Erasing the Shadow: Sanitization of Images with Malicious Payloads using Deep Autoencoders

Angelica Liguori[1], Marco Zuppelli[2], Daniela Gallo[1,3], Massimo Guarascio[1], Luca Caviglione[2]

[1] Institute for High Performance Computing and Networking of Italian National Research Council, Rende, Italy
[2] Institute for Applied Mathematics and Information Technologies of Italian National Research Council, Genova, Italy
[3] University of Salento, Lecce, Italy

ISMIS 2024
17-19 JUNE 2024
POITIERS / FUTUROSCOPE, FRANCE

Outline

01 Scenario

02 Problem Definition

03 Methodology

04 Case Study

05 Experiments

06 Conclusions

Scenario

Problem

- Threat actors take advantage of **information-hiding techniques** to, e.g.,:
  - exfiltrate secret information
  - distribute malicious code
- Hiding malicious data in **digital images** through **steganography** is a preferred offensive technique

Goal

- Mitigation of attacks targeting digital images to leak relevant information and hide malicious data via information-hiding techniques

Scenario

Idea

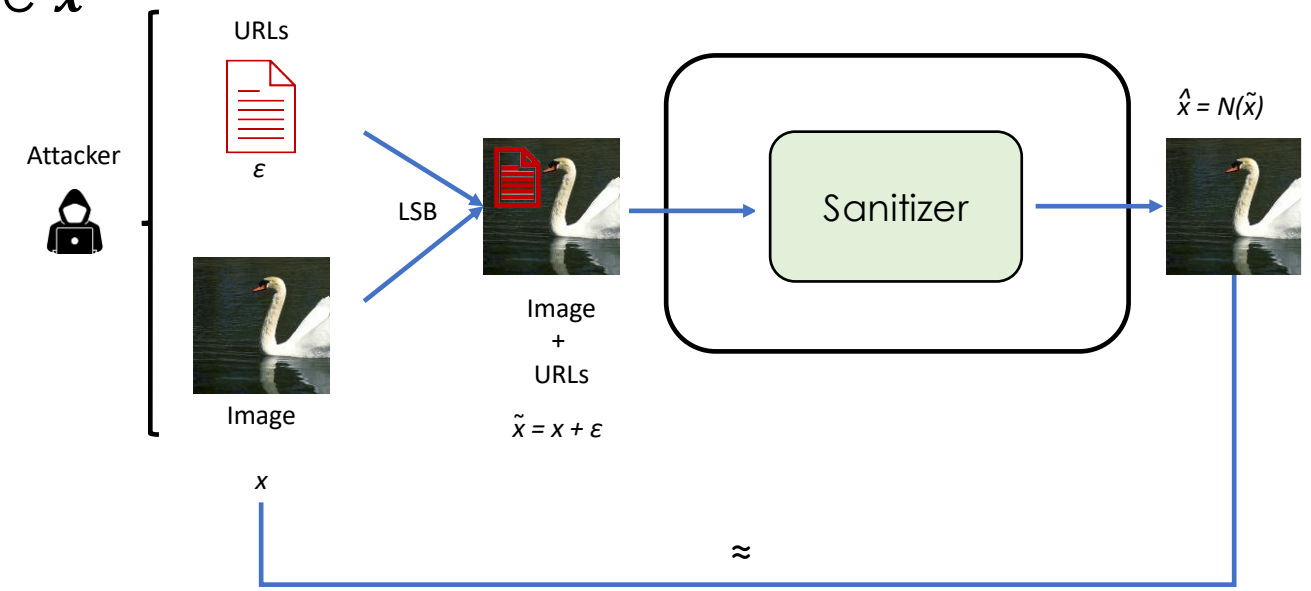
- Exploiting image processing methods to **destroy** hidden content while **preserving** the hosting data

Solution

- Defining a deep learning-based approach to **sanitizing** compromised image, i.e., disrupting the hidden content without altering the overall perceived quality

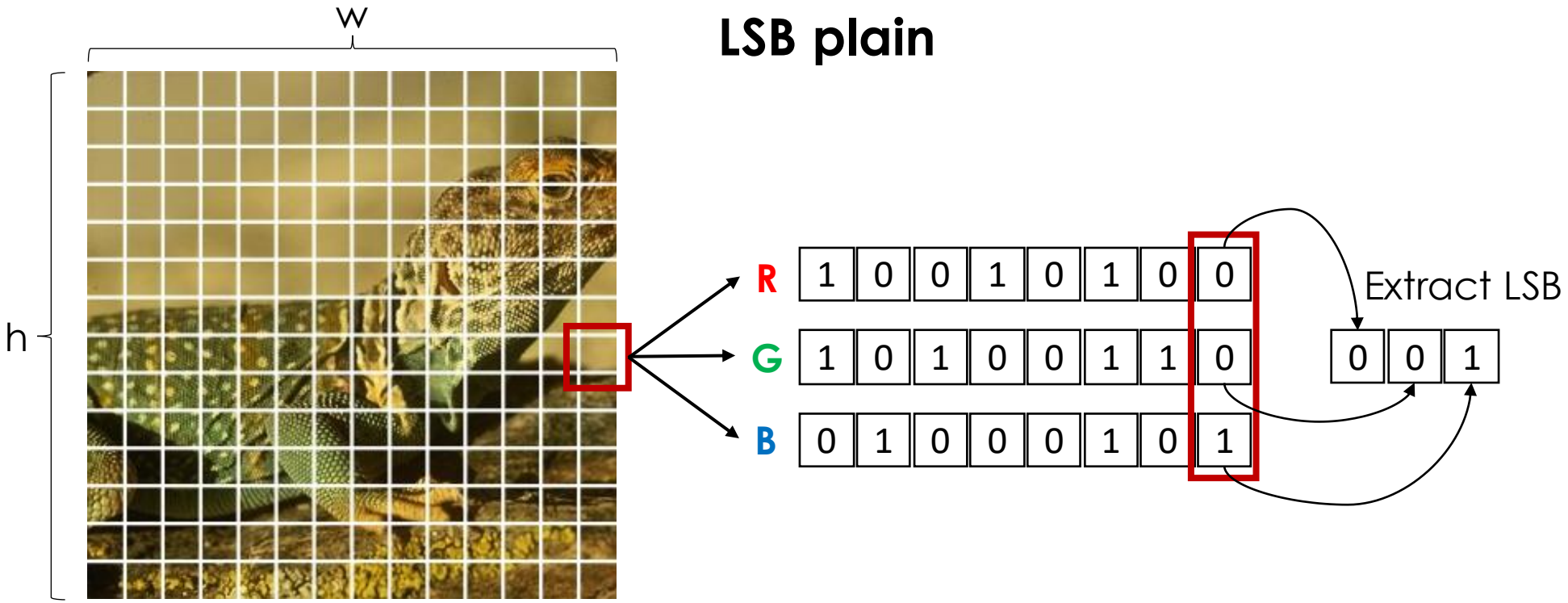
Problem Definition

- Let $\boldsymbol{x}$ be a **legitimate** image where an attacker wants to hide arbitrary (and potentially **malicious**) **information** $\boldsymbol{\epsilon}$ generating the compromised image $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\epsilon}$
- We aim at finding an estimate $\hat{\boldsymbol{x}}$ that is as close as possible to the legitimate (non-compromised) image $\boldsymbol{x}$



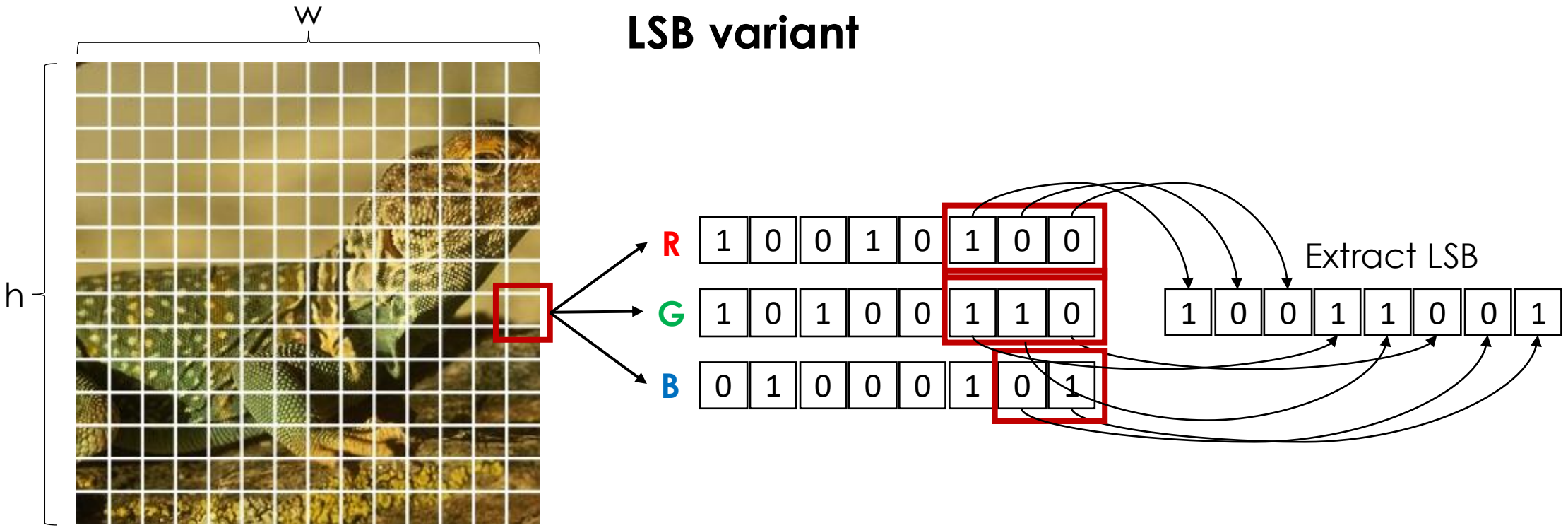
Problem Definition

- The threat actor conceal the information via two steganographic strategies:



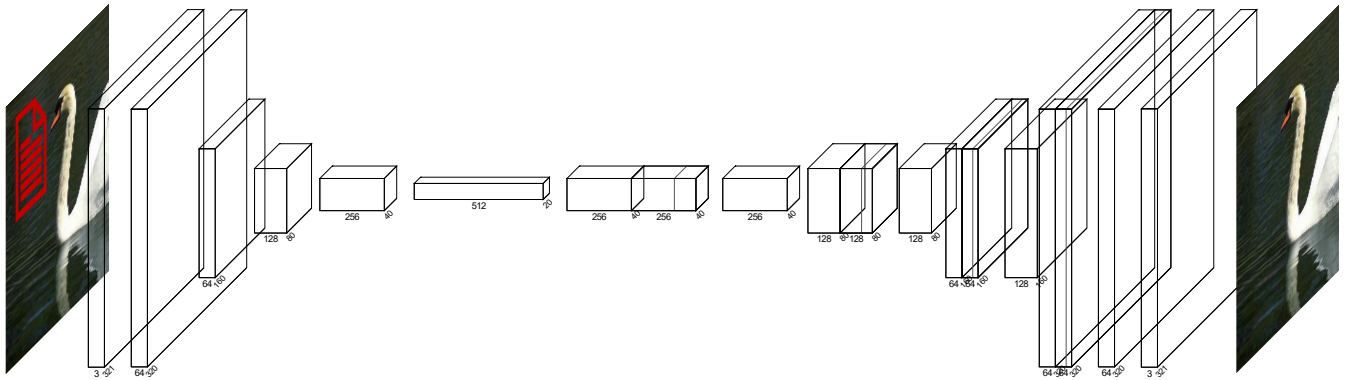
Problem Definition

- The threat actor conceal the information via two steganographic strategies:



Methodology Overview

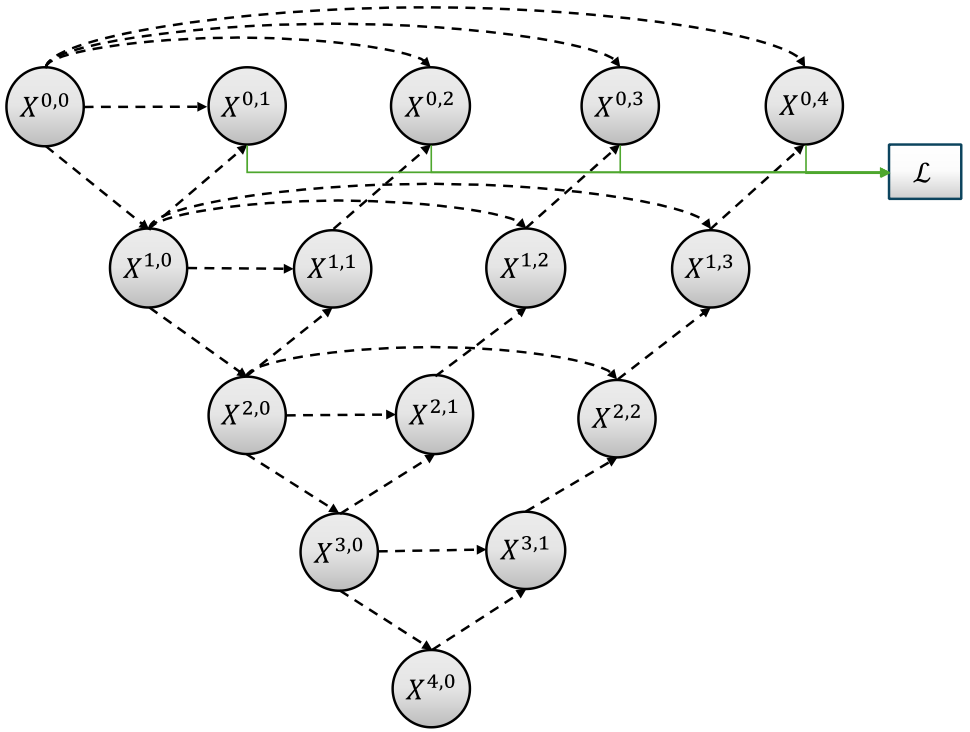
- Devising a neural functional N to map the compromised image $\tilde{\boldsymbol{x}}$ to its corresponding estimate $\hat{\boldsymbol{x}}$
  - N is trained to minimize the dissimilarity between $\hat{\boldsymbol{x}}$ and the legitimate image $\boldsymbol{x}$



  - N acts as **sanitizer** and takes the form of an Autoencoder U-net-like architecture

Methodology Overview

- We use a variant of the U-Net+ architecture [1]



[1] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. IEEE Trans. on Medical Imaging, 39:1856–1867, 2019.

Methodology Overview

- The neural model N is learned on a set of image pair

$$D = \{(\tilde{x}_1, x_1), (\tilde{x}_2, x_2), \dots, (\tilde{x}_n, x_n)\}$$

where $\tilde{x}_i$ is the compromised image and x_i is the legitimate one

- The aim is to produce for each $\tilde{x}_i$ its sanitized counterpart $\hat{x}_i$, i.e., the image in which the hidden information is destroyed

- The learning phase aims at optimizing the network weights by minimizing the reconstruction loss between $\hat{x}_i$ and x_i

Case Study

- To assess the effectiveness of the proposed approach, we created a dataset of image to model the malware cloaking the URLs
- Each image contained a payload composed of 70 randomly picked URLs

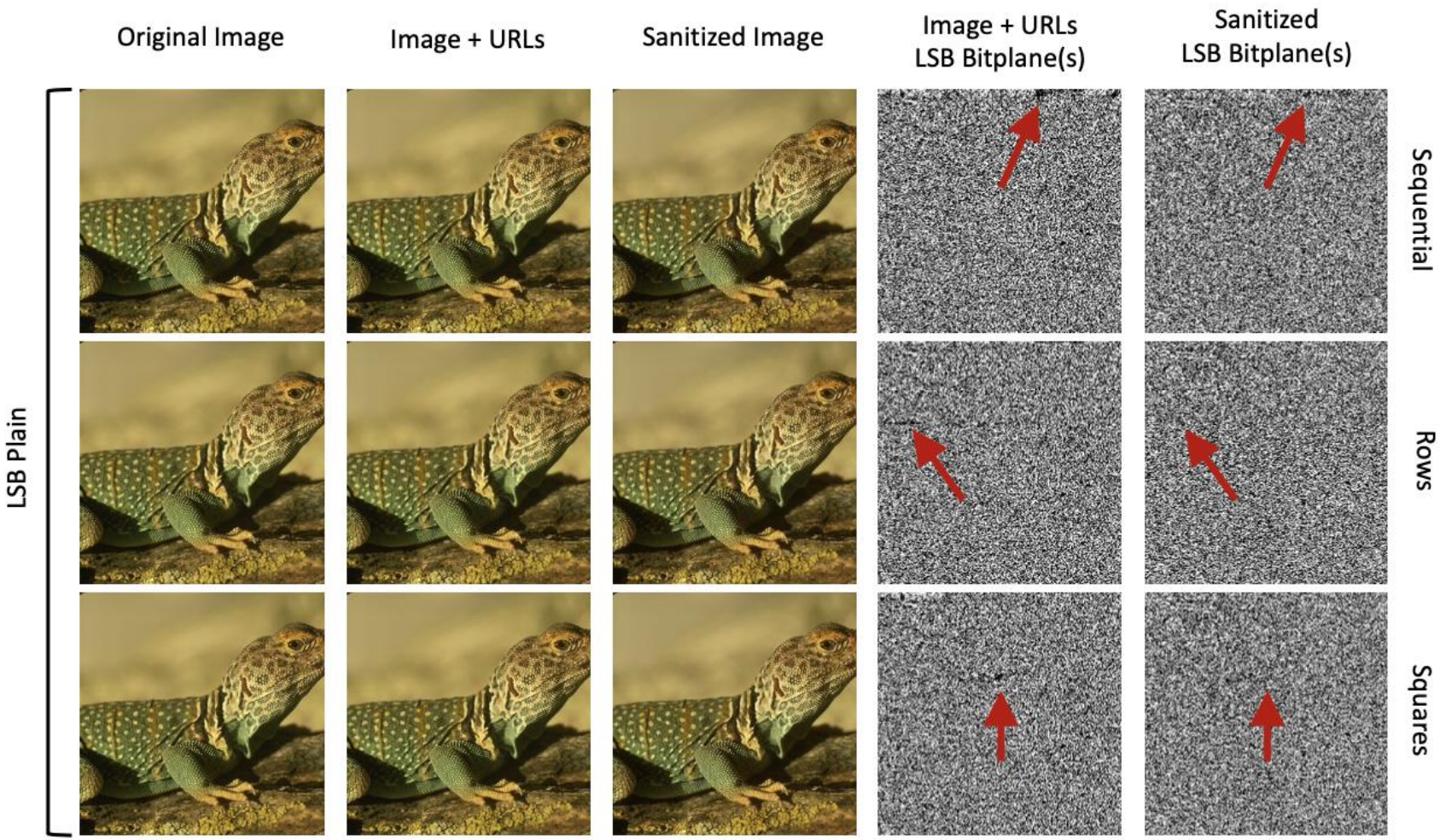
- To hide the content, we use the following patterns:
  - **sequential**
  - **rows**
  - **squares**

Results

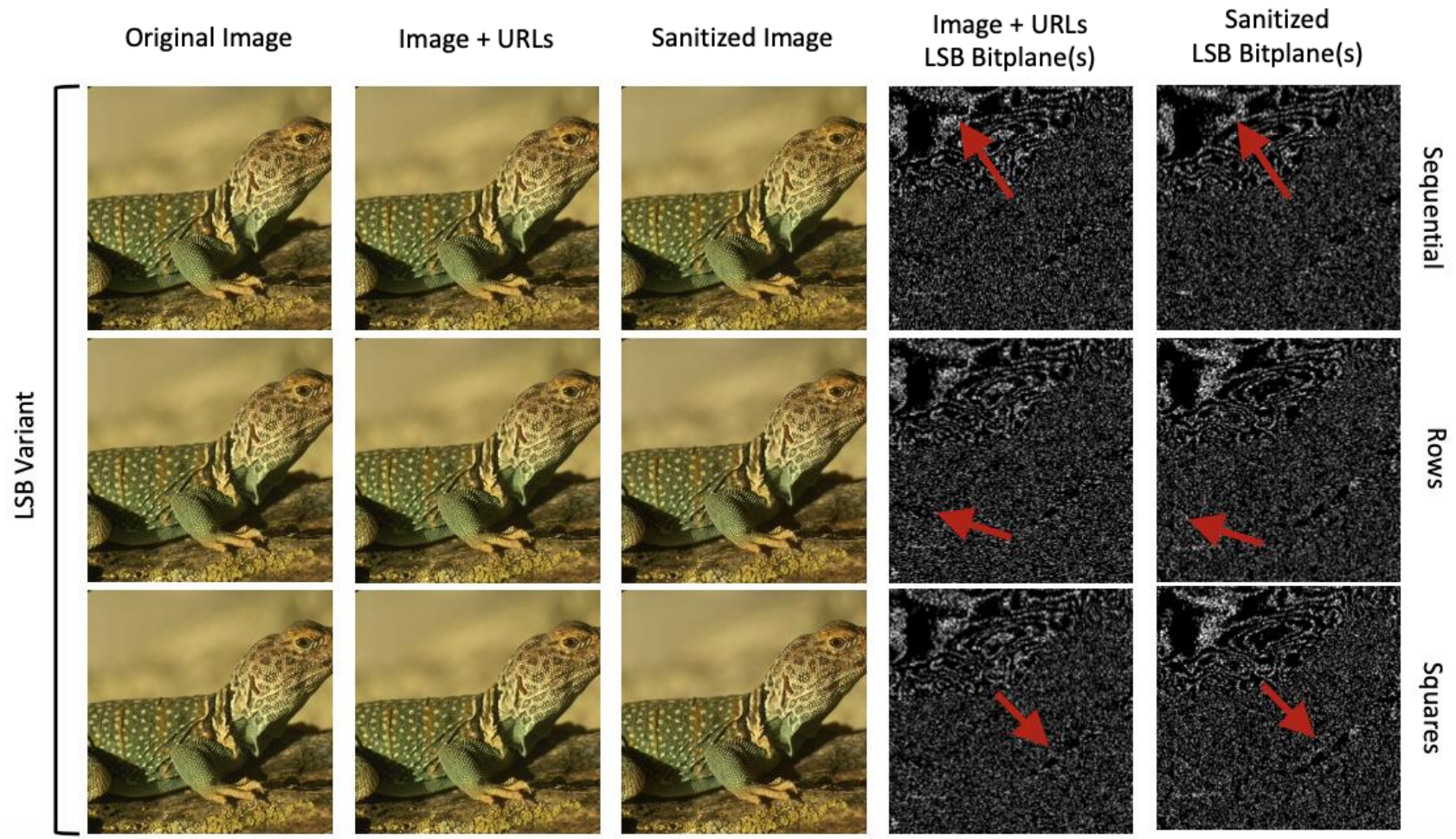
- We compared the legitimate images included in the test set with their sanitized counterpart

Neural Model	LSB Technique	Pattern	MSE	PSNR	Training [s]	Inference [s]
DeepAE	LSB plain	Sequential	6.36e-3	21.98	**4231**	**0.0016**
		Rows				
		Squares				
	LSB variant	Sequential				
		Rows				
		Squares				
U-Net	LSB plain	Sequential	2e-4	36.95	13648	0.0074
		Rows				
		Squares				
	LSB variant	Sequential				
		Rows				
		Squares				
U-Net+	LSB plain	Sequential	**1.6e-4**	**38.05**	18646	0.0067
		Rows				
		Squares				
	LSB variant	Sequential				
		Rows				
		Squares				

Results


Original Image
Image + URLs
Sanitized Image
Image + URLs
LSB Bitplane(s)
Sanitized
LSB Bitplane(s)
LSB Plain
Sequential
Rows
Squares

Results


Original Image
Image + URLs
Sanitized Image
Image + URLs
LSB Bitplane(s)
Sanitized
LSB Bitplane(s)
LSB Variant
Sequential
Rows
Squares

Conclusions

- We devised a U-Net-like model for sanitizing images, cloaking a list of URLs
- Experiments demonstrate the effectiveness of the proposed approach in disrupting embedded information while restoring the original images

- One of the limitations relies on the need of a suitable amount of images containing hidden data
  - In future works, we plan to investigate semi-supervised methods to train reliable models with scarce data

Thank You for your attention!

Questions?

ISMIS 2024
17-19 JUNE 2024
POITIERS / FUTUROSCOPE, FRANCE

Angelica Liguori
ICAR-CNR
angelica.liguori@icar.cnr.it