Suppressing High-frequency Artifacts for Generative Model Watermarking by Anti-aliasing

L. Zhang, Y. Liu, X. Zhang and Hanzhou Wu

Shanghai University

ACM IH&MMSec'24, Vigo, Spain

Outline

1 Introduction

DNN Watermarking/Model Watermarking

- Embed watermarks into DNN models
- Protect the intellectual property of DNN models

Computing resources

Professional skills

High-quality datasets

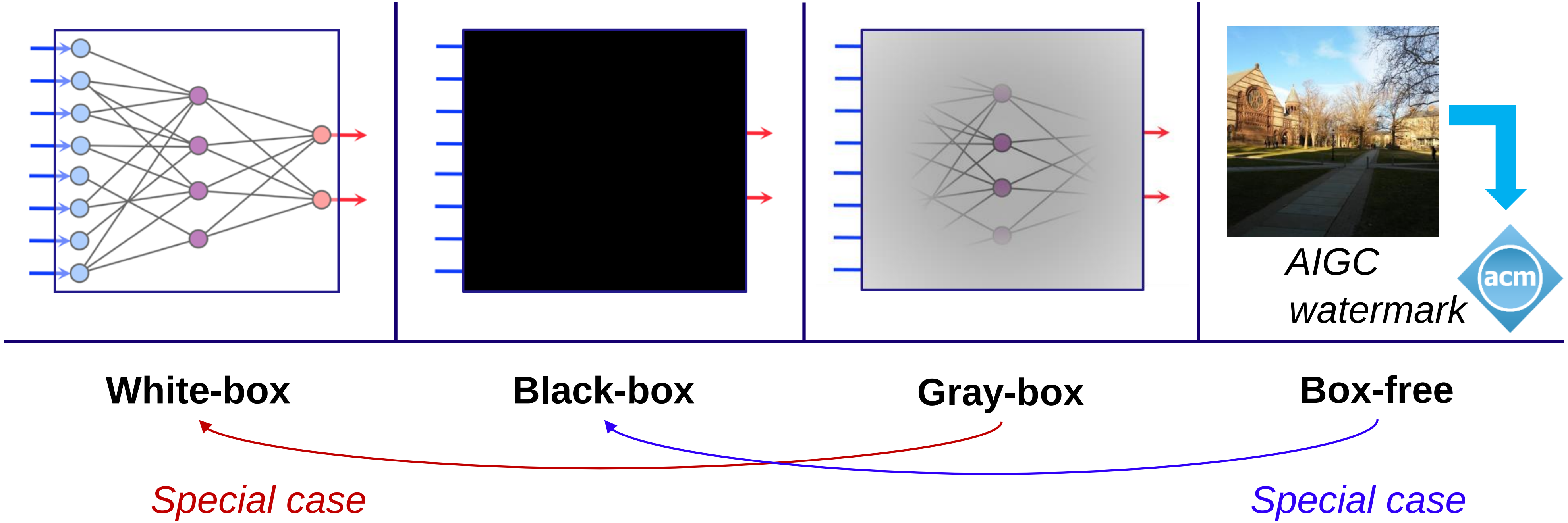
Machine Learning as a Service (MLaaS)

Many threats

1 Introduction

Categories

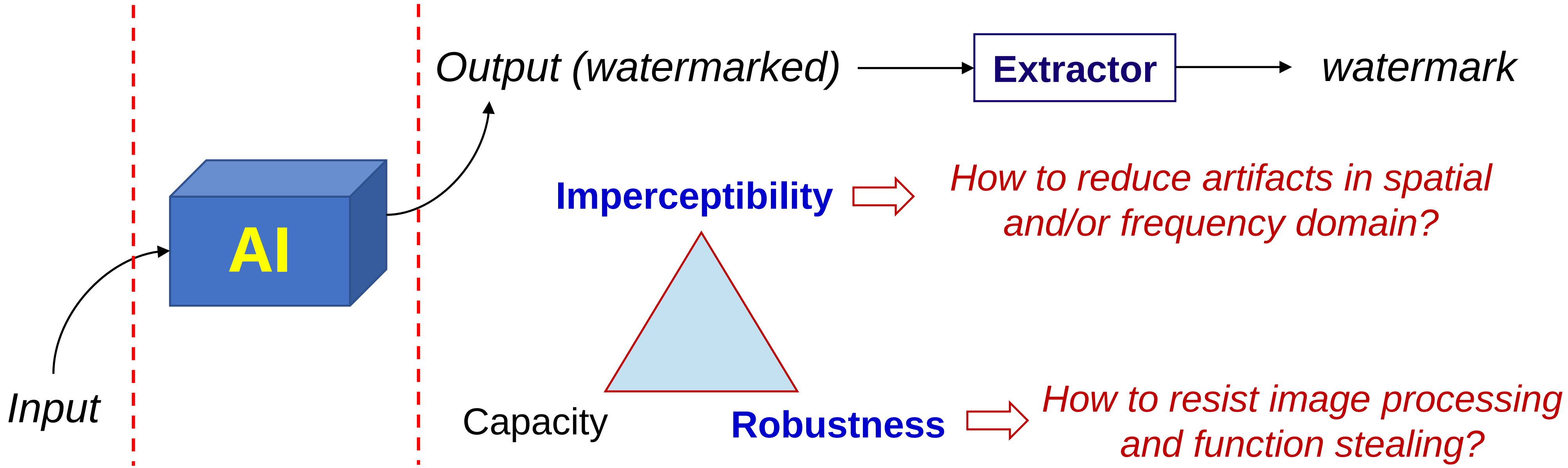
- *Whether the extractor can access or interact with the model?*
- White-box/Black-box/Gray-box/Box-free DNN Watermarking



1 Introduction

Motivation

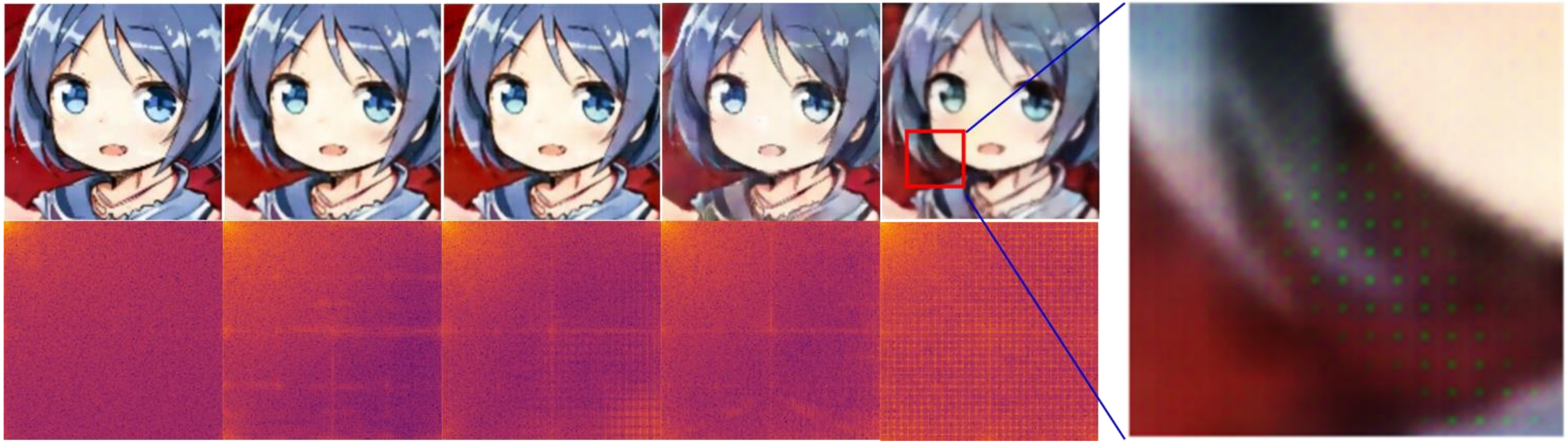
- **Box-free: Generative Model Watermarking**
- E.g., any image generated by a certain DNN model must contain a pre-determined watermark



1 Introduction

Motivation

- **Imperceptibility**: existing works easily introduce high-frequency artifacts which impair the concealment of the hidden watermark



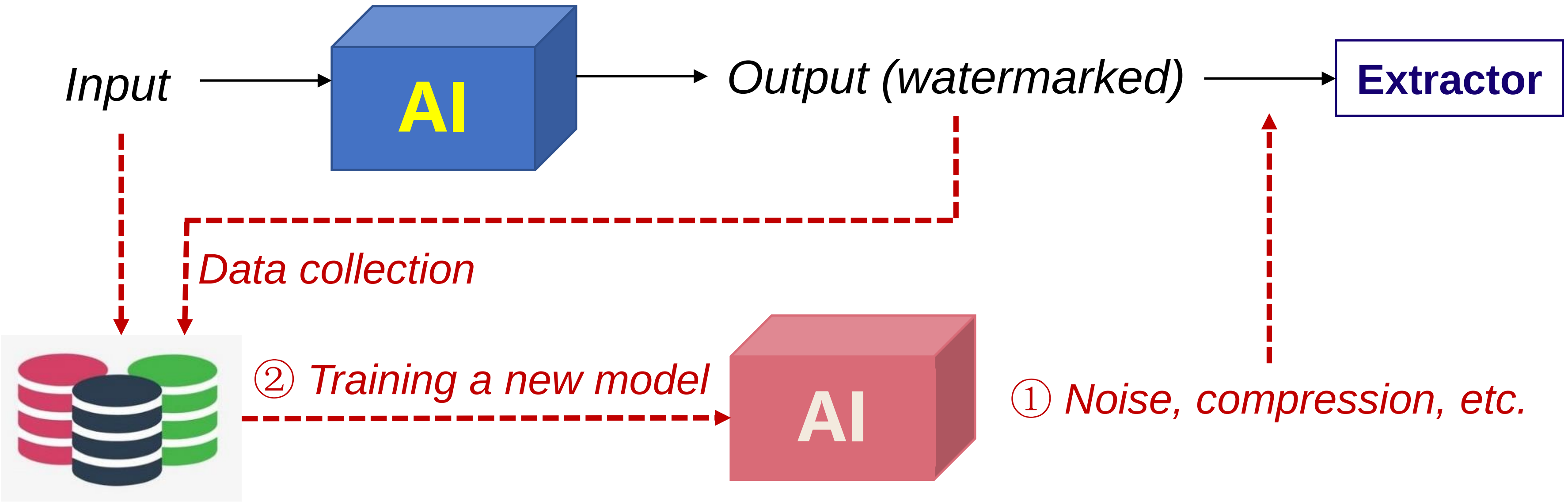
(a) Natural | (b) Non-marked | (c) Marked **Method #1** | (d) Marked **Method #2** | (e) Marked **Method #3** | Partial zoom-in

1 Introduction

Motivation

Robustness

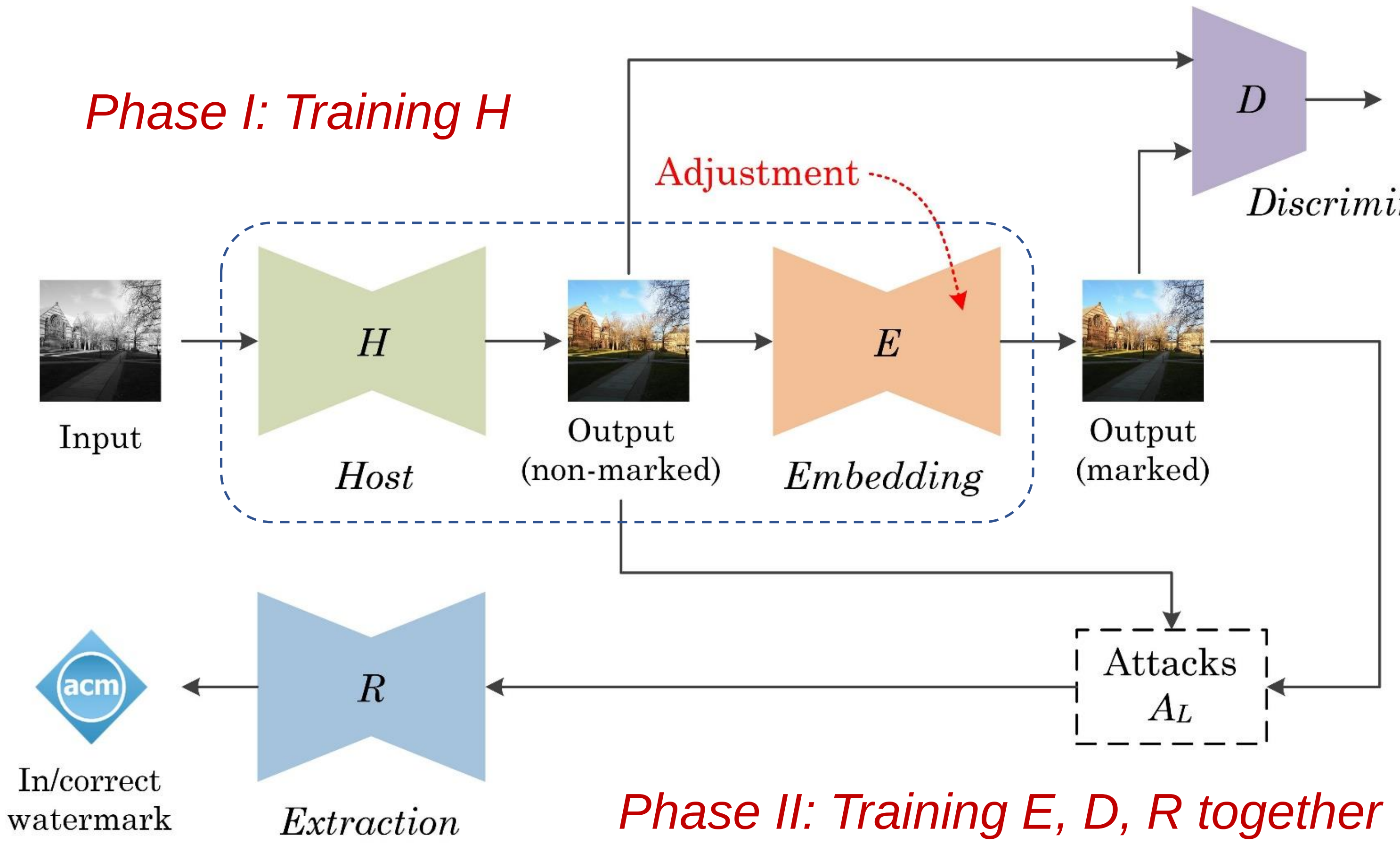
- ① Marked images may be attacked before watermark extraction
- ② Attackers may collect a set of input-output pairs to train a new model



Outline

2 Proposed Method

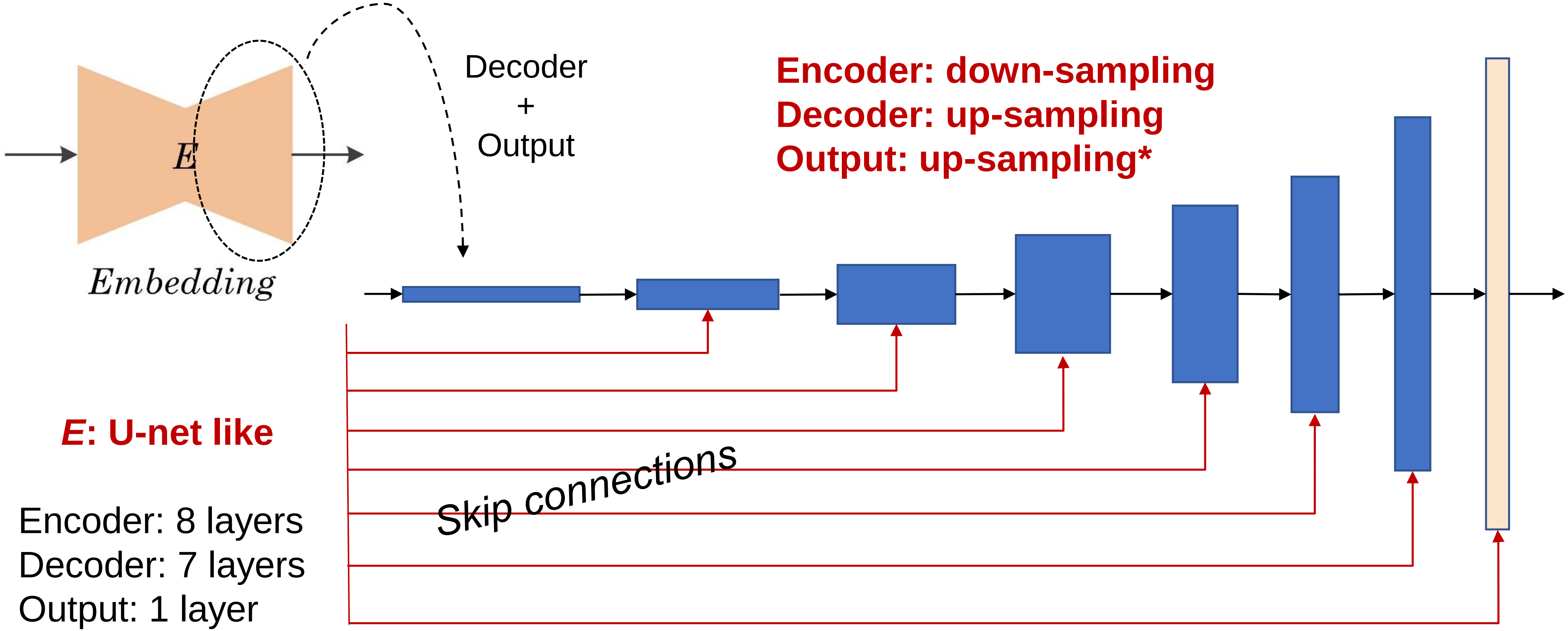
❑ General Framework



- ✓ *E*: to reduce high-frequency artifacts
- ✓ A_L: to enhance robustness of *R*

2 Proposed Method

❑ Structural Design of *E*



***E*: U-net like**

Encoder: 8 layers
Decoder: 7 layers
Output: 1 layer

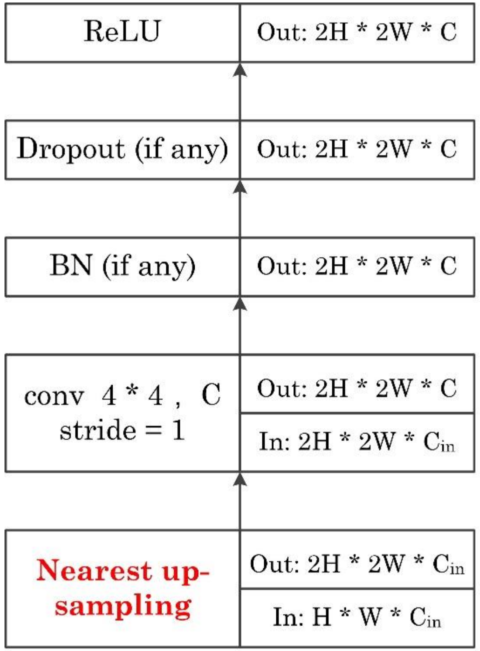
2 Proposed Method

Structural Design of *E*

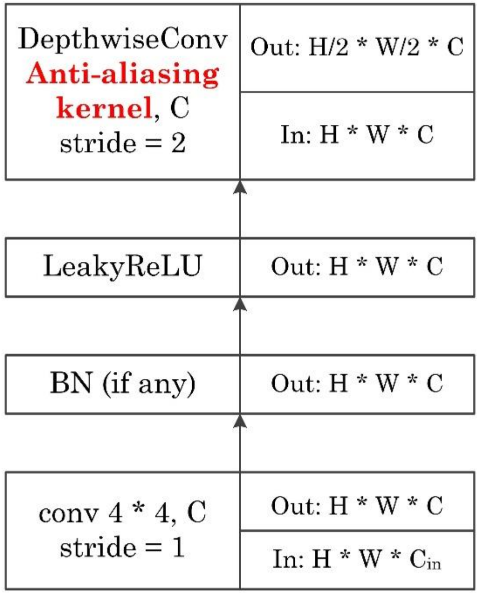
- *Up-sampling (left), down-sampling (middle), output layer (right)*

Anti-aliasing: low-pass filtering

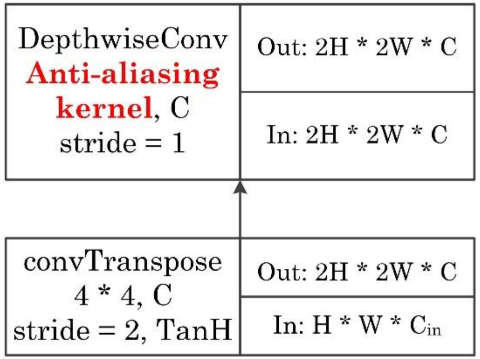
$(1, 5, 10, 10, 5, 1)^T(1, 5, 10, 10, 5, 1)$



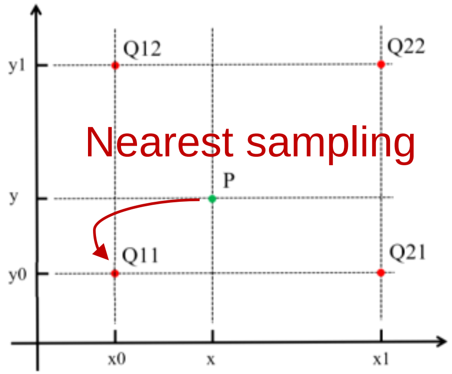
Up-sampling module



Down-sampling module

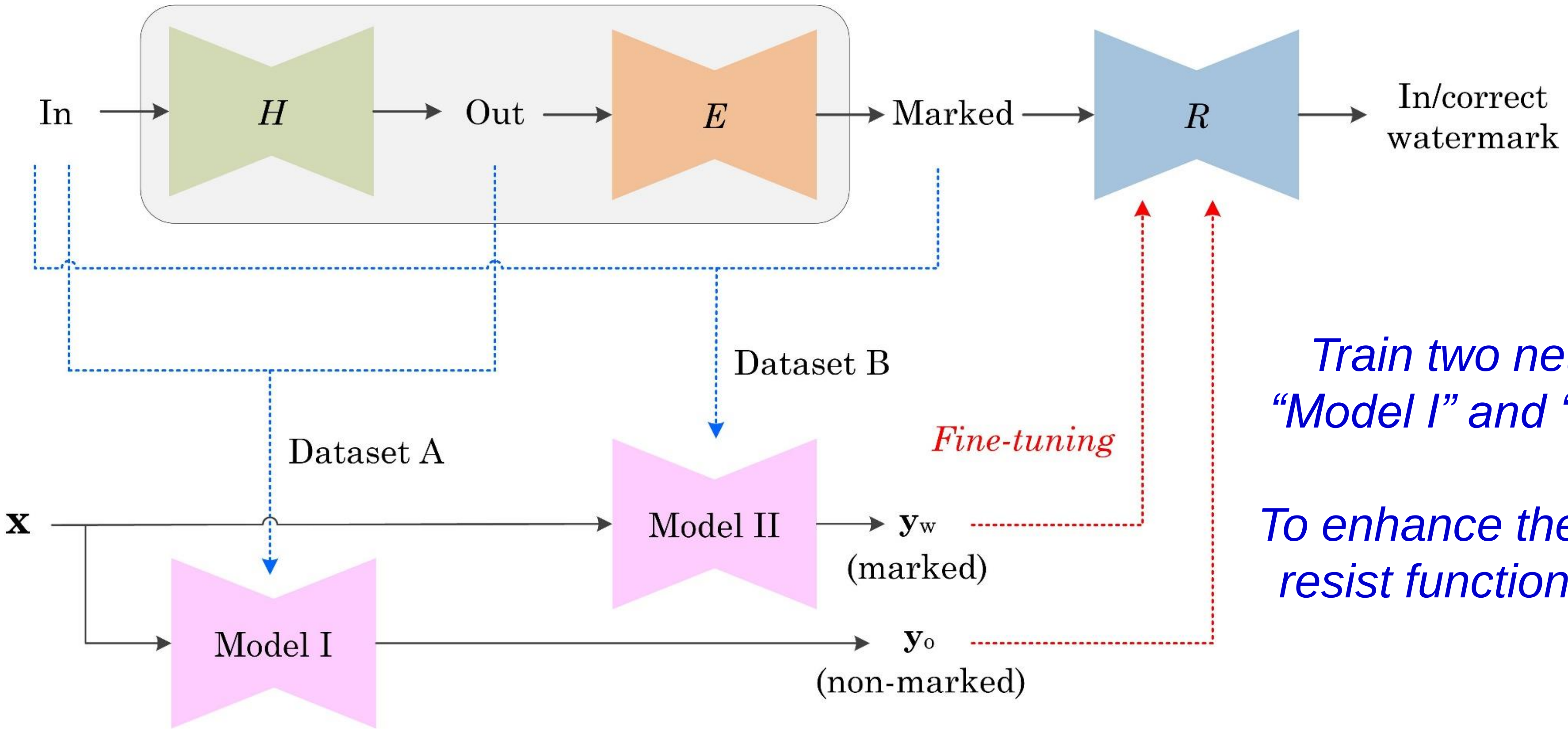


Output layer



2 Proposed Method

❑ Adversarial Fine-tuning to Resist Function Stealing



Train two networks: "Model I" and "Model II"

To enhance the ability to resist function stealing

Outline

3 Experimental Results and Analysis

Qualitative Results

- Tasks: paint transfer (left) & style transfer (right)

Non-marked

Marked

Watermark

DCT heat maps

3 Experimental Results and Analysis

Quantitative Results

The marked images are of high quality

TABLE I

QUALITY ASSESSMENT FOR THE MARKED IMAGES AND THE EXTRACTED COLOR WATERMARKS OVER THE TEST SET. ALL EXPERIMENTAL RESULTS SHOWN IN THIS TABLE ARE MEAN VALUES. $PSNR_w$ MEASURES THE QUALITY OF THE EXTRACTED COLOR WATERMARKS.

Task	Watermark	Mean PSNR	Mean SSIM	Mean MS-SSIM	Mean VIF	Mean $PSNR_w$	SR
Paint transfer	*Lena*	35.08	0.987	0.999	0.913	50.73	100%
Paint transfer	*Baboon*	35.29	0.988	0.999	0.917	40.14	100%
Paint transfer	*Peppers*	35.02	0.986	0.999	0.914	44.97	100%
Style transfer	*Lena*	41.61	0.998	0.999	0.948	53.74	100%
Style transfer	*Baboon*	41.93	0.998	0.999	0.954	42.72	100%
Style transfer	*Peppers*	41.53	0.998	0.999	0.951	48.68	100%

TABLE II

QUALITY ASSESSMENT FOR THE MARKED IMAGES AND THE EXTRACTED BINARY WATERMARKS OVER THE TEST SET. ALL EXPERIMENTAL RESULTS SHOWN IN THIS TABLE ARE MEAN VALUES. BER MEASURES THE QUALITY OF THE EXTRACTED BINARY WATERMARKS.

Task	Watermark	Mean PSNR	Mean SSIM	Mean MS-SSIM	Mean VIF	Mean BER	SR
Paint transfer	*IEEE*	34.98	0.986	0.999	0.913	0	100%
Style transfer	*IEEE*	41.03	0.997	0.999	0.948	0	100%

上海大学 SHANGHAI UNIVERSITY

3 Experimental Results and Analysis

Quantitative Results

Robust against common image processing operations

(a) (b) (c) (d) (e)

(a) (b) (c) (d) (e)

TABLE III

SR AGAINST DIFFERENT PREPROCESSING OPERATIONS. "PT" MEANS "PAINT TRANSFER" AND "ST" MEANS "STYLE TRANSFER".

Noise
Resizing
JPEG
Flipping
…

Task	Watermark	Noise addition			Resizing			JPEG compression			Flipping	
		$\sigma = 0.1$	$\sigma = 0.15$	$\sigma = 0.2$	$128^2 \times 3$	$196^2 \times 3$	$512^2 \times 3$	QF = 50	QF = 70	QF = 90	horizontal	vertical
PT	*Lena*	100%	100%	99.60%	78.20%	98.60%	100%	98.20%	98.60%	99.40%	100%	99.00%
PT	*Baboon*	100%	98.80%	97.40%	97.20%	99.80%	100%	99.40%	99.60%	100%	100%	98.80%
PT	*Peppers*	100%	99.40%	96.60%	89.20%	100%	100%	99.20%	99.40%	100%	100%	99.60%
PT	*IEEE*	99.60%	99.20%	98.20%	84.20%	99.60%	100%	98.20%	98.40%	99.40%	100%	99.60%
ST	*Lena*	99.73%	99.33%	96.40%	97.20%	99.80%	99.93%	93.73%	96.67%	99.07%	100%	99.47%
ST	*Baboon*	99.47%	98.13%	94.67%	83.80%	99.73%	100%	95.60%	98.93%	99.80%	100%	99.80%
ST	*Peppers*	100%	99.53%	97.33%	98.07%	100%	100%	95.27%	98.07%	99.00%	100%	100%
ST	*IEEE*	100%	99.93%	99.33%	99.20%	100%	100%	96.00%	99.54%	100%	100%	100%

3 Experimental Results and Analysis

Quantitative Results

Robust against function stealing

TABLE IV

SR AGAINST THE SURROGATE NETWORK ATTACK, WHERE THE ℓ_1 LOSS FUNCTION WAS USED FOR NETWORK TRAINING.

Task	Watermark	ConvGen	ResGen	UnetGen
Paint transfer	*Lena*	100%	100%	100%
Paint transfer	*IEEE*	99.60%	100%	99.60%
Style transfer	*Lena*	100%	100%	100%
Style transfer	*IEEE*	99.54%	99.80%	99.54%

Different networks:

ConvGen: CNN
ResGen: ResNet-like
UnetGen: Unet-like

TABLE V

SR AGAINST THE SURROGATE NETWORK ATTACK, WHERE DIFFERENT LOSS FUNCTIONS WERE USED FOR NETWORK TRAINING.

Task	Watermark	UnetGen					
		ℓ_1	$\ell_1 + \ell_{per}$	$\ell_1 + \ell_{per} + \ell_{adv}$	ℓ_2	$\ell_2 + \ell_{per}$	$\ell_2 + \ell_{per} + \ell_{adv}$
Paint transfer	*Lena*	100%	100%	100%	100%	100%	100%
Paint transfer	*IEEE*	99.60%	100%	100%	100%	99.80%	100%
Style transfer	*Lena*	100%	100%	100%	100%	100%	100%
Style transfer	*IEEE*	99.54%	99.87%	99.80%	99.07%	99.67%	99.80%

Different loss functions

3 Experimental Results and Analysis

Quantitative Results

Better than previous methods

TABLE VI
COMPARISON BETWEEN DIFFERENT MODEL WATERMARKING METHODS IN TERMS OF ROBUSTNESS AND IMPERCEPTIBILITY.

Method	Robustness against surrogate attack	Imperceptibility	
		Spatial domain	Frequency domain
Ref. [28]	Yes	Yes	
Ref. [16]		Partially	
Ref. [29]		Partially	
Proposed	Yes	Yes	Yes

TABLE VII
MEAN PSNRS (FOR COLOR WATERMARKS, DB) AND MEAN BERS (FOR BINARY WATERMARKS) BEFORE AND AFTER FILTERING OUT THE HIGH-FREQUENCY COMPONENTS OF THE MARKED IMAGES. "PT" MEANS "PAINT TRANSFER" AND "ST" MEANS "STYLE TRANSFER". THE SUPERSCRIPT "*" MEANS TO APPLY THE FILTERING OPERATION.

Task	Watermark	Ref. [28]	Ref. [16]	Ref. [29]	Proposed
PT	*Lena*	35.22 dB	26.16 dB	29.34 dB	50.73 dB
PT*	*Lena*	12.67 dB	16.53 dB	10.29 dB	48.56 dB
PT	*IEEE*	0.0027	0.0031	0.0015	0
PT*	*IEEE*	0.5260	0.4237	0.5726	0.0002
ST	*Lena*	34.05 dB	29.48 dB	26.01 dB	53.74 dB
ST*	*Lena*	12.65 dB	12.93 dB	14.10 dB	50.04 dB
ST	*IEEE*	0.0001	0.0004	0.0003	0
ST*	*IEEE*	0.5376	0.4386	0.4515	0.0001

After filtering out the high-frequency components of the marked image, the watermark can be accurately extracted, while previous arts cannot achieve this goal.

4 Conclusion and Discussion

Conclusion

- Reduce high-frequency artifacts of model watermarking by adjusting the structure of the watermark embedding network
- Enhance the robustness of the watermark extraction network through adversarial training and fine-tuning

Discussion

- Instead of network design, watermarking strategy (e.g., *use DWT to force the network to embed watermark into the low frequency area*) can be optimized to further reduce artifacts

Many Thanks!