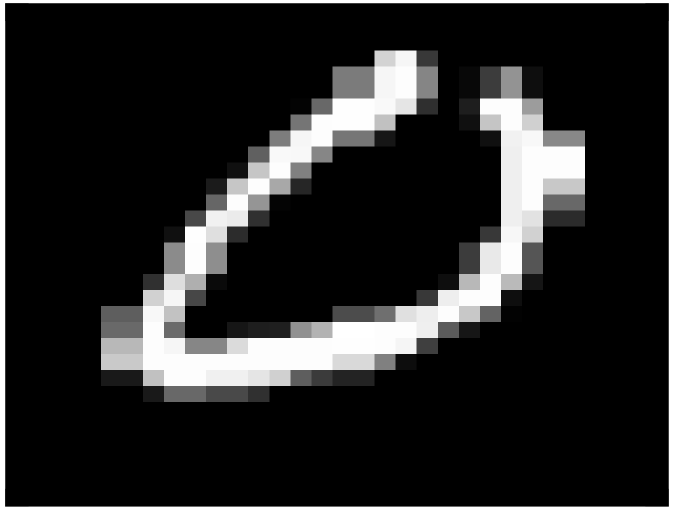
Robust Detectors

Neil Gong

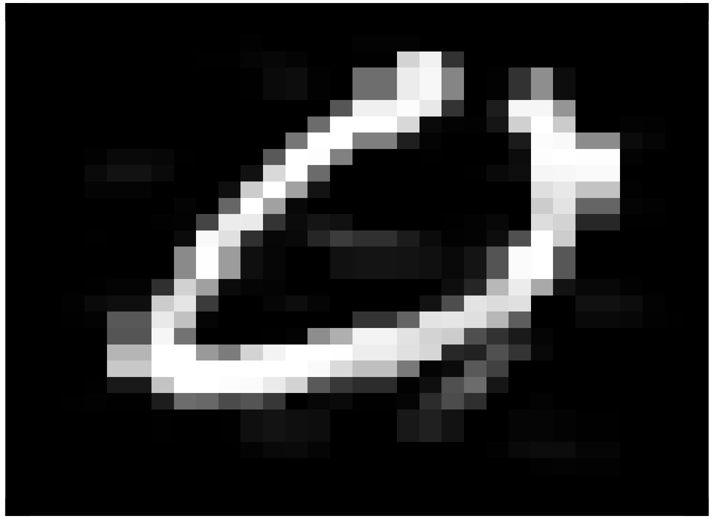
AI-generated image detectors

- Passive
- Watermark-based
- Robustness issues
  - Fake → real
    - Removal
  - Real → fake
    - Forgery

Adversarial Examples



Normal example: digit 0



Adversarial example: predicted to be 9

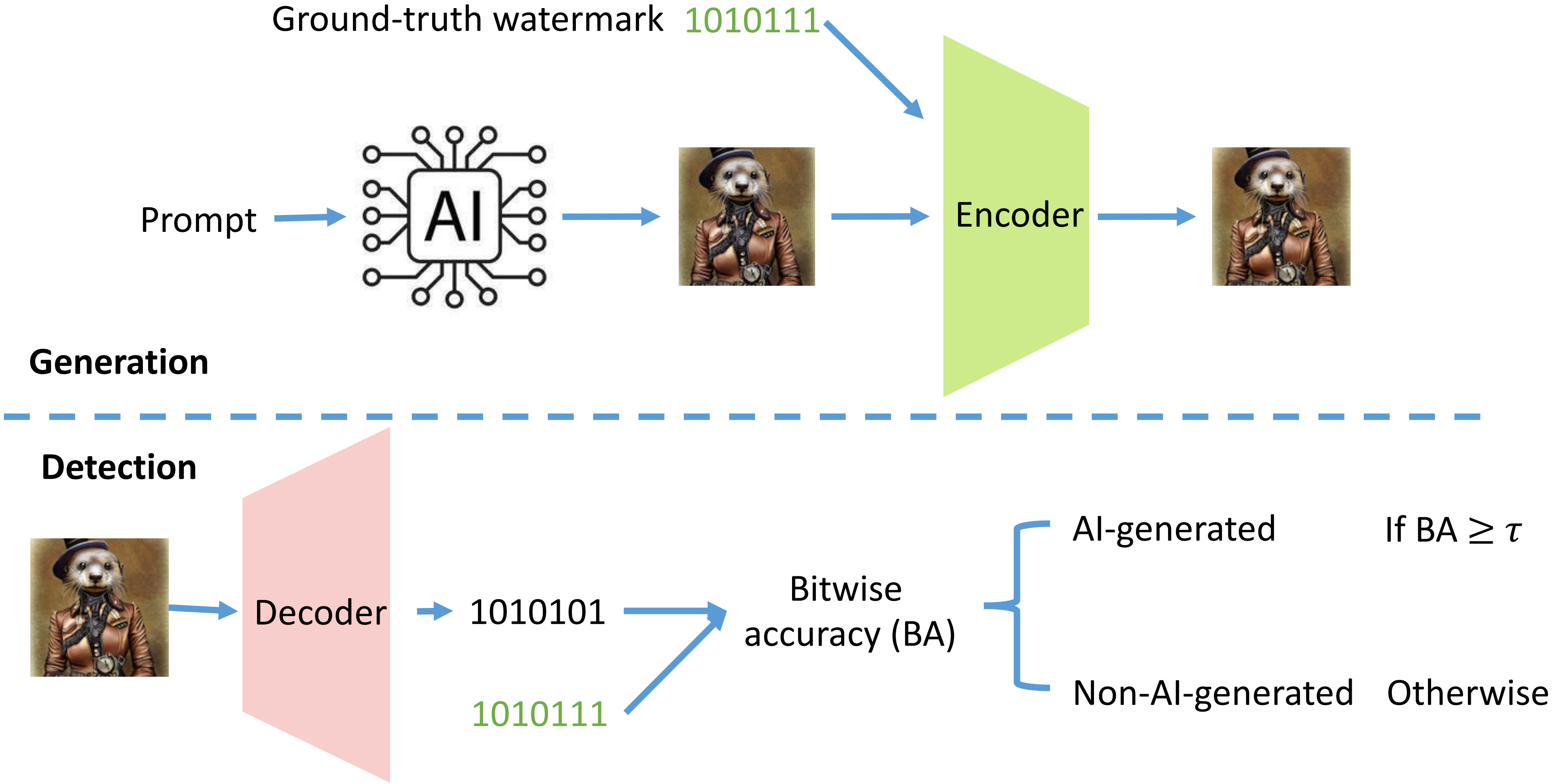
Building robust detectors

- Adversarial training

- Certifiably robust detectors
  - Randomized smoothing

Adversarial training – passive detector

Watermark-based detector



Ground-truth watermark 1010111

Prompt

Encoder

Generation

Detection

Decoder

1010101

1010111

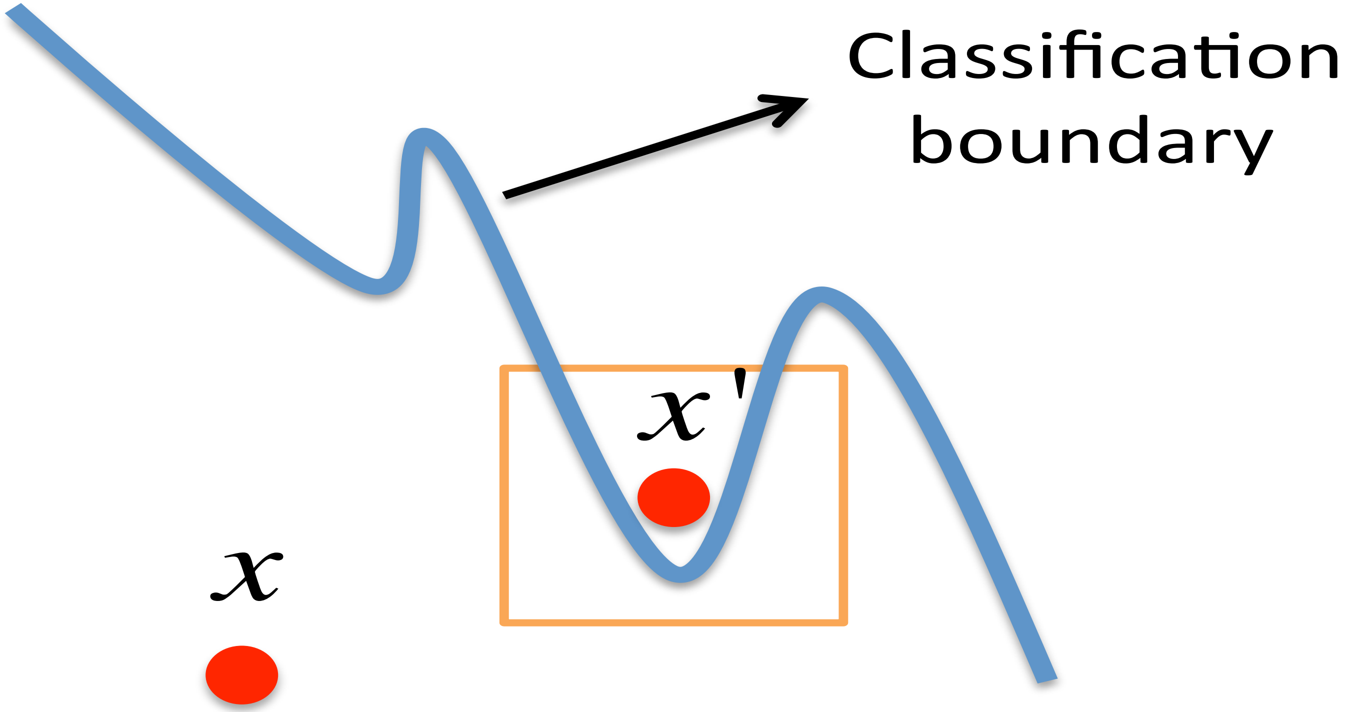
Bitwise accuracy (BA)

AI-generated If BA $\geq \tau$

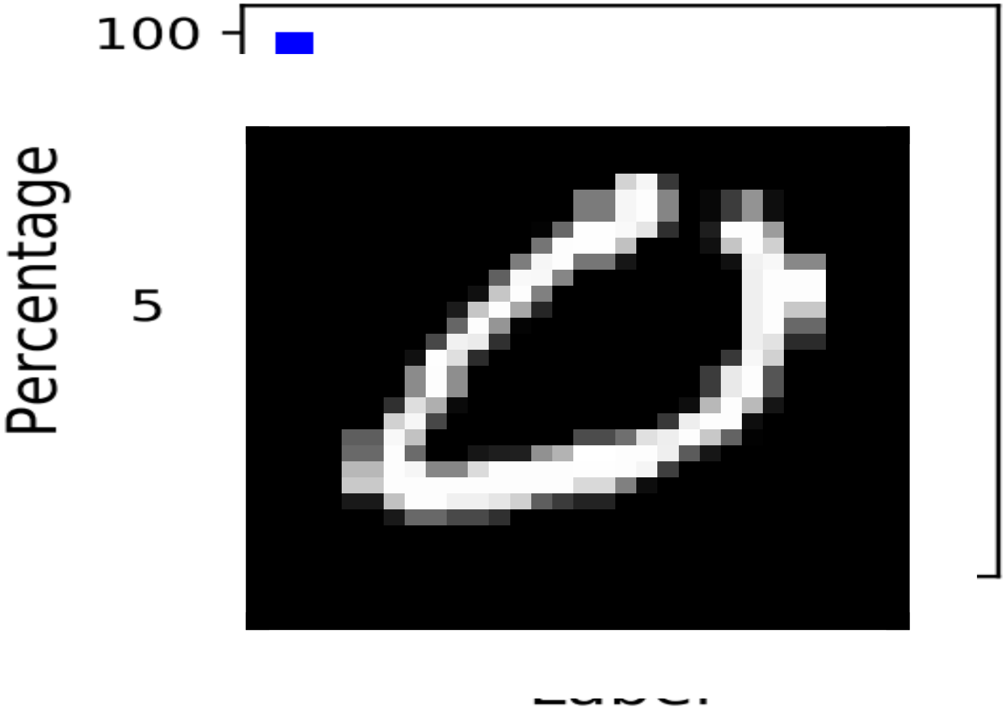
Non-AI-generated Otherwise

Adversarial training – watermark-based detector

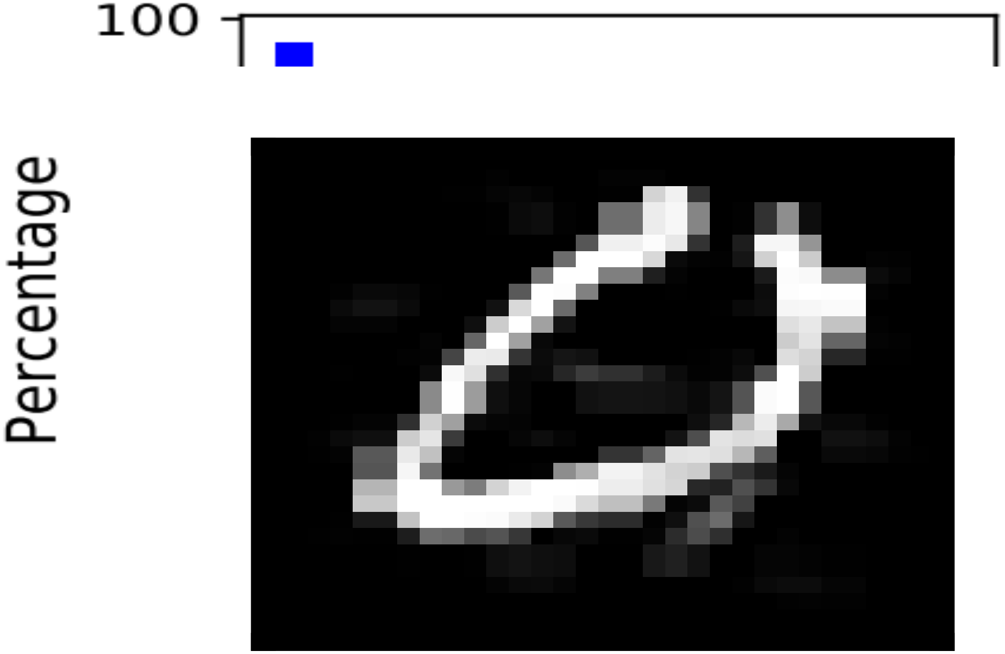
Adversarial example is close to classification boundary?



Measuring Adversarial Examples

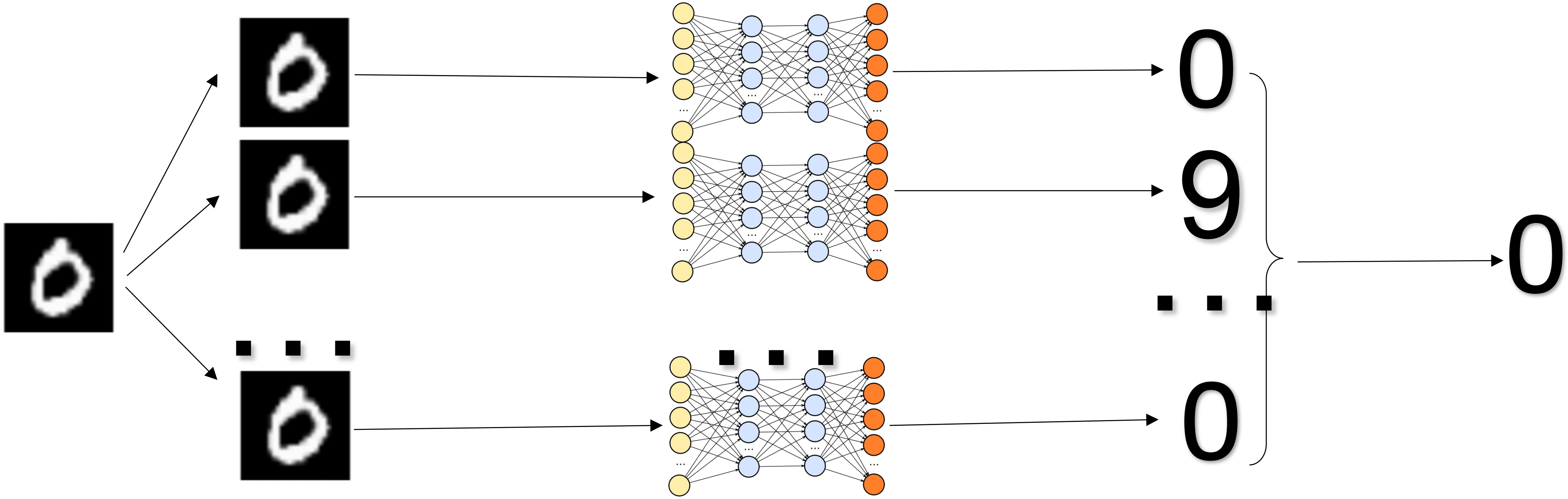


A normal example: digit 0



An adversarial example
with a target label 9

Randomized smoothing



Formal definition of randomized smoothing

- Input
    - a classifier f
    - an example x
    - a noise distribution

- Output
    - $\mathrm{g}(x) = \underset{c}{\mathrm{argmax}}\,\Pr(f(x + r) = c)$

Robustness guarantee

- Noise is isotropic Gaussian distribution
- $\mathrm{g}(x+\delta)=C_A$ when $|\delta|_2 \leq \mathcal{E}$

$$\mathcal{E}= \frac{\sigma}{2}\left(\Phi^{-1}(\underline{p_A})-\Phi^{-1}(\overline{p_B})\right)$$

Certified radius

Tightness of the bound

- Given
  - No assumptions on the classifier f
  - Randomized smoothing with Gaussian noise

- The derived bound is tight

Estimating the label probabilities

- Sampling a large number of noise

- Predicting labels for the noisy examples

- Estimating label probabilities with probabilistic guarantees

Randomized smoothing

- Strengths
    - Applicable to any classifier
    - Scalable to large classifier

- Limitations
    - Efficiency – need many predictions
    - Probabilistic guarantees

Variants of randomized smoothing

- Multi-label

- Regression

Certifiably robust passive detector

Testing Robustness of Image Watermarks

Watermark removal

Watermarked

+

Perturbation

=

Non-watermark

$\text{BA} < \tau$

Watermark forgery

Non-watermarked

+

Perturbation

=

Watermarked

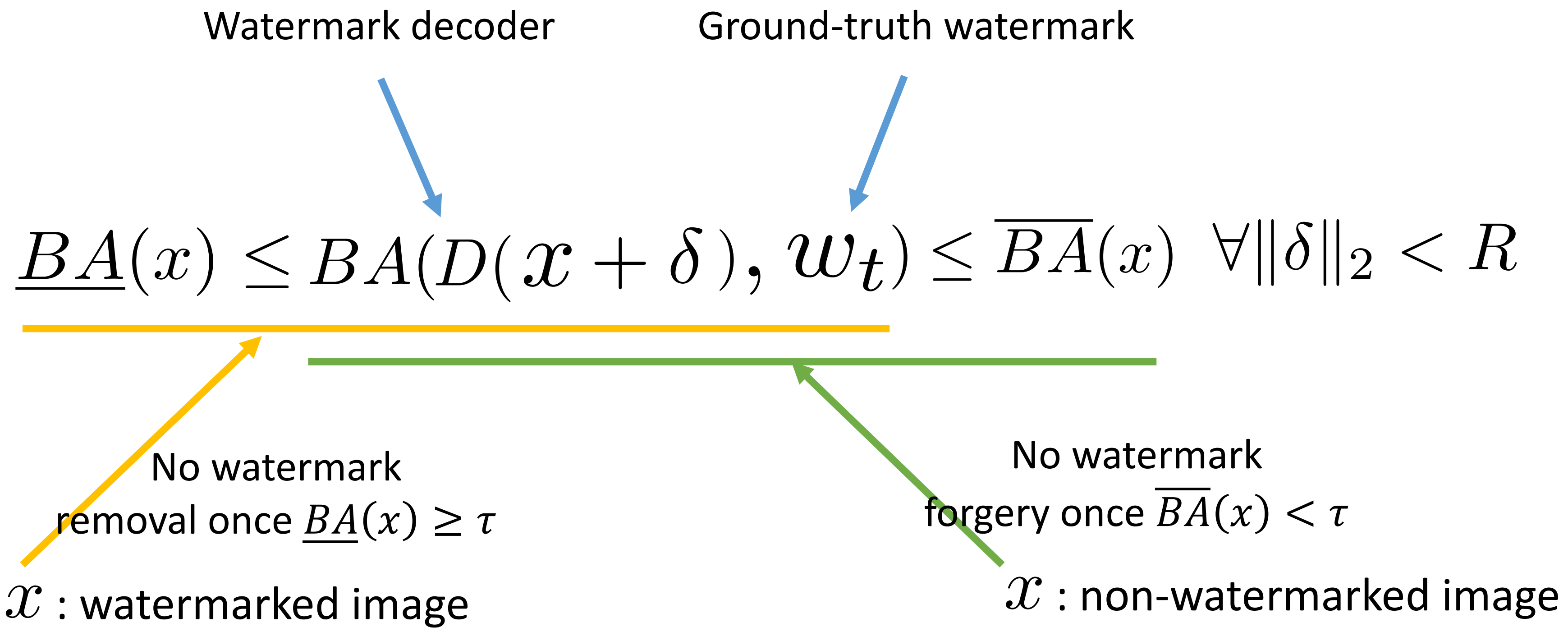
$\text{BA} \geq \tau$

Certifiably Robust Image Watermark - Definition

$$\underline{BA}(x) \leq BA(\overset{\substack{\text{Watermark decoder}\\ \downarrow}}{D}(x+\delta), \overset{\substack{\text{Ground-truth watermark}\\ \downarrow}}{w_t}) \leq \overline{BA}(x) \quad \forall \|\delta\|_2 < R$$

Jiang et al. "Certifiably Robust Image Watermark". In *European Conference on Computer Vision (ECCV),* 2024.

Certifiably Robust Image Watermark - Definition



Watermark decoder

Ground-truth watermark

$$\underline{BA}(x) \leq BA(D(x+\delta), w_t) \leq \overline{BA}(x) \quad \forall \|\delta\|_2 < R$$

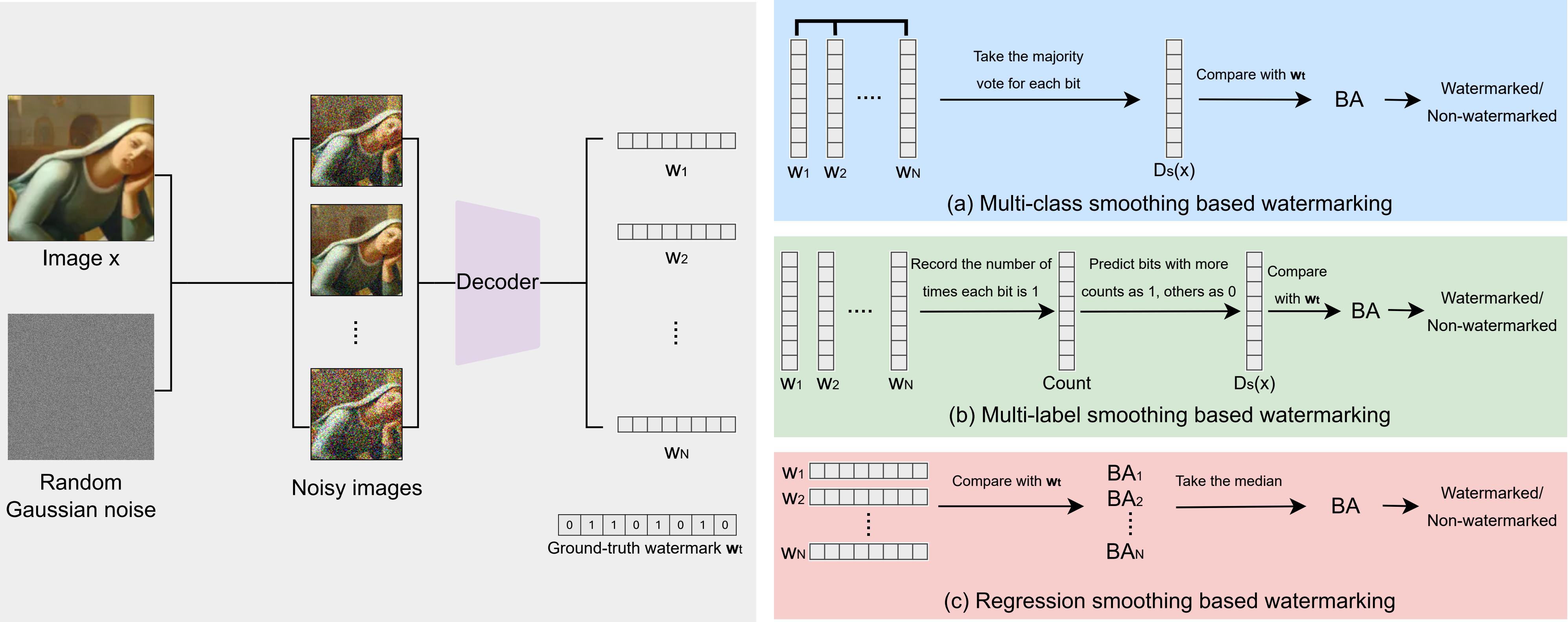
No watermark removal once $\underline{BA}(x) \geq \tau$

x : watermarked image

No watermark forgery once $\overline{BA}(x) < \tau$

x : non-watermarked image

Certifiably robust watermark-based detector



(a) Multi-class smoothing based watermarking

(b) Multi-label smoothing based watermarking

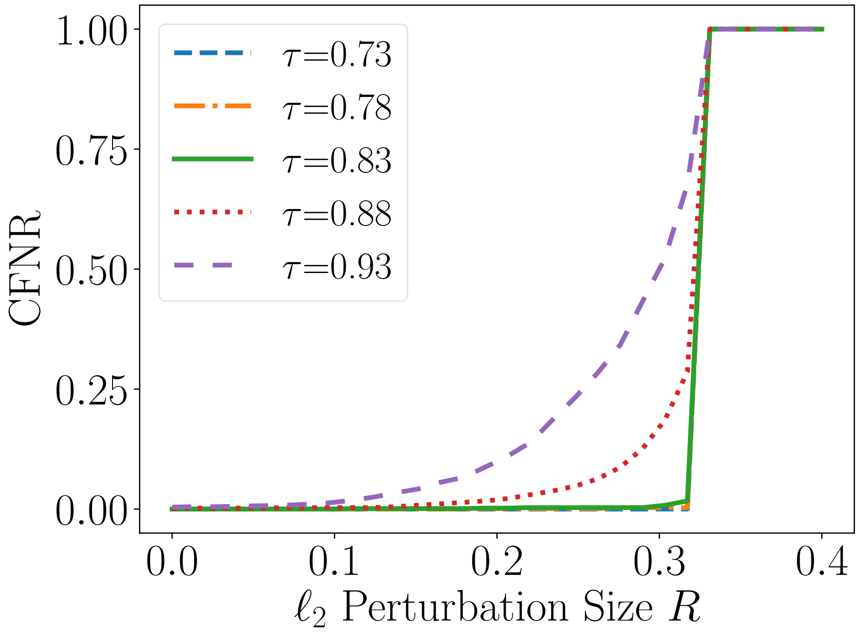
(c) Regression smoothing based watermarking

Evaluation metrics

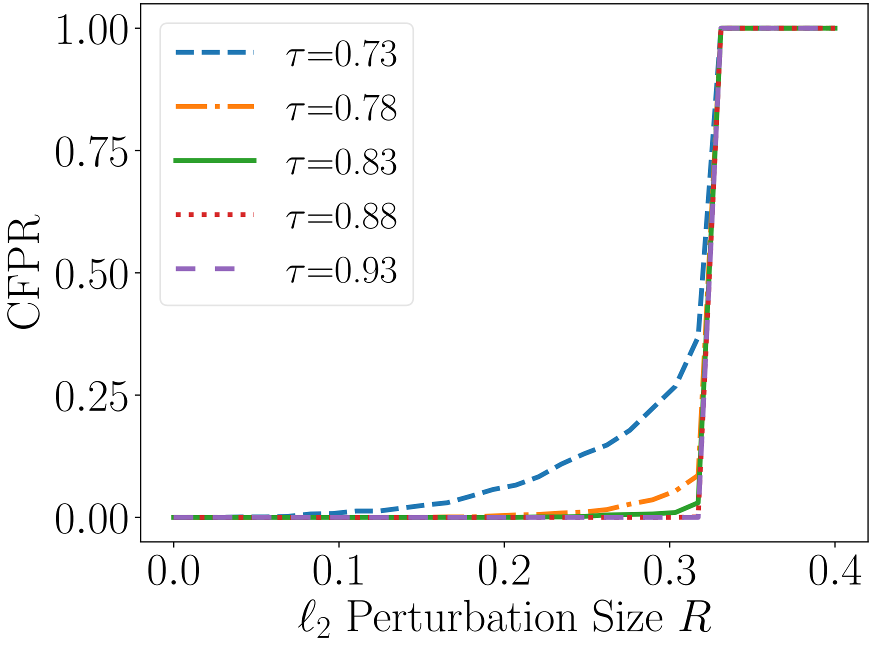
$$CFNR = \frac{1}{|X_w|} \sum_{x_w \in X_w} \mathbb{I}(\underline{BA}(x_w) < \tau)$$

$$CFPR \quad = \frac{1}{|X_n|} \sum_{x_n \in X_n} \mathbb{I}(\overline{BA}(x_n) \geq \tau)$$

Experimental Results on Stable Diffusion



Watermark removal



Watermark forgery

Summary

- Building robust detectors
  - Adversarial training
  - Randomized smoothing