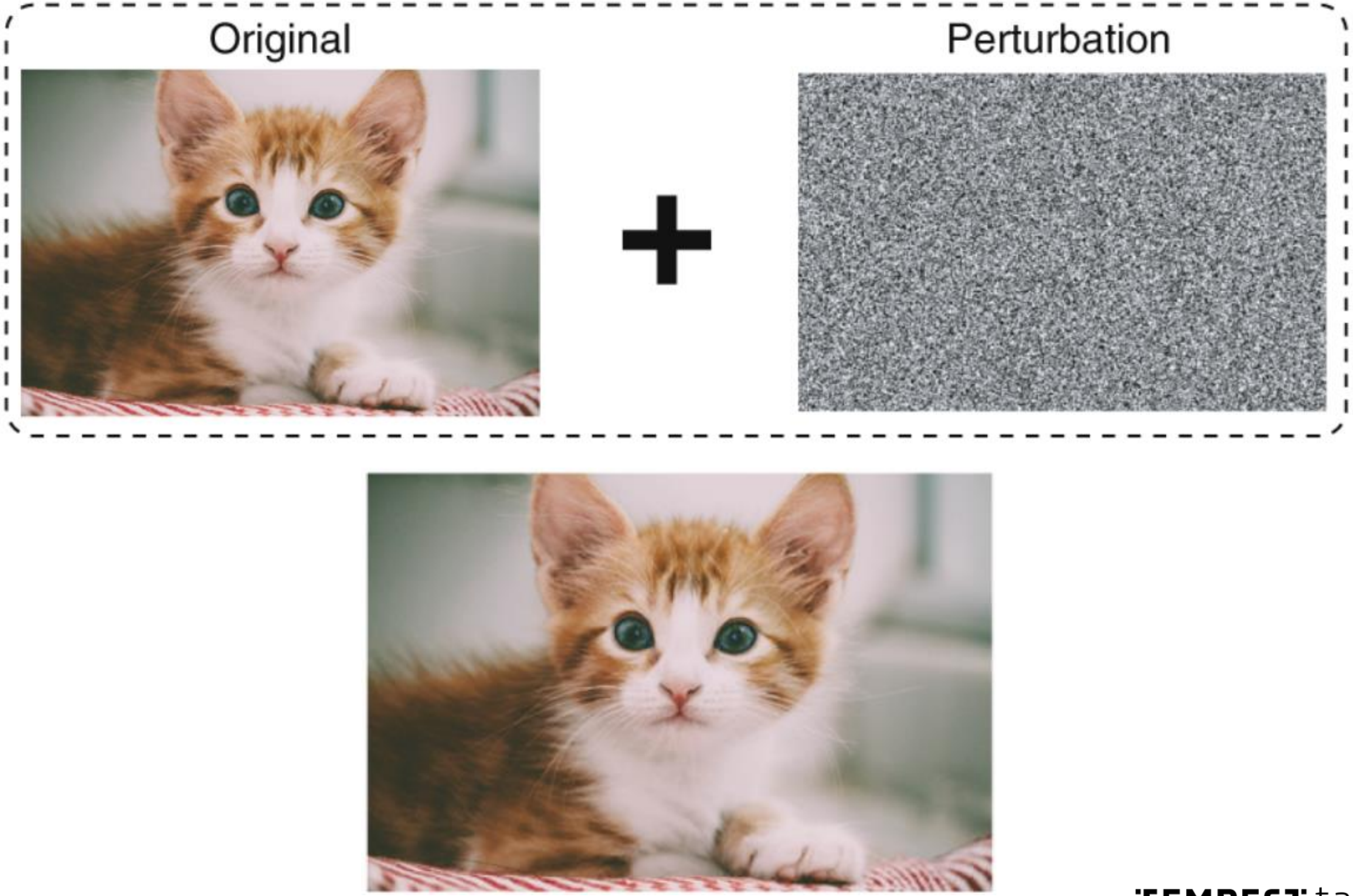
Cat (99.6)

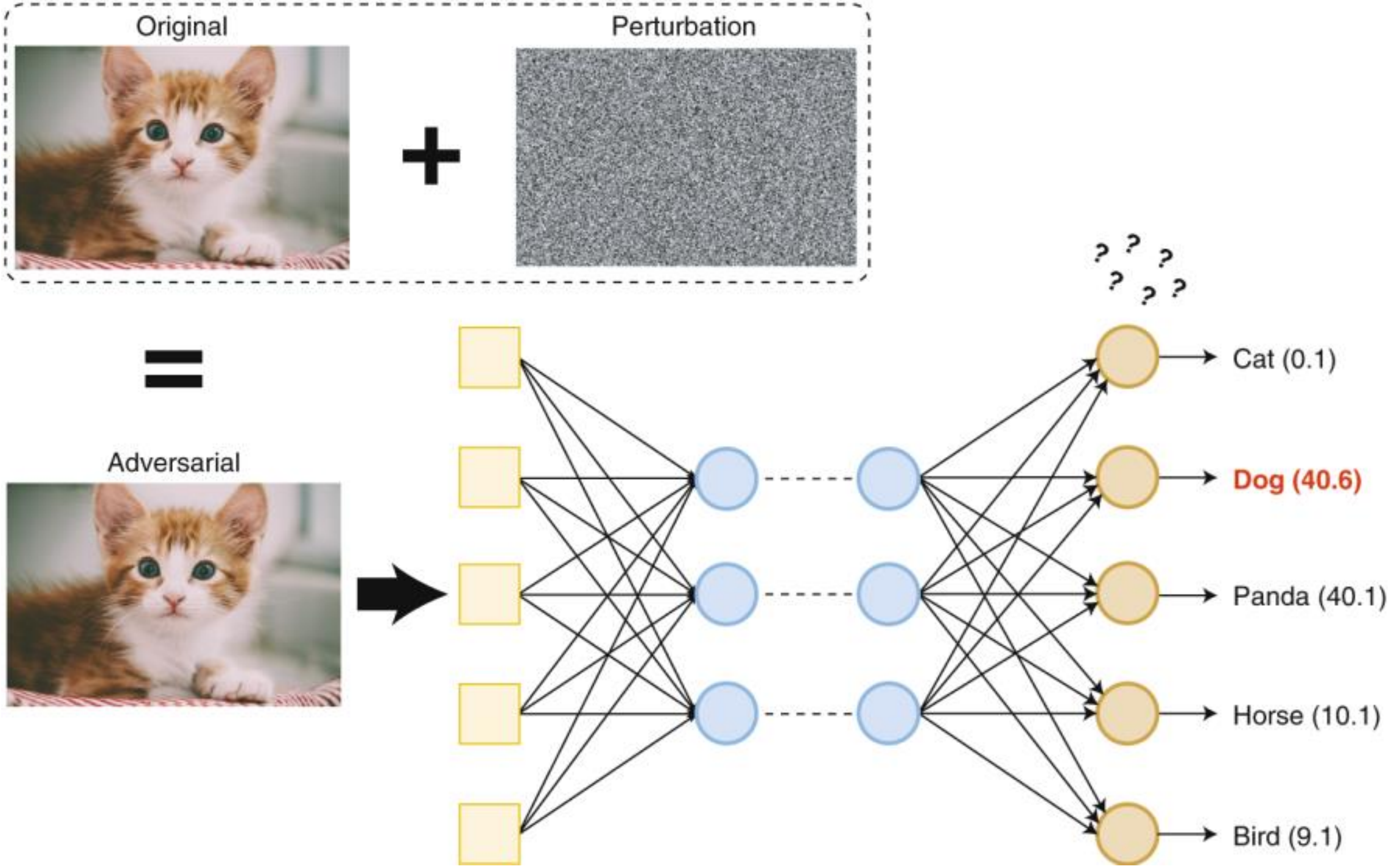
Dog (0.1)

Panda (0.1)

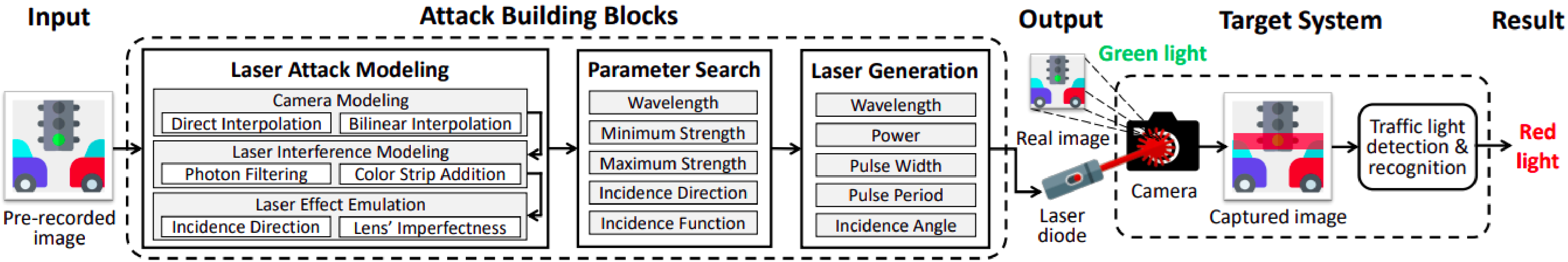
Horse (0.1)

Bird (0.1)


Original
+
Perturbation



TEMPEST talks
2022



Input

Pre-recorded image

Attack Building Blocks

Laser Attack Modeling

- Camera Modeling
  - Direct Interpolation
  - Bilinear Interpolation
- Laser Interference Modeling
  - Photon Filtering
  - Color Strip Addition
- Laser Effect Emulation
  - Incidence Direction
  - Lens' Imperfectness

Parameter Search

- Wavelength
- Minimum Strength
- Maximum Strength
- Incidence Direction
- Incidence Function

Laser Generation

- Wavelength
- Power
- Pulse Width
- Pulse Period
- Incidence Angle

Output

Green light

Real image

Laser diode

Target System

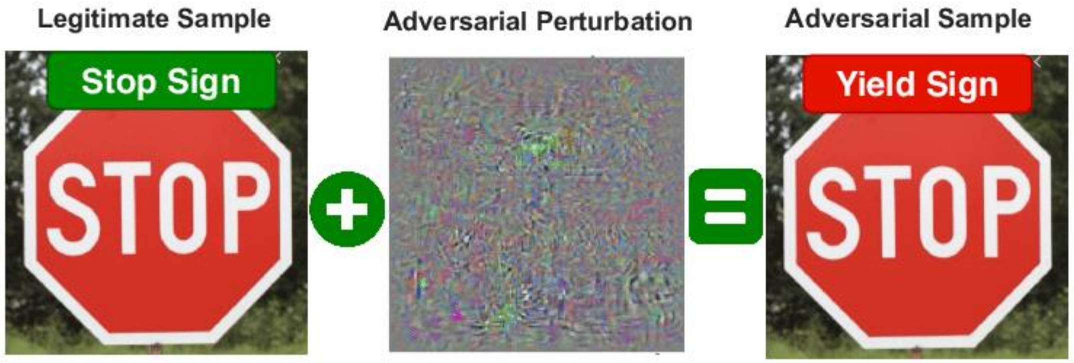
Camera

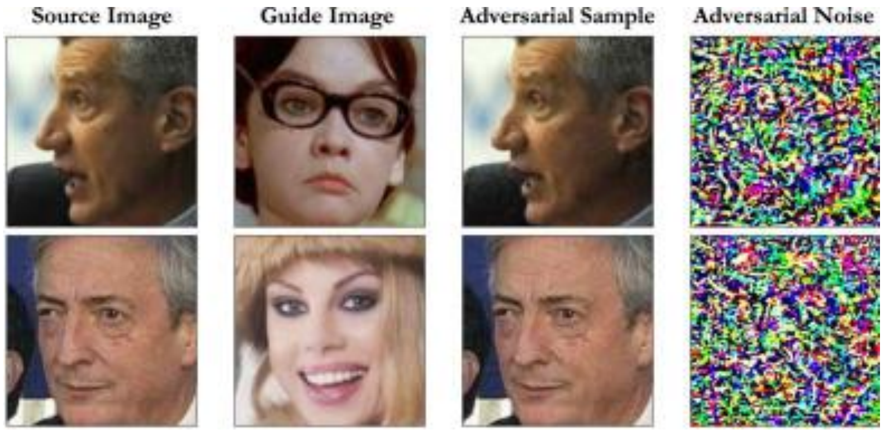
Captured image

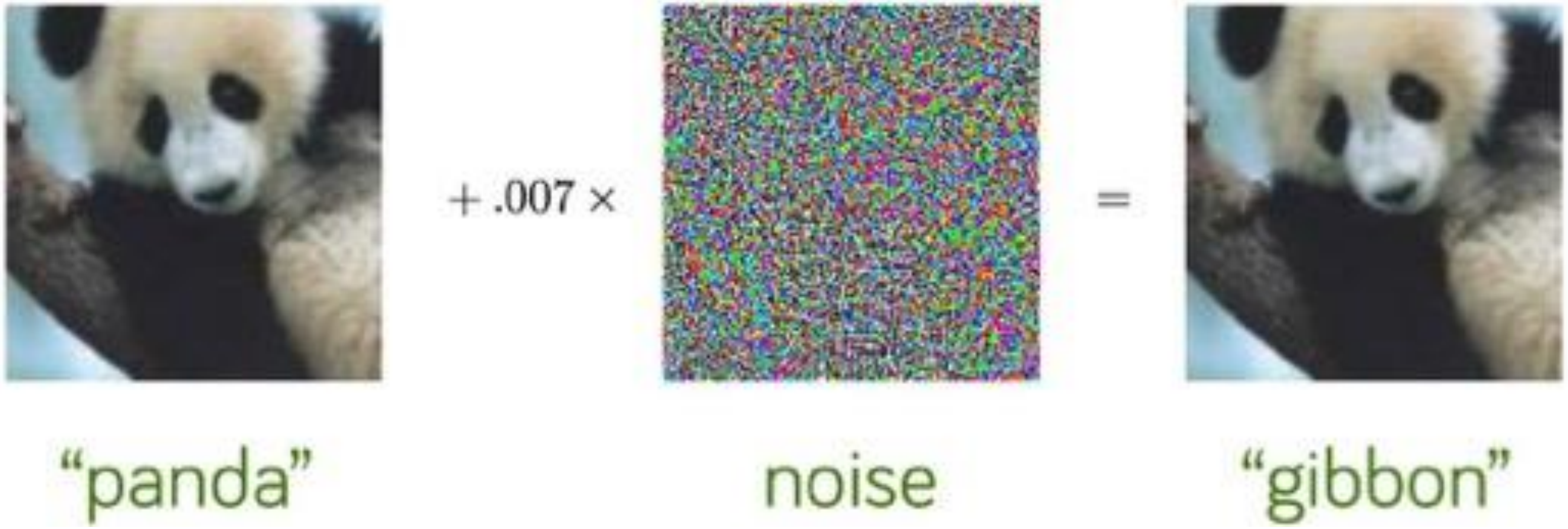
Traffic light detection & recognition

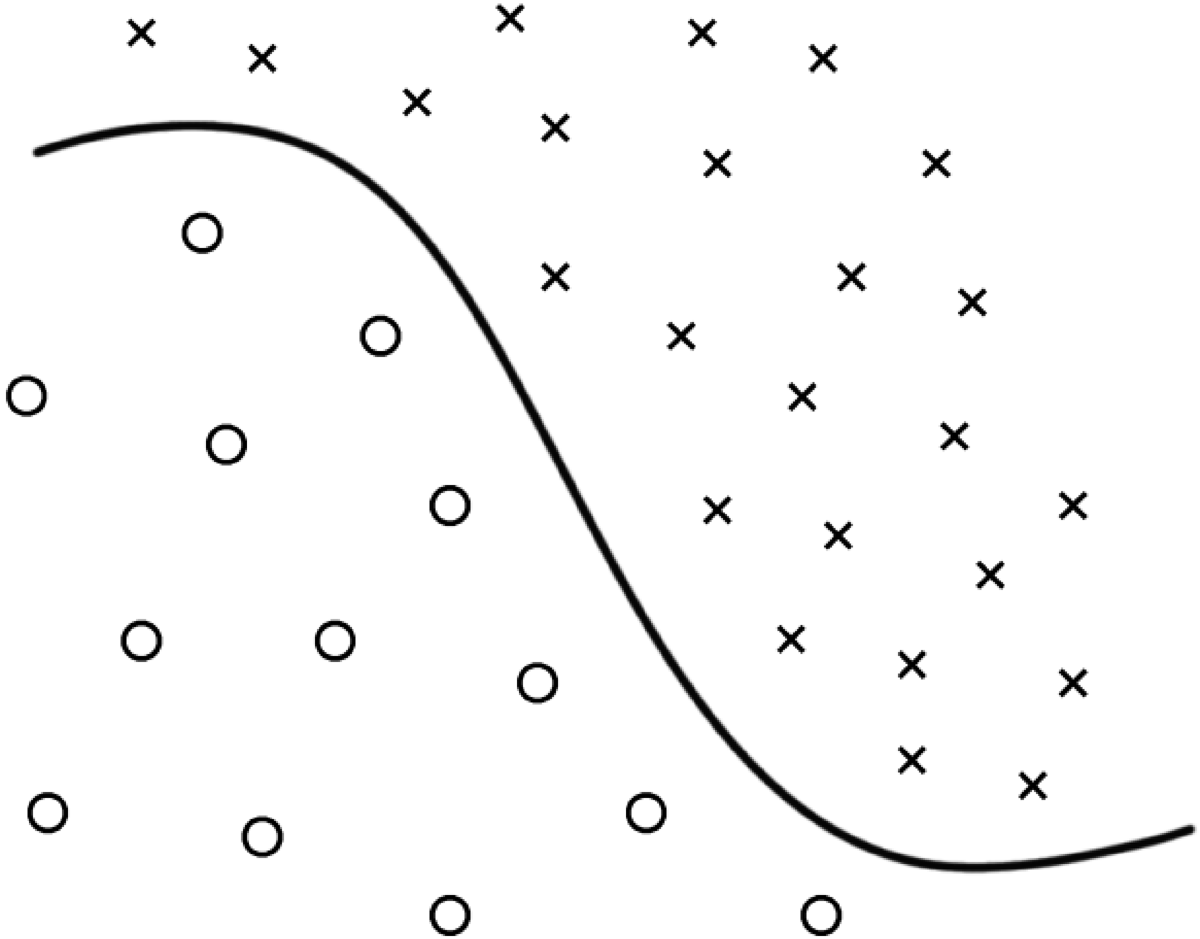
Result

Red light

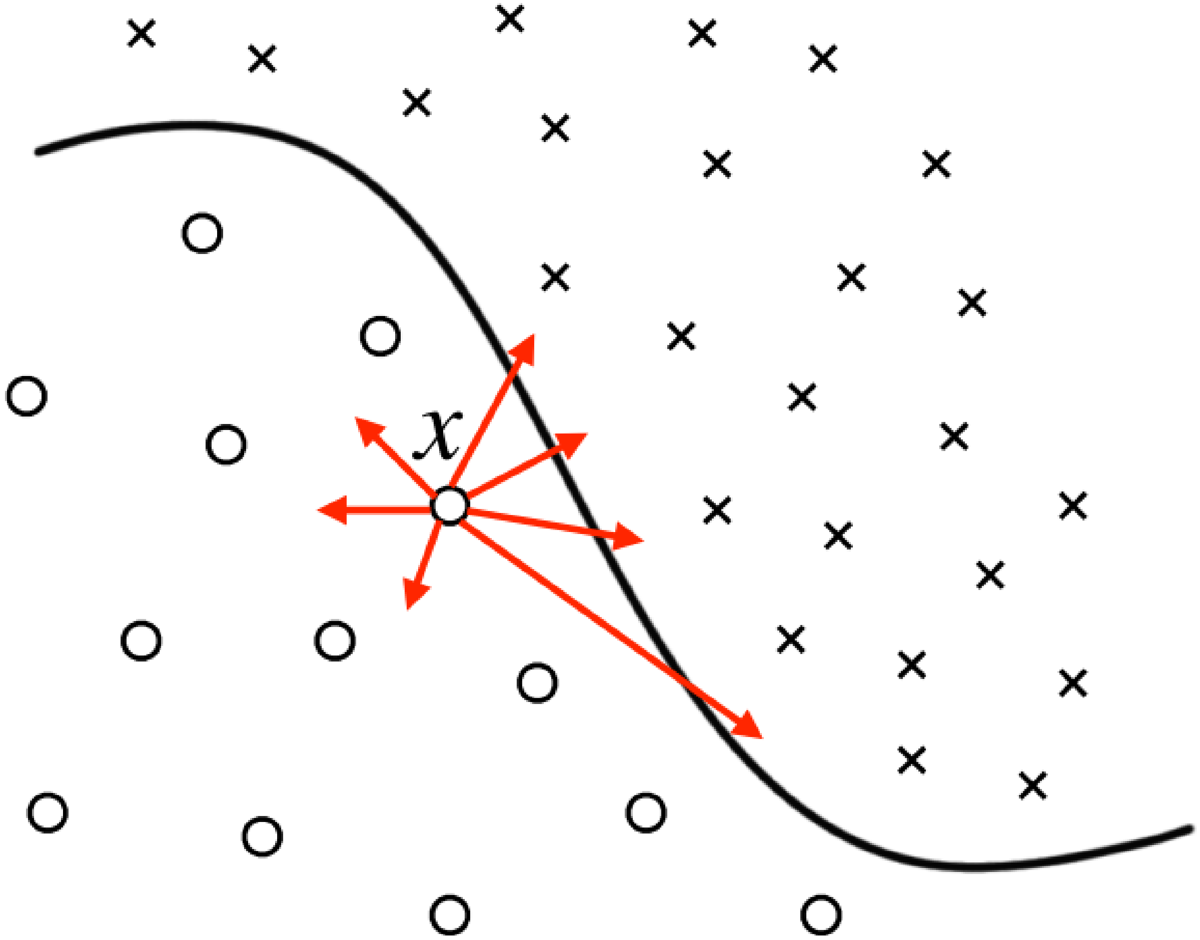


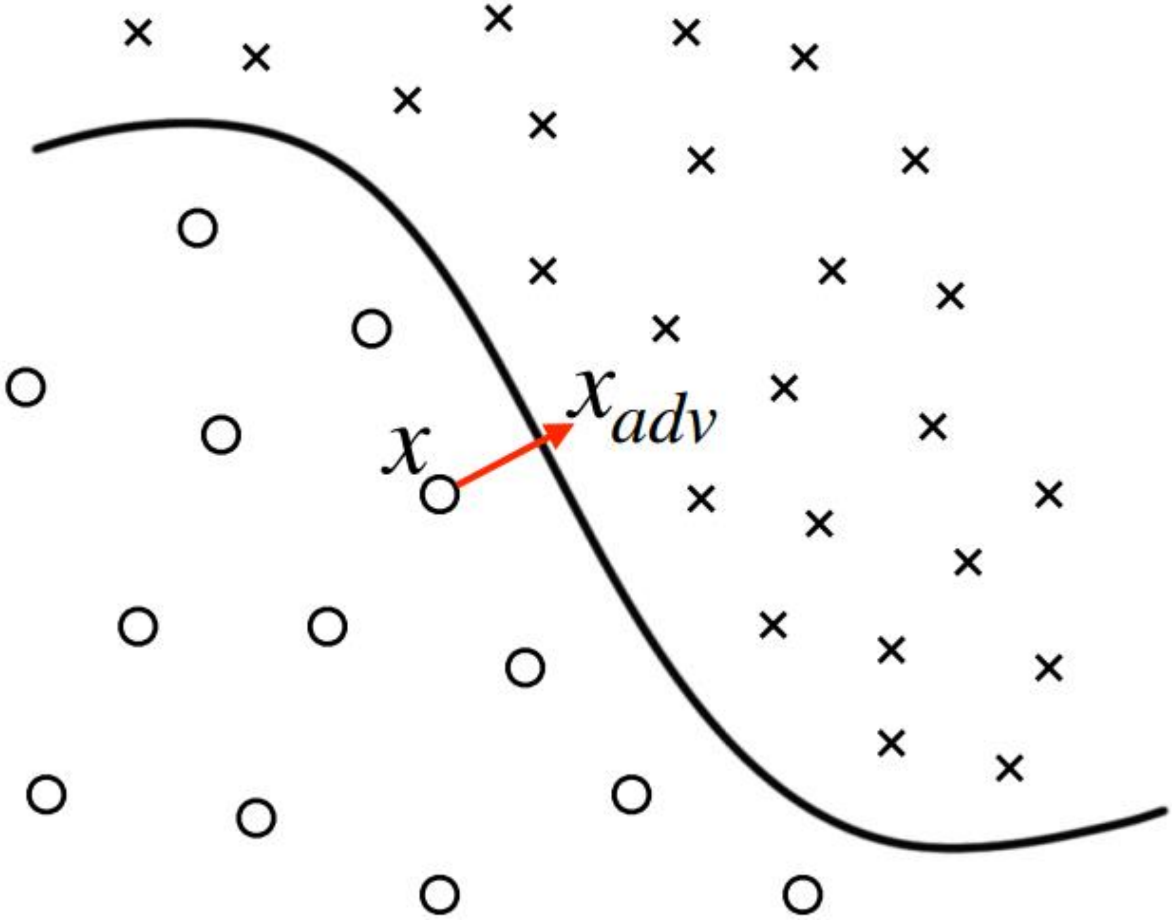



+ .007 ×
=
“panda”
noise
“gibbon”



TEMPEST talks 2022

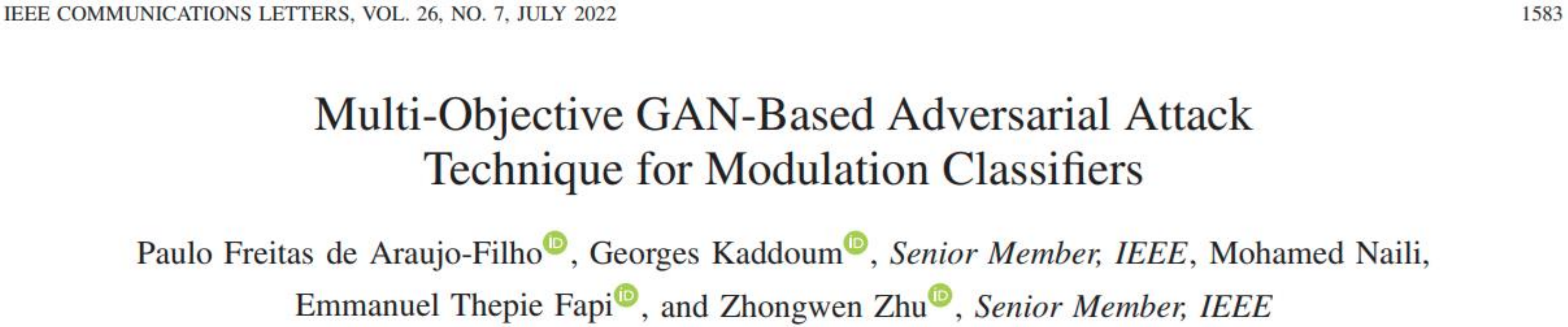

x



$$x_{adv} = x + \delta$$

$$\min||x_{adv} - x|| < \rho$$

$$f(x_{adv}) \neq f(x)$$



IEEE COMMUNICATIONS LETTERS, VOL. 26, NO. 7, JULY 2022 1583

Multi-Objective GAN-Based Adversarial Attack Technique for Modulation Classifiers

Paulo Freitas de Araujo-Filho, Georges Kaddoum, *Senior Member, IEEE*, Mohamed Naili, Emmanuel Thepie Fapi, and Zhongwen Zhu, *Senior Member, IEEE*

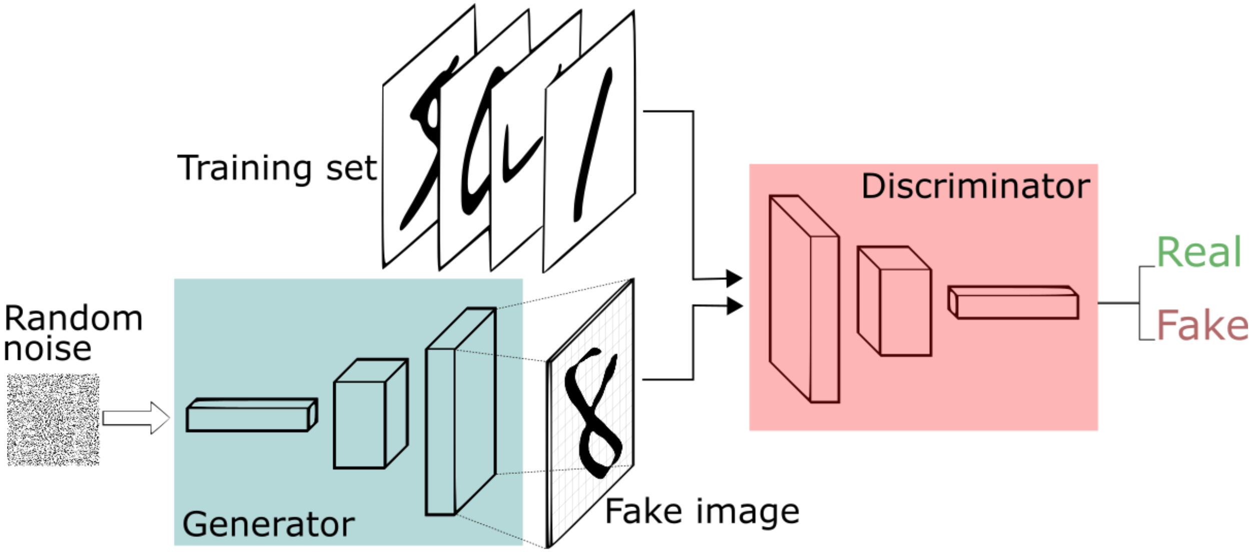
$$f(x_{adv}) \neq f(x)$$

$$\min||x_{adv} - x|| < \rho$$

- Generative Adversarial Networks (GANs)
    - Treina simultaneamente duas redes neurais que competem entre si

- Gerador G
    - Treinado para produzir amostras sintéticas de dados que sejam reconhecidos para reais
    - Aprende a distribuição de probabilidade de dos dados reais
    - Implicitamente modela o sistema

- Discriminador D
    - Treinado para distinguir as amostras reais daquelas produzidas pelo gerador



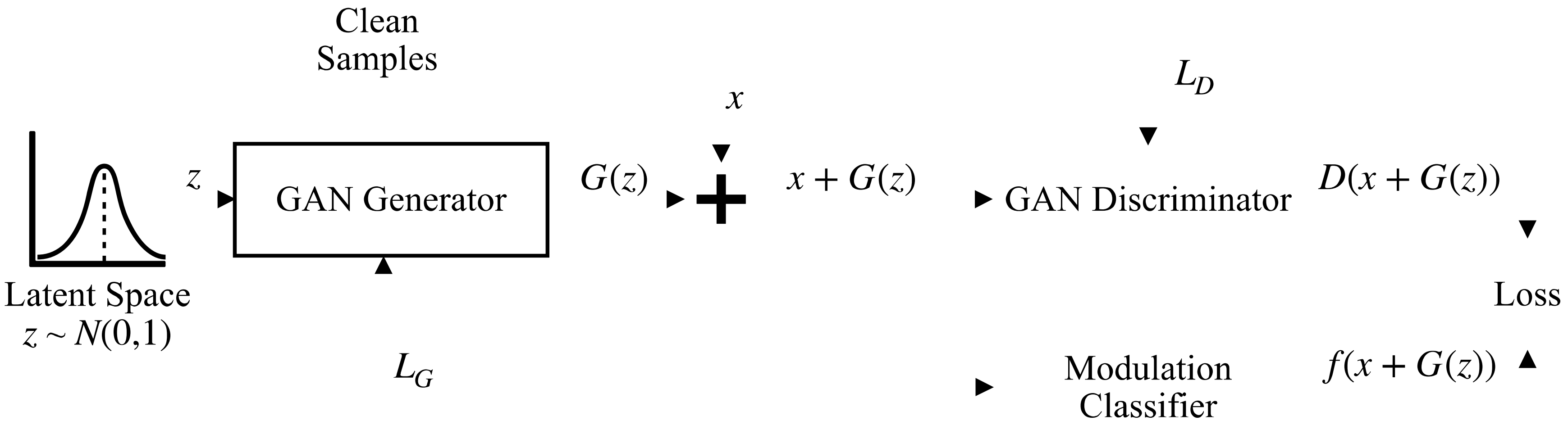
$$L_G = -D(G(z))$$

$$L_D = D\big(G(z)\big) - D(x)$$

- Modificamos a estrutura da GAN para que o gerador produza perturbações adversariais

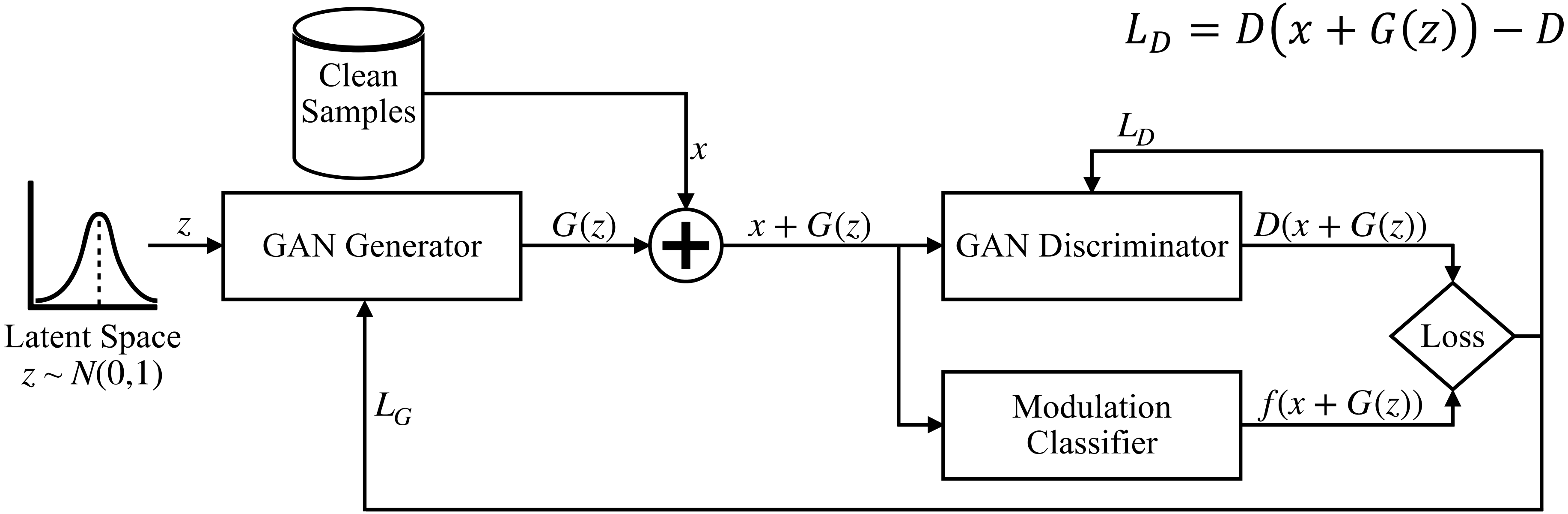
$$\delta = G(z)$$
$$x_{adv} = x + G(z)$$

$$L_G = -D(x + G(z))$$
$$L_D = D\big(x + G(z)\big) - D(x)$$



- Modificamos a estrutura da GAN para que o gerador produza perturbações adversariais

$$L_G = -D(x + G(z))$$
$$L_D = D\big(x + G(z)\big) - D(x)$$



$$L_{G2} = CE(f\big(x + G(z)\big), y) = -\sum_{i=1}^{n} y_i \log(f_i(x + G(z)))$$

- Modificamos a estrutura da GAN para que o gerador produza perturbações adversariais

$$L_{G1} = -D(x + G(z))$$

$$L_{G2} = CE\big(f\big(x + G(z)\big), y\big) = -\sum_{i=1}^{n} y_i \log(f_i(x + G(z)))$$

- Modificamos a estrutura da GAN para que o gerador produza perturbações adversariais

$$L_{G1} = -D(x + G(z))$$

$$L_{G2} = CE\big(f\big(x + G(z)\big), y\big) = -\sum_{i=1}^{n} y_i \log(f_i(x + G(z)))$$

$$L_G = \alpha L_{G1} + \beta L_{G2}$$

- Multi-Task Loss

$$p(\boldsymbol{y}|\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x})) = Softmax(\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x}))$$

- Multi-Task Loss

$$p(\boldsymbol{y}|\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x})) = Softmax(\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x}))$$

$$\log p(\boldsymbol{y}|\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x})) \propto -\frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{f}^{\boldsymbol{W}}(\mathbf{x})||^2 - \log\sigma$$

- Multi-Task Loss

$$p(\boldsymbol{y}|\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x})) = Softmax(\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x}))$$

$$\log p(\boldsymbol{y}|\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x})) \propto -\frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{f}^{\boldsymbol{W}}(\mathbf{x})||^2 - \log\sigma$$

$$\log p(\boldsymbol{y_1}, \boldsymbol{y_2}|\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x})) = p\left(\boldsymbol{y_1}\middle|\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x})\right) \cdot p(\boldsymbol{y_2}|\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x}))$$

- Multi-Task Loss

$$p(\boldsymbol{y}|\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x})) = Softmax(\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x}))$$

$$\log p(\boldsymbol{y}|\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x})) \propto -\frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{f}^{\boldsymbol{W}}(\mathbf{x})||^2 - \log\sigma$$

$$\log p(\boldsymbol{y_1}, \boldsymbol{y_2}|\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x})) = p\left(\boldsymbol{y_1}\middle|\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x})\right) \cdot p(\boldsymbol{y_2}|\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x}))$$

$$\begin{aligned} \mathrm{L}(\mathbf{W}, \sigma_1, \sigma_2) &= -\log p\left(\boldsymbol{y_1}, \boldsymbol{y_2}\middle|\boldsymbol{f}^{\boldsymbol{W}}(\boldsymbol{x})\right) \\ &\propto \frac{1}{2\sigma_1^2}||\boldsymbol{y_1} - \boldsymbol{f}^{\boldsymbol{W}}(\mathbf{x})||^2 + \frac{1}{2\sigma_2^2}||\boldsymbol{y_2} - \boldsymbol{f}^{\boldsymbol{W}}(\mathbf{x})||^2 + \log\sigma_1\sigma_2 \\ &= \frac{1}{2\sigma_1^2}L_1(\boldsymbol{W}) + \frac{1}{2\sigma_2^2}L_2(\boldsymbol{W}) + \log\sigma_1\sigma_2 \end{aligned}$$

- Multi-Task Loss

$$L_{G1} = -D(x + G(z))$$

$$L_{G2} = CE\left(f\left(x + G(z)\right), y\right) = -\sum_{i=1}^{n} y_i \log(f_i(x + G(z)))$$

$$L_G = \frac{1}{2\sigma_1^2} L_{G1} + \frac{1}{2\sigma_2^2} L_{G2} + \log(\sigma_1 \sigma_2)$$

- Multi-Task Loss

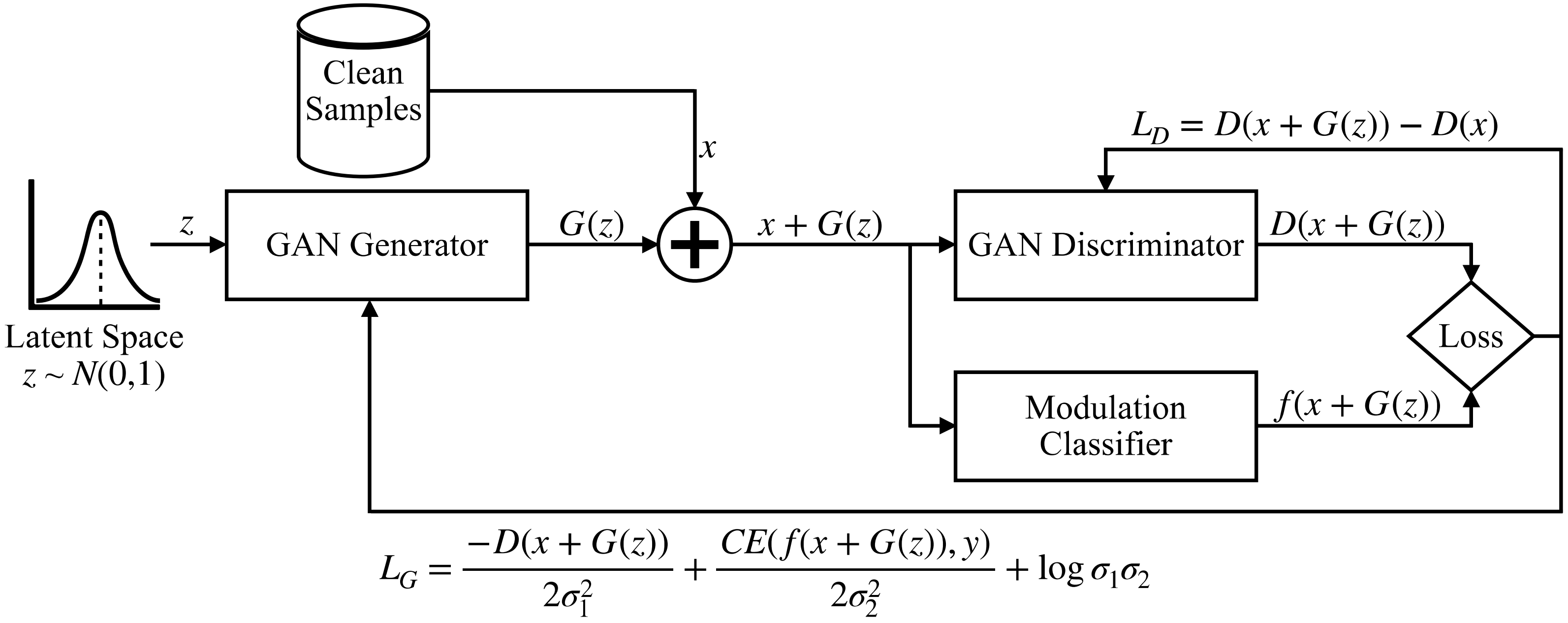
$$L_{G1} = -D(x + G(z))$$

$$L_{G2} = CE(f\big(x + G(z)\big), y) = -\sum_{i=1}^{n} y_i \log(f_i(x + G(z)))$$

$$L_G = \frac{1}{2\sigma_1^2} L_{G1} + \frac{1}{2\sigma_2^2} L_{G2} + \log(\sigma_1 \sigma_2)$$

$$L_G = -\frac{D\big(x+G(z)\big)}{2\sigma_1^2} + \frac{CE\big(f\big(x+G(z)\big),y\big)}{2\sigma_2^2} + \log(\sigma_1 \sigma_2)$$

- Modificamos a estrutura da GAN e a combinamos com a Multi-Task Loss

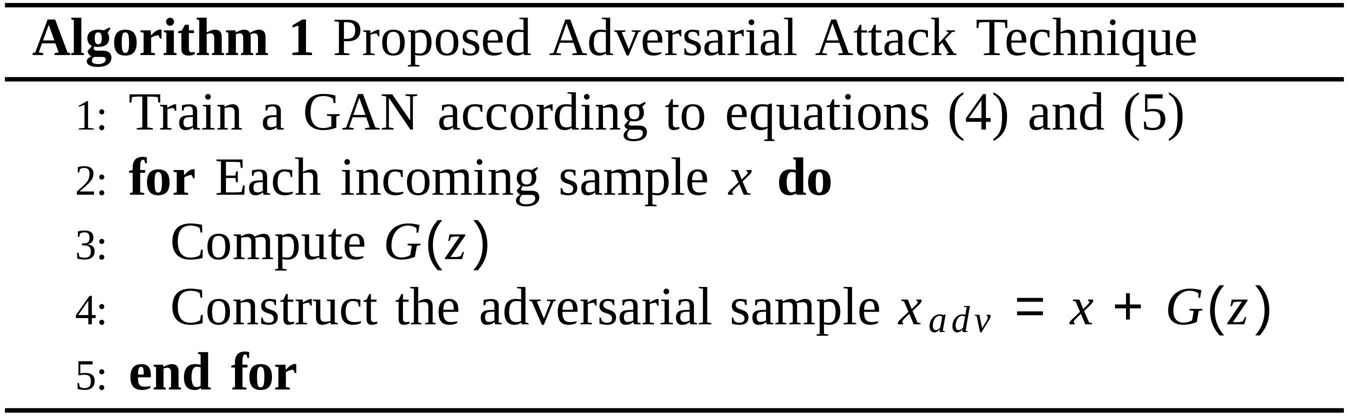


$$L_D = D\big(x + G(z)\big) - D(x)$$

$$L_G = -\frac{D\big(x+G(z)\big)}{2\sigma_1^2} + \frac{CE\big(f\big(x+G(z)\big),y\big)}{2\sigma_2^2} + \log(\sigma_1\sigma_2)$$

TEMPEST talks 2022

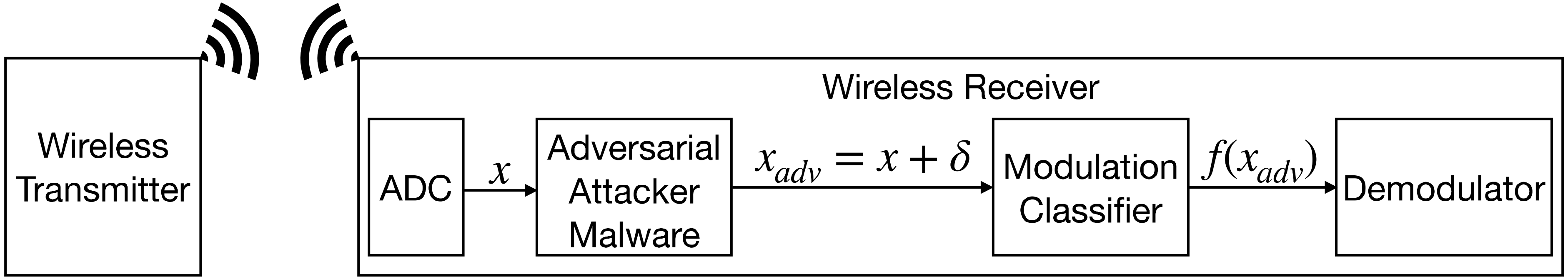
- Multi-Objective GAN-Based Adversarial Attack

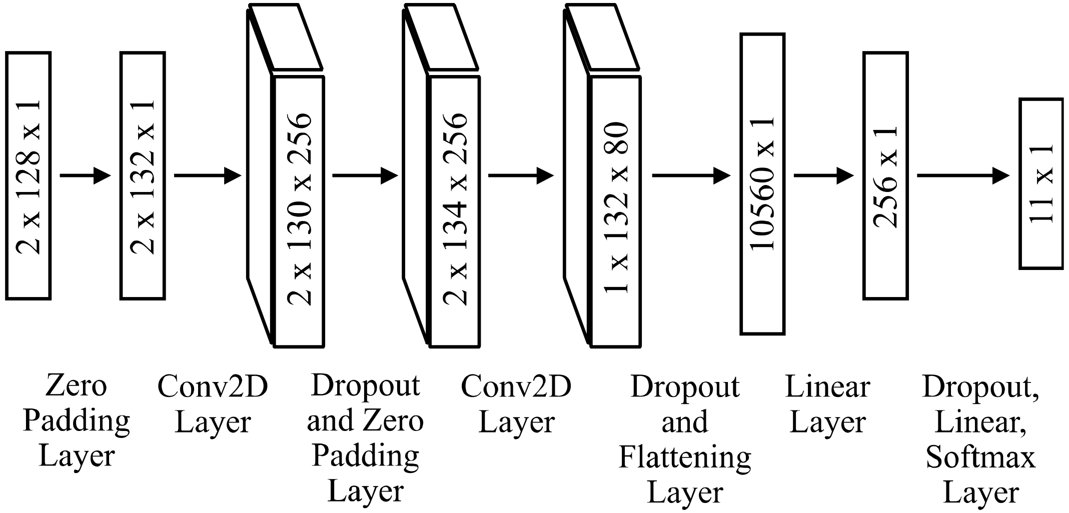


Algorithm 1 Proposed Adversarial Attack Technique

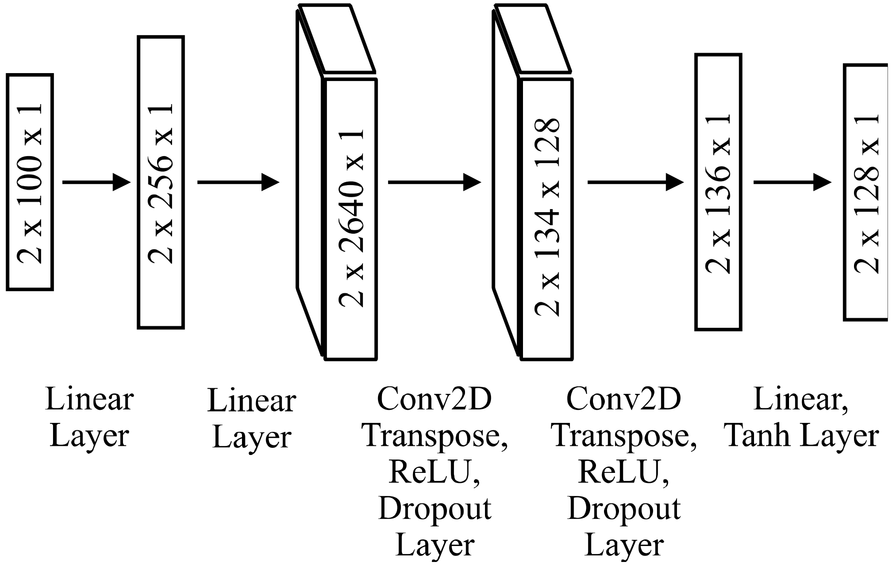
```
Train a GAN according to equations (4) and (5)
for Each incoming sample $x$ do
    Compute $G(z)$
    Construct the adversarial sample $x_{adv} = x + G(z)$
end for
```

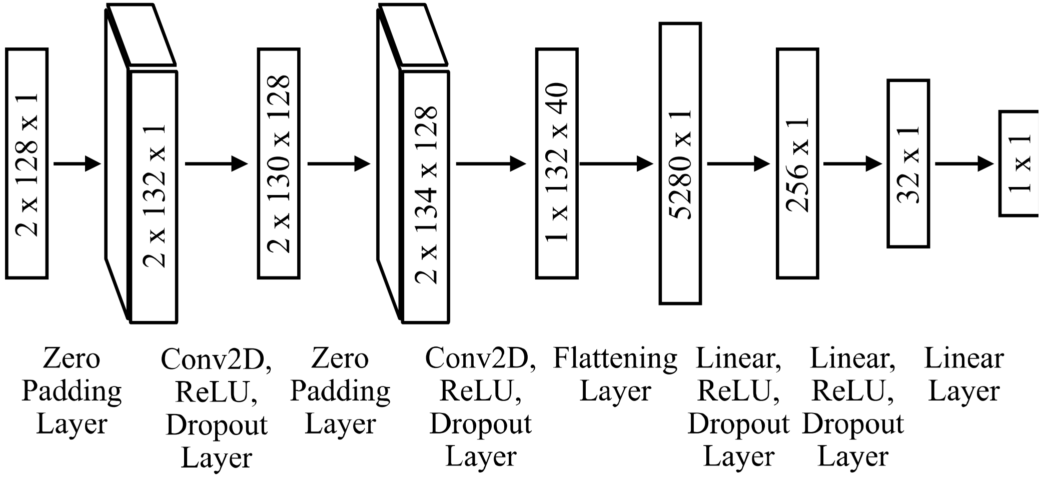
- Multi-Objective GAN-Based Adversarial Attack



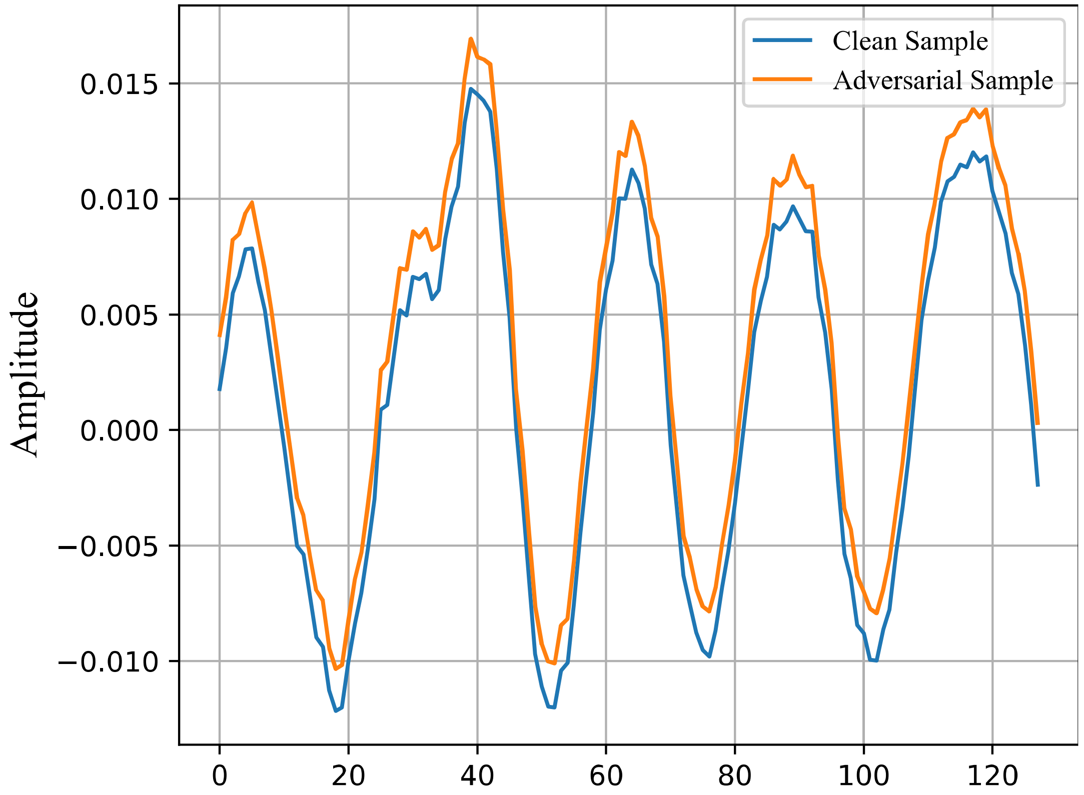


- Multi-Objective GAN-Based Adversarial Attack



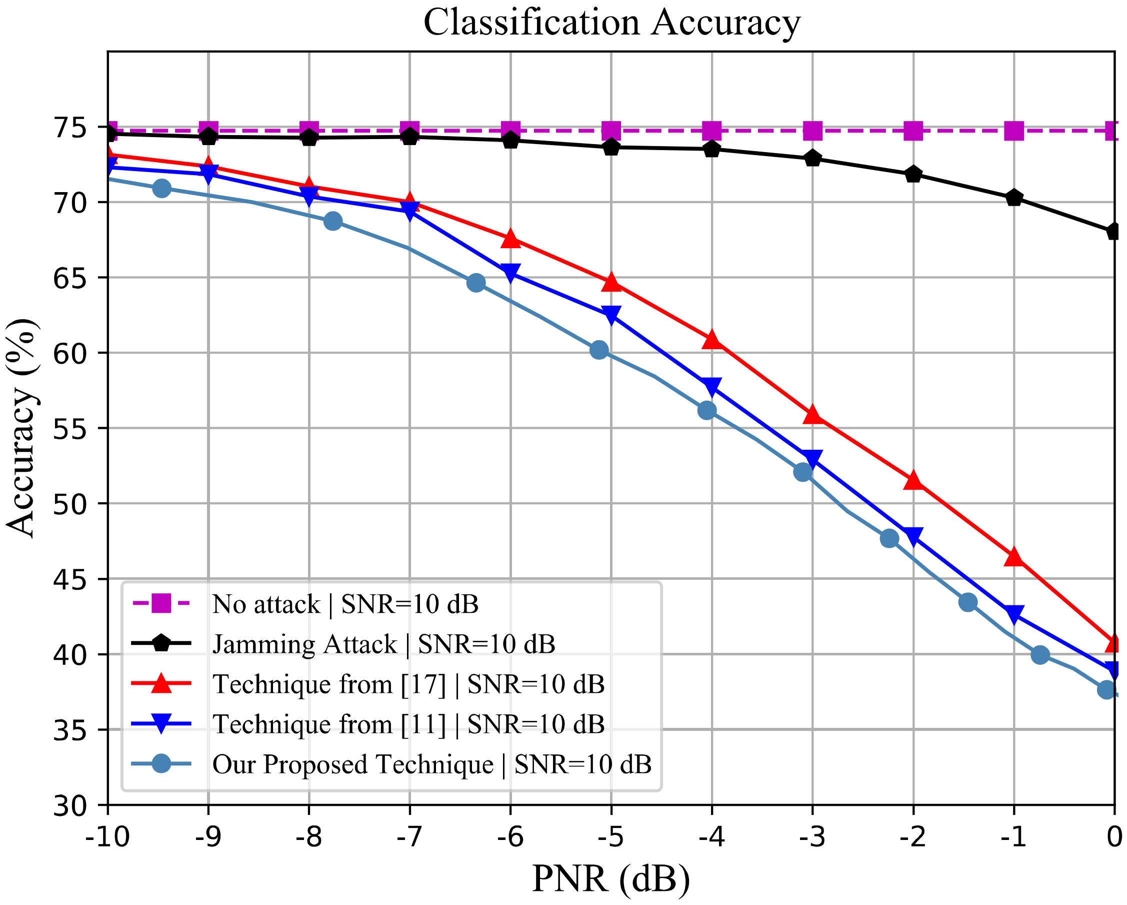


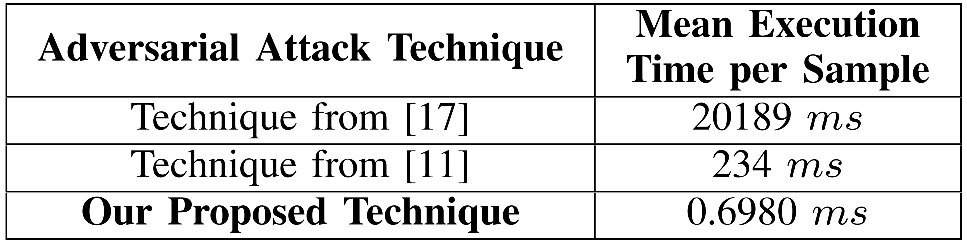
- Multi-Objective GAN-Based Adversarial Attack



TEMPEST talks 2022

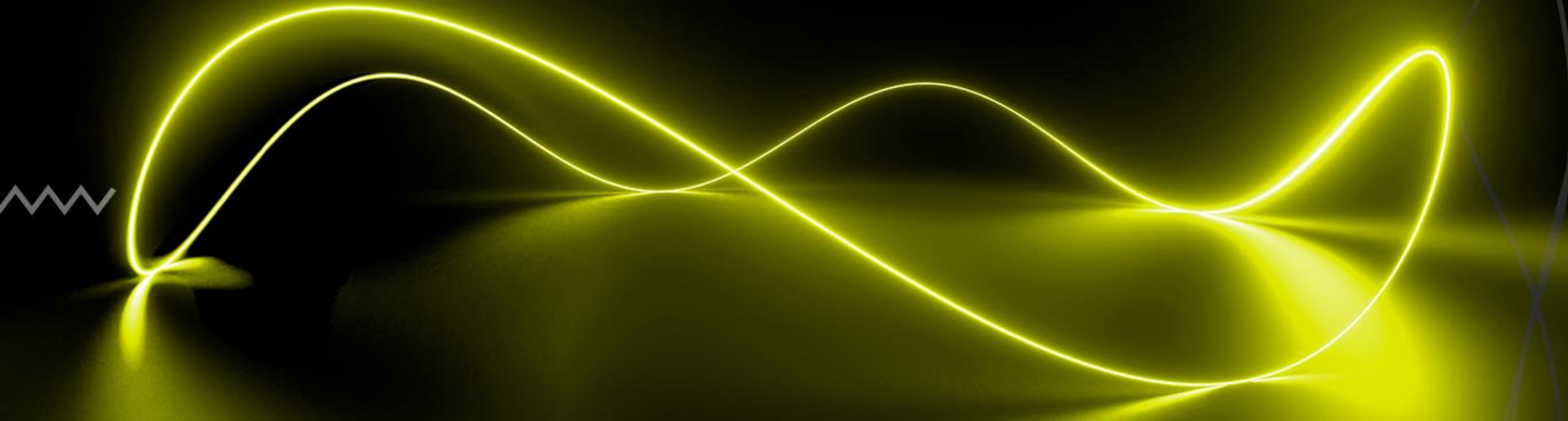
- Multi-Objective GAN-Based Adversarial Attack



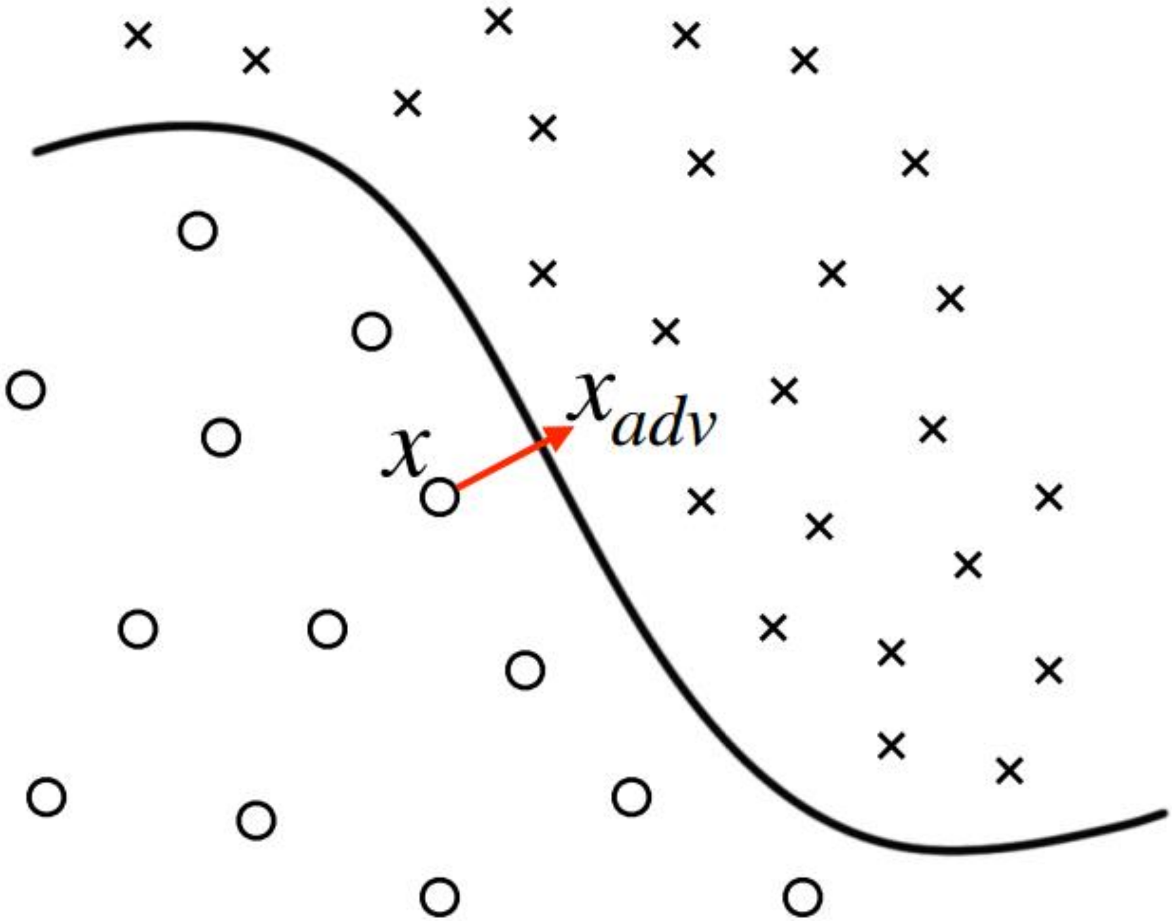


Adversarial Attack Technique	**Mean Execution Time per Sample**
Technique from [17]	20189 ms
Technique from [11]	234 ms
Our Proposed Technique	0.6980 ms

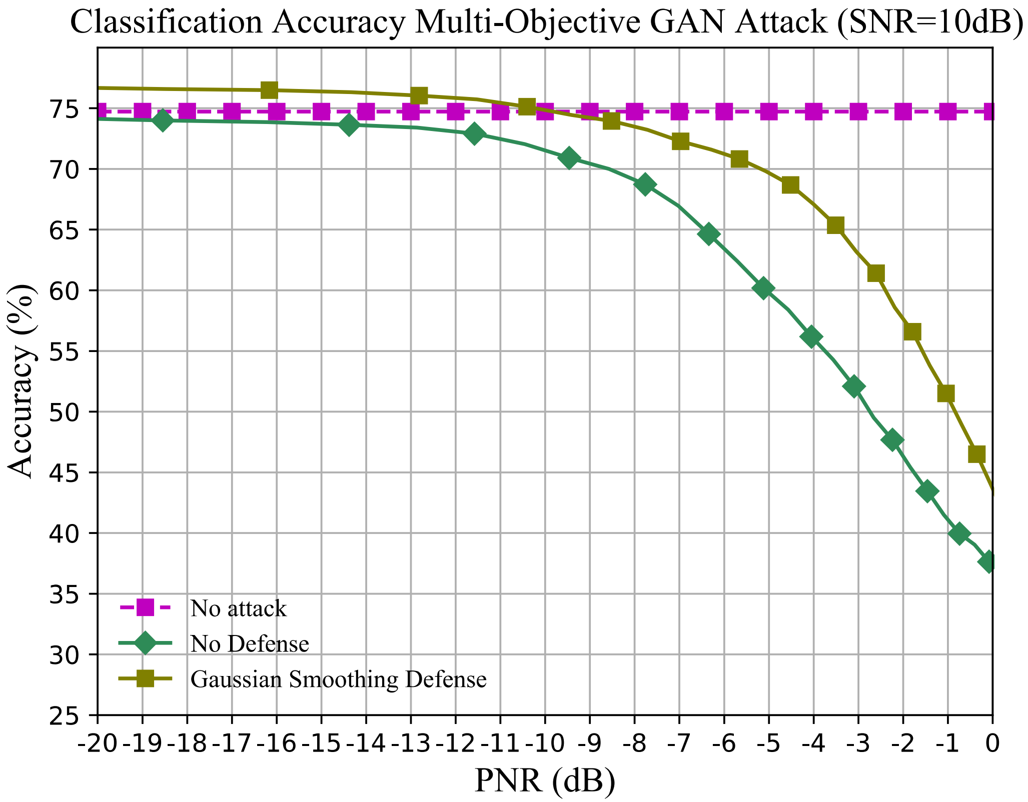
E agora? O que fazer?



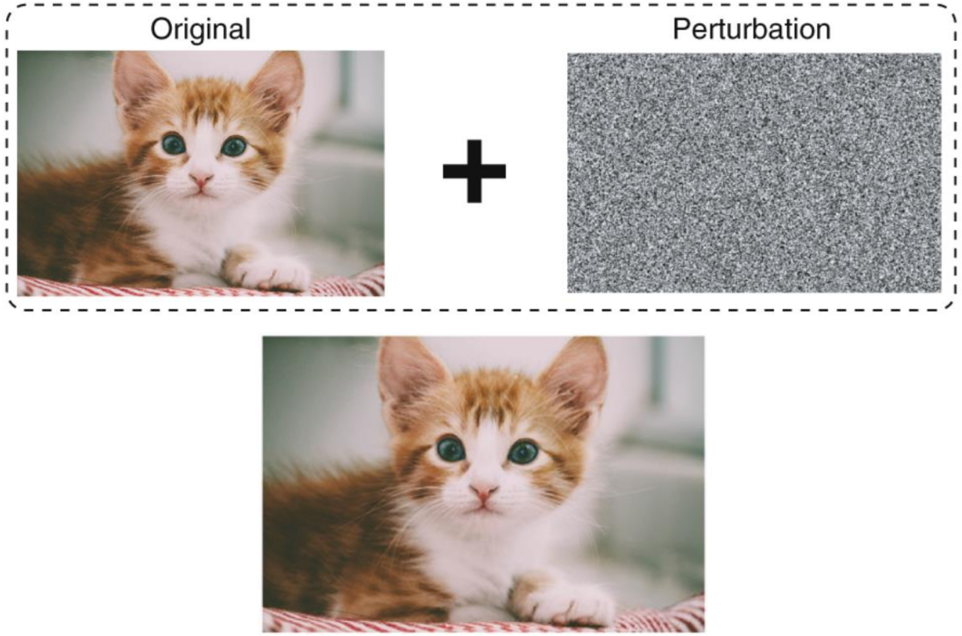
- Diminuir a sensibilidade das fronteiras de decisão

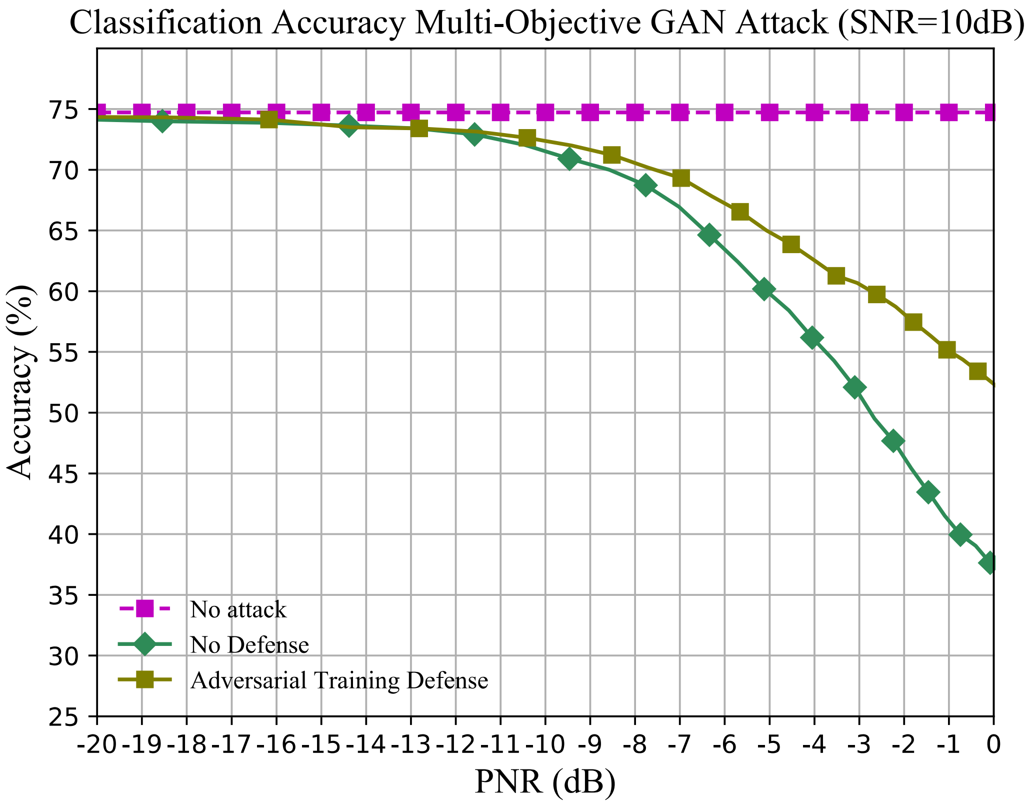


- Diminuir a sensibilidade das fronteiras de decisão

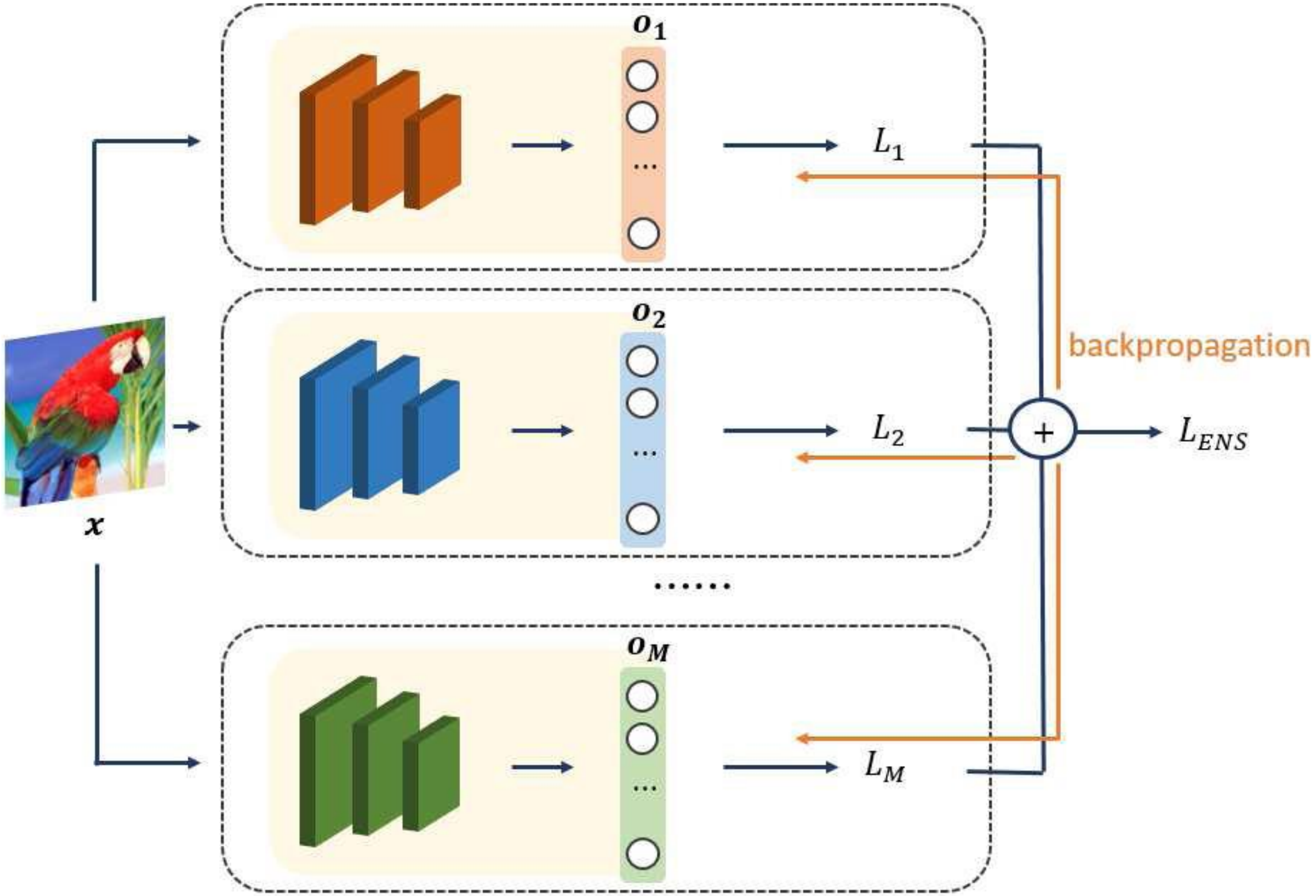


- Diminuir a sensibilidade das fronteiras de decisão



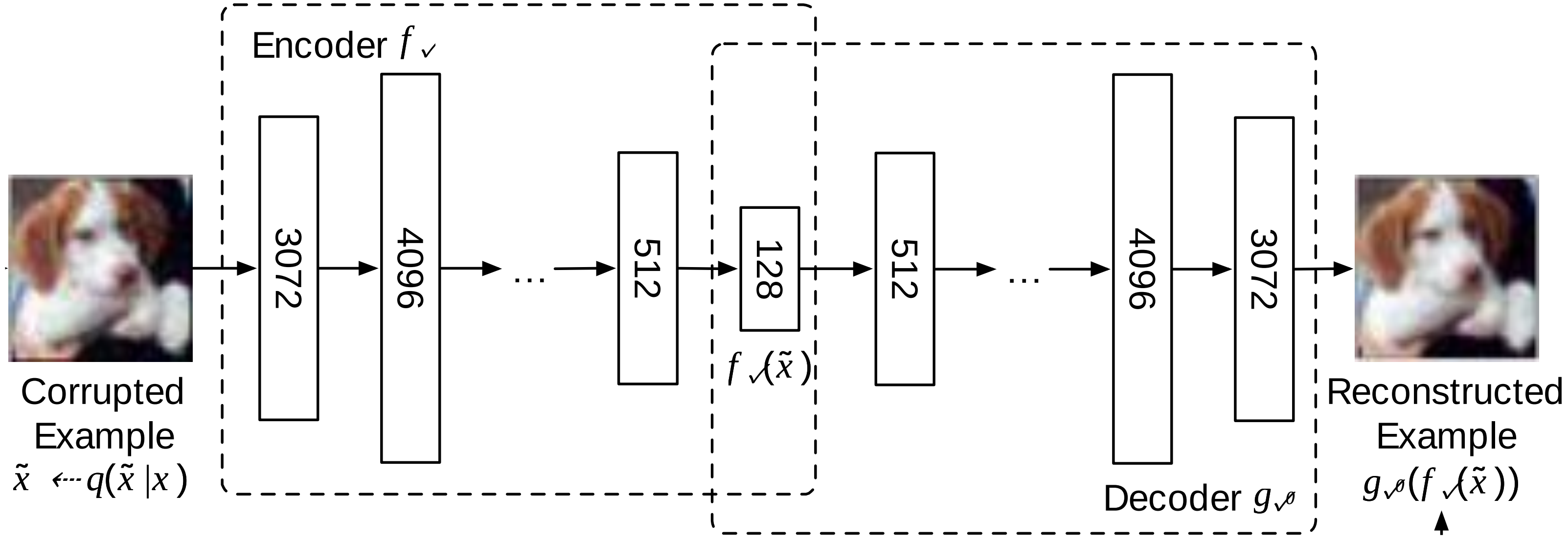


- Combinação de modelos



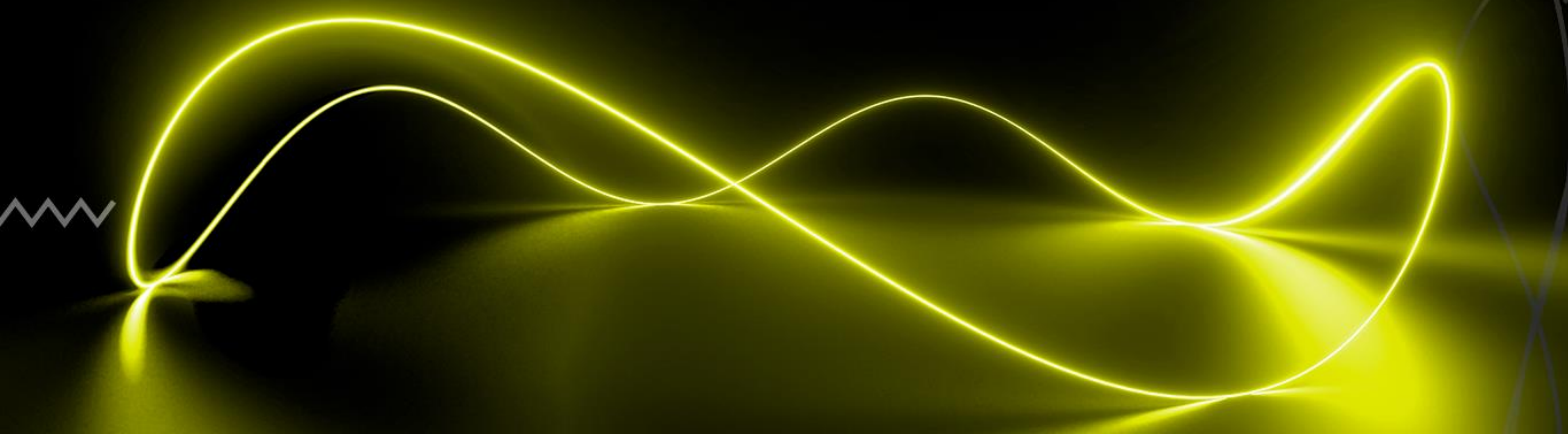
TEMPEST talks 2022

- Remoção de ruído e perturbações adversariais



TEMPEST talks 2022

Com grandes poderes vêm grandes responsabilidades!



Obrigado!

TEMPEST talks 2022

P. Freitas de Araujo-Filho, G. Kaddoum, M. Naili, E. T. Fapi and Z. Zhu, "Multi-Objective GAN-Based Adversarial Attack Technique for Modulation Classifiers," in IEEE Communications Letters, vol. 26, no. 7, pp. 1583-1587, July 2022, doi: 10.1109/LCOMM.2022.3167368.

Paulo Freitas de Araujo Filho
paulo.freitas@tempest.com.br