Generative Autoencoders as Watermark Attackers: Analyses of Vulnerabilities and Threats

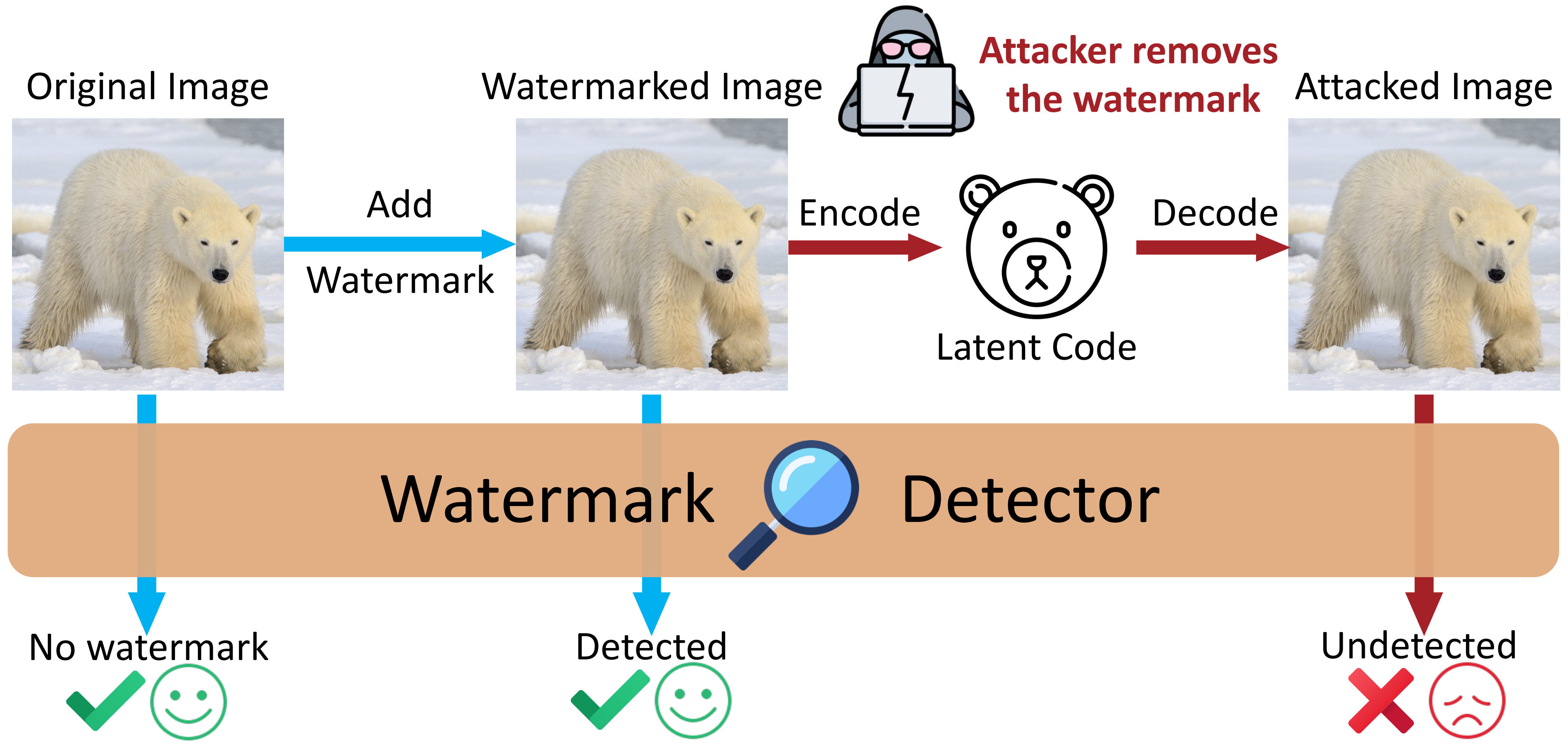
Xuandong Zhao* Kexun Zhang* Yu-Xiang Wang Lei Li
xuandongzhao@ucsb.edu

UC SANTA BARBARA

ICML
International Conference
On Machine Learning

Overview

Generative autoencoders can be used to remove most existing invisible image watermarks.



Method

Encode and decode to remove the watermark.

$$\hat{x} = Dec(Enc(\text{Watermark}(x))) \approx x$$

Why it works?

- The encoding compresses the image (**VAE**) and adds some noise to it (**Diffusion**) to remove the watermark.
- Generative autoencoders are trained to sample from a distribution where most images are **watermark-free**.
- The decoding process **reconstructs** the image and preserves the image quality.

Attacked Samples

Experiment Results

- **Our method is the most effective in removing watermarks.**
- **Diffusion is better than VAE in retaining image quality.**

Attacker	PSNR↑	SSIM↑	FID↓	Bit Acc↓	Word Acc↓
DCT-DWT-SVD based watermarking:					
Brightness	12.07	0.707	21.16	**0.443**	0.03
Contrast	18.34	0.801	16.99	**0.443**	0.02
JPEG	31.93	0.906	35.00	0.688	0.00
BM3D	33.37	0.896	90.41	0.576	0.00
DPIR	34.84	0.945	18.47	0.918	0.21
Bmshj2018	31.02	0.873	77.11	0.526	**0.00**
Cheng2020	31.96	0.887	69.03	**0.525**	**0.00**
Diffusion	24.88	0.712	41.37	0.643	**0.00**
RivaGAN watermarking:					
Brightness	12.05	0.705	30.87	0.992	0.87
Contrast	18.34	0.802	25.43	0.995	0.89
JPEG	32.05	0.906	35.92	0.959	0.41
BM3D	33.43	0.896	91.59	0.950	0.34
DPIR	34.96	0.945	18.77	0.996	0.87
Bmshj2018	31.07	0.873	78.45	**0.648**	**0.00**
Cheng2020	32.03	0.888	67.82	**0.636**	**0.00**
Diffusion	24.82	0.706	45.25	**0.629**	**0.00**

Attacker	PSNR↑	SSIM↑	FID↓	Bit Acc↓	Word Acc↓
SSL watermarking:					
Brightness	12.08	0.705	32.38	0.999	0.99
Contrast	18.37	0.803	29.80	1.000	1.00
JPEG	32.05	0.904	43.19	0.805	0.01
BM3D	33.67	0.897	93.13	**0.671**	0.00
DPIR	35.10	0.945	26.94	0.937	0.26
Bmshj2018	31.14	0.874	80.63	**0.640**	**0.00**
Cheng2020	32.10	0.887	71.08	**0.634**	**0.00**
Diffusion	24.83	0.707	47.42	0.719	**0.00**
StegaStamp watermarking:					
Brightness	12.03	0.727	51.72	1.000	1.00
Contrast	17.97	0.805	51.29	1.000	1.00
JPEG	26.89	0.840	72.04	1.000	1.00
BM3D	28.57	0.874	118.14	1.000	1.00
DPIR	27.67	0.876	58.17	1.000	1.00
Bmshj2018	27.75	0.847	93.36	1.000	1.00
Cheng2020	28.33	0.868	85.72	1.000	1.00
Diffusion	22.63	0.622	69.65	**0.648**	**0.50**

Generative autoencoders are shaded. Top-3 attacks are **bolded**. Best image among top-3 is underlined.

Takeaway

- **Pixel-level post-processing watermarks are not reliable and can be easily moved. Consider using stronger watermarks for your images.**