

# Srishti Singh

## thesisfinal

 Thesis Mtech

---

### Document Details

Submission ID

trn:oid:::27535:140805278

Submission Date

May 28, 2026, 6:47 PM GMT+5:30

Download Date

Oct 5, 2026, 1:36 PM GMT+5:30

File Name

thesisfinal.pdf

File Size

1.5 MB

53 Pages

12,176 Words

69,763 Characters

# 6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

## Filtered from the Report

- Bibliography
- Small Matches (less than 10 words)

## Exclusions

- 1 Excluded Source

## Match Groups

-  **31 Not Cited or Quoted 5%**  
Matches with neither in-text citation nor quotation marks
-  **1 Missing Quotations 0%**  
Matches that are still very similar to source material
-  **1 Missing Citation 0%**  
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**  
Matches with in-text citation present, but no quotation marks

## Top Sources

- 5%  Internet sources
- 1%  Publications
- 4%  Submitted works (Student Papers)

## Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

## Match Groups

- 31 Not Cited or Quoted 5%**  
Matches with neither in-text citation nor quotation marks
- 1 Missing Quotations 0%**  
Matches that are still very similar to source material
- 1 Missing Citation 0%**  
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**  
Matches with in-text citation present, but no quotation marks

## Top Sources

- 5% Internet sources
- 1% Publications
- 4% Submitted works (Student Papers)

## Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

- 1** Student papers  
Delhi Technological University on 2026-05-20 <1%
- 2** Student papers  
Delhi Technological University on 2026-05-26 <1%
- 3** Student papers  
Delhi Technological University on 2026-05-25 <1%
- 4** Internet  
www.coursehero.com <1%
- 5** Student papers  
Delhi Technological University on 2026-05-24 <1%
- 6** Student papers  
Delhi Technological University on 2019-06-27 <1%
- 7** Internet  
arxiv.org <1%
- 8** Student papers  
Galgotias University, Greater Noida on 2026-05-02 <1%
- 9** Student papers  
Delhi Technological University on 2026-05-24 <1%
- 10** Internet  
gitarattan.edu.in <1%

|    |                |                                                             |     |
|----|----------------|-------------------------------------------------------------|-----|
| 11 | Internet       | iieta.org                                                   | <1% |
| 12 | Student papers | Delhi Technological University on 2026-04-09                | <1% |
| 13 | Internet       | www.ophtalmologyscience.org                                 | <1% |
| 14 | Student papers | Al-Ibdaa Institute - Iraq on 2026-06-03                     | <1% |
| 15 | Internet       | pub.tik.ee.ethz.ch                                          | <1% |
| 16 | Internet       | vtechworks.lib.vt.edu                                       | <1% |
| 17 | Student papers | Delhi Technological University on 2026-05-25                | <1% |
| 18 | Internet       | kups.ub.uni-koeln.de                                        | <1% |
| 19 | Internet       | mdpi-res.com                                                | <1% |
| 20 | Student papers | University of Engineering & Technology Mardan on 2025-08-12 | <1% |
| 21 | Student papers | Vrije Universiteit Amsterdam on 2026-01-26                  | <1% |
| 22 | Student papers | Vrije Universiteit Brussel on 2025-06-03                    | <1% |
| 23 | Internet       | armonisnord.ro                                              | <1% |
| 24 | Internet       | dspace.vutbr.cz                                             | <1% |

|                                                   |                |     |
|---------------------------------------------------|----------------|-----|
| 25                                                | Internet       |     |
| ece.uwaterloo.ca                                  |                | <1% |
| 26                                                | Student papers |     |
| Galgotias University, Greater Noida on 2026-05-24 |                | <1% |
| 27                                                | Student papers |     |
| Kean University on 2025-08-29                     |                | <1% |
| 28                                                | Student papers |     |
| Vrije Universiteit Amsterdam on 2026-01-26        |                | <1% |
| 29                                                | Internet       |     |
| zjzac.github.io                                   |                | <1% |

# Attention-Guided Dual-Domain Embedding for Intellectual Property Protection of Image Processing Model

A Thesis Submitted

In Partial Fulfillment of the Requirements for the  
Degree of

MASTERS OF TECHNOLOGY

in

Information Technology

by

Srishti Singh

(24/ISY/03)

Under the Supervision of

Dr. Varsha Sisaudia

Asst. Professor, Information Technology



Department of Information Technology

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

May, 2026



## DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

### ACKNOWLEDGEMENT

I've taken efforts in this research. However, it would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them. I'm highly indebted to **Dr. Varsha Sisaudia** for her guidance and constant supervision. I'm extremely indebted to **Prof. Dinesh Kumar Vishwakarma (HoD-IT)**, Department of Information Technology, Delhi Technological University, Delhi for their valuable suggestions and constant support throughout the tenure.

I would also like to express my sincere thanks to all faculty and staff members of the Department of Information Technology, Delhi Technological University, Delhi for their support. My thanks and appreciation also go to my friends and all the people who have willingly helped me out with their abilities.

**Srishti Singh**

**24/ISY/03**



## DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

### CANDIDATE'S DECLARATION

I **Srishti Singh**, Roll No. 24/ISY/03 student of M.Tech (Information Technology), hereby certify that the work which is being presented in the thesis entitled “**Attention-Guided Dual-Domain Embedding for Intellectual Property Protection of Image Processing Model**” in partial fulfillment of the requirements for the award of the Degree of Masters of Technology, submitted in the Department of Information Technology, Delhi Technological University is an authentic record of my own work carried out during the period from August, 2024 to May, 2026 under the supervision of **Dr. Varsha Sisaudia** (Asst. Professor, IT).

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

**Srishti Singh**



# DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

## CERTIFICATE BY THE SUPERVISOR

1 Certified that **Srishti Singh (24/ISY/03)** has carried out her research work presented in this thesis entitled **“Attention-Guided Dual-Domain Embedding for Intellectual Property Protection of Image Processing Model”** for the award of **Master of Technology** from the Department of Information Technology, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself, and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

2

**Dr. Varsha Sisaudia**

Asst. Professor, IT

Delhi, India

Date:

## ABSTRACT

Deep Learning based image processing models such as denoising, super-resolution and deraining consume significant computational resources, curated data and engineering effort. However when models such as those, are available through publicly accessible APIs, it renders them vulnerable to Surrogacy Model Extraction attacks, in which an adversary sends queries to the model, collects the output and training a functionally equivalent replica of the model without acquiring its parameters. Watermarking schemes typically fail to protect the model as, its reconstruction objective, is trained on the premise that the pixel level watermark embedded within it is random noise, and averaging it will result in elimination of such information.

This thesis proposes an Attention Guided Dual Domain Embedding (AGDDE) network framework for the IP protection of image processing models, in the form of Frequency Aware Structure Consistent Watermarking (FA-SCW) system. A dual branch attention mechanism which is much simpler and replacing the fixed Sobel detector used in previous work, learning a differentiable mask identifying both edges and flat regions for embedding. A dual domain embedding network which combines a correct U-shape Network in spatial domain and CNN in middle frequency DCT coefficients domain. A self-supervised frequency consistency loss can be achieved to maintain spectrum invariance under geometric perturbation providing robustness to crop, rotation, JPEG compression without knowing the surrogate model. Evaluation has been performed on Color Berkeley Segmentation Dataset of 68 images (CBSD68) and DIVERse 2K resolution dataset (DIV2K), to show the imperceptibility is more than 35 dB PSNR, surrogate extraction Success Rate (SR) is higher than 85% and Bit Error Rate (BER) is less than 0.10 which is 22.9% higher than nearest competitor at higher quality. Complete system takes 30 minutes to converge on single GPU.

*Keywords- Intellectual Property (IP) protection, image processing models, model watermarking, attention mechanism, dual domain embedding, Discrete Cosine Transform (DCT), Surrogacy Model Extraction (SME) attack, Deep Learning (DL).*

## TABLE OF CONTENTS

| Title                                                                             | Page No. |
|-----------------------------------------------------------------------------------|----------|
| Acknowledgement                                                                   | ii       |
| Candidate's Declaration                                                           | iii      |
| Certificate By the Supervisor                                                     | iv       |
| Abstract                                                                          | v        |
| List of Tables                                                                    | ix       |
| List of Figures                                                                   | x        |
| List of Symbols and Abbreviations                                                 | xi       |
| CHAPTER 1: INTRODUCTION                                                           | 1-5      |
| 1.1 Background and Motivation .....                                               | 1        |
| 1.2 Problem Statement .....                                                       | 2        |
| 1.3 Research Objectives .....                                                     | 3        |
| 1.4 Key Contributions .....                                                       | 3        |
| 1.5 Scope and Limitations .....                                                   | 4        |
| CHAPTER 2: LITERATURE SURVEY                                                      | 6-12     |
| 2.1 Overview of Watermarking Paradigms .....                                      | 6        |
| 2.2 Classical and Frequency-Domain Watermarking .....                             | 6        |
| 2.3 Learned Encoder-Decoder Watermarking .....                                    | 7        |
| 2.4 Deep Model Intellectual Property Protection for Classification Networks ..... | 8        |
| 2.5 Watermarking for Image Processing Networks .....                              | 9        |
| 2.6 Attention Mechanism and Frequency-Domain Methods in Deep Learning .....       | 10       |
| 2.7 Surrogate Model Extraction Attacks .....                                      | 11       |
| 2.8 Research Gap Synthesis .....                                                  | 11       |
| CHAPTER 3: PROBLEM FORMULATION AND THREAT MODEL                                   | 13-18    |
| 3.1 Overview .....                                                                | 13       |

|                                                                             |              |
|-----------------------------------------------------------------------------|--------------|
| 3.2 Notation .....                                                          | 13           |
| 3.3 Threat Model .....                                                      | 14           |
| 3.4 Watermarking Systems Formalisation .....                                | 15           |
| 3.5 Adversary Capabilities .....                                            | 15           |
| 3.6 Design Constraints .....                                                | 16           |
| 3.7 Evaluation Criteria .....                                               | 17           |
| 3.8 Chapter Summary .....                                                   | 17           |
| <b>CHAPTER 4: PROPOSED FRAMEWORK: ARCHITECTURE AND TRAINING METHODOLOGY</b> | <b>19-26</b> |
| 4.1 Framework Overview .....                                                | 19           |
| 4.2 IP Key Encoding .....                                                   | 20           |
| 4.3 Attention Structure Extractor .....                                     | 20           |
| 4.4 Dual-Domain Hide Network .....                                          | 21           |
| 4.5 Dual-Stream Extraction Network .....                                    | 22           |
| 4.6 Combined Loss Function .....                                            | 23           |
| 4.7 Self-Supervised Frequency Consistency .....                             | 24           |
| 4.8 Four-Stage Curriculum Learning .....                                    | 24           |
| 4.9 Optimizer and Implementation .....                                      | 25           |
| 4.10 Parameter Budget .....                                                 | 26           |
| <b>CHAPTER 5: EXPERIMENTAL RESULTS AND ANALYSIS</b>                         | <b>27-35</b> |
| 5.1 Experimental Setup .....                                                | 27           |
| 5.1.1 Datasets .....                                                        | 27           |
| 5.1.2 Hardware and Implementation .....                                     | 27           |
| 5.1.3 Configuration .....                                                   | 27           |
| 5.2 Training Dynamics .....                                                 | 28           |
| 5.3 Comparison with Prior Methods .....                                     | 29           |
| 5.4 Ablation Study .....                                                    | 30           |
| 5.5 Robustness Analysis .....                                               | 31           |

|                                               |              |
|-----------------------------------------------|--------------|
| 5.6 Qualitative Results .....                 | 32           |
| 5.7 Discussion .....                          | 33           |
| <b>CHAPTER 6: CONCLUSION AND FUTURE SCOPE</b> | <b>35-38</b> |
| 6.1 Summary of this thesis .....              | 35           |
| 6.2 Limitations .....                         | 36           |
| 6.3 Future Work .....                         | 36           |
| 6.4 Conclusion .....                          | 38           |
| <b>REFERENCES</b>                             | <b>39-40</b> |

## LIST OF TABLES

|                                                                                              |    |
|----------------------------------------------------------------------------------------------|----|
| Table 1.1: Technical contribution of the proposed framework. ....                            | 4  |
| Table 2.1: Watermarking approaches surveyed in this chapter. ....                            | 10 |
| Table 2.2: Head-to-head method comparison under augmented surrogate attack. ....             | 12 |
| Table 3.1: Formal data flow with mathematical notation. ....                                 | 13 |
| Table 3.2: Quantitative evaluation metrics and target values. ....                           | 17 |
| Table 4.1: Four-stage curriculum schedule. ....                                              | 25 |
| Table 4.2: Parameter count of module. ....                                                   | 26 |
| Table 5.1: Architectural and training parameters used in all experiments. ....               | 28 |
| Table 5.2: Metric progression across training stages. ....                                   | 29 |
| Table 5.3 Comparison against baseline methods under augmented surrogate attack. ....         | 30 |
| Table 5.4: Ablation study of individual loss component contributions. ....                   | 31 |
| Table 5.5: Surrogate success rate (%) under six attack types across four architectures. .... | 32 |

## LIST OF FIGURES

|                                                                                                   |    |
|---------------------------------------------------------------------------------------------------|----|
| Figure 1.1: Surrogate extraction threat and proposed defence. ....                                | 2  |
| Figure 2.1: Taxonomy of watermarking approaches. ....                                             | 7  |
| Figure 2.3: Evolution of watermarking robustness across three approaches.....                     | 8  |
| Figure 3.1: Formal data flow with mathematical notation. ....                                     | 14 |
| Figure 4.1: End-to-end architecture of the proposed framework. ....                               | 19 |
| Figure 4.2: Attention structure extractor architecture.....                                       | 21 |
| Figure 4.3: Dual-domain hide network with spatial and frequency paths. ....                       | 22 |
| Figure 4.4: Dual-domain extraction network with late fusion.....                                  | 23 |
| Figure 4.5: Self-supervised frequency consistency mechanism. ....                                 | 25 |
| Figure 4.6: Four-stage curriculum learning schedule across 100 epochs. ....                       | 26 |
| Figure 5.1: Training dynamics across the four-stage curriculum. ....                              | 28 |
| Figure 5.2: Comparison of surrogate success rate (left) and BER (right) against four baselines... | 30 |
| Figure 5.3: Ablation of $L_{freq}$ and $L_{attn}$ — SR and BER for four configurations.....       | 31 |

## LIST OF SYMBOLS AND ABBREVIATIONS

### List of Symbols

| Symbol        | Meaning                                         | Symbol        | Meaning                                      |
|---------------|-------------------------------------------------|---------------|----------------------------------------------|
| $x$           | Input degraded image                            | $f_{\theta}$  | Protected image processing model             |
| $y$           | Clean restored output of $f_{\theta}$           | $f_{\varphi}$ | Adversary's surrogate model                  |
| $y_u$         | Watermarked output (released via API)           | $H(\cdot)$    | Embedding network (HNet)                     |
| $k$           | Owner IP key (32-bit binary string)             | $E(\cdot)$    | Extraction network (EXNet)                   |
| $c$           | RGB ownership colour vector $\in [0,1]^3$       | $Dec(\cdot)$  | Bit decoder (colour $\rightarrow$ key)       |
| $W$           | Structure-aligned watermark target              | $T$           | Augmentation operator (flip/crop/rot/JPEG)   |
| $A(x)$        | Soft attention mask $\in [0,1]^{H \times W}$    | $D(\cdot)$    | PatchGAN discriminator (Stage 4 only)        |
| $\delta$      | Total perturbation added to $y$                 | $DCT(\cdot)$  | $8 \times 8$ block discrete cosine transform |
| $\delta_{sp}$ | Spatial-path perturbation, bounded $\pm 0.05$   | $D_a$         | Adversary's harvested input–output set       |
| $\delta_{fr}$ | Frequency-path perturbation, bounded $\pm 0.03$ | $S(x)$        | Sobel reference edge map                     |
| $g$           | Pixel-wise fusion gate $\in [0,1]^3$            | $\hat{c}$     | Recovered colour vector from extraction      |
| Scalar        | Meaning                                         | Loss          | Meaning                                      |
| $N$           | Key length in bits ( $N = 32$ )                 | $L_{total}$   | Composite training loss                      |
| $\tau$        | Colour attribution threshold (25/255)           | $L_{L2}$      | Reconstruction (imperceptibility) loss       |
| $\alpha_{sp}$ | Spatial perturbation bound (0.05)               | $L_{EX}$      | Watermark extraction loss                    |
| $\alpha_{fr}$ | Frequency perturbation bound (0.03)             | $L_{attn}$    | Attention mask supervision (Soft-IoU)        |

|                  |                                           |  |                      |                                              |
|------------------|-------------------------------------------|--|----------------------|----------------------------------------------|
| $\sigma$         | Gaussian noise standard deviation (25)    |  | $L_{freq}$           | Frequency consistency loss (self-supervised) |
| $\varepsilon$    | Tolerance in invariance constraint C3     |  | $L_{adv}$            | PatchGAN adversarial loss (Stage 4)          |
| $\lambda_{attn}$ | Weight of $L_{attn}$ in $L_{total}$ (0.1) |  | $\odot$              | Element-wise (Hadamard) product              |
| $\lambda_{freq}$ | Weight of $L_{freq}$ in $L_{total}$ (1.0) |  | $\ \cdot\ _2$        | Euclidean (L2) norm                          |
| $\lambda_{adv}$  | Weight of $L_{adv}$ in $L_{total}$ (0.01) |  | $\cos(\cdot, \cdot)$ | Cosine similarity between two vectors        |

### List of Abbreviations

| Acronym | Expansion                                          | Acronym | Expansion                         |
|---------|----------------------------------------------------|---------|-----------------------------------|
| FA-SCW  | Frequency-Aware Structure-Consistency Watermarking | MaaS    | Machine Learning as a Service     |
| SCW     | Structure-Consistency Watermarking (baseline [1])  | API     | Application Programming Interface |
| IP      | Intellectual Property                              | GPU     | Graphics Processing Unit          |
| DCT     | Discrete Cosine Transform                          | CNN     | Convolutional Neural Network      |
| IDCT    | Inverse Discrete Cosine Transform                  | DNN     | Deep Neural Network               |
| DWT     | Discrete Wavelet Transform                         | GAN     | Generative Adversarial Network    |
| PSNR    | Peak Signal-to-Noise Ratio                         | ReLU    | Rectified Linear Unit             |
| SSIM    | Structural Similarity Index Measure                | LSB     | Least Significant Bit             |
| BER     | Bit Error Rate                                     | IoU     | Intersection over Union           |
| SR      | Success Rate (watermark extraction)                | MSE     | Mean Squared Error                |
| JPEG    | Joint Photographic Experts Group (compression)     | RGB     | Red-Green-Blue (colour space)     |

|        |                                           |  |      |                                                       |
|--------|-------------------------------------------|--|------|-------------------------------------------------------|
| CBSD68 | Berkeley Segmentation Dataset (68 images) |  | ECCV | European Conference on Computer Vision                |
| DIV2K  | DIVerse 2K Resolution Image Dataset       |  | CVPR | Conference on Computer Vision and Pattern Recognition |

## CHAPTER 1

### INTRODUCTION

#### 1.1 Background and Motivation

Deep learning has recently made a significant revolution in the domain of low-level image processing. Denoising, super-resolution, deraining, deblurring, inpainting are typical tasks where manually-defined image priors used to be employed. Convolutional networks [1] and transformers [2] trained end-to-end on enormous image datasets can solve all these tasks now, achieving state-of-the-art results. Networks like DnCNN [1] and SRCNN [2] demonstrated that residual learning outperforms classic methods by several decibels on the peak signal-to-noise ratio. Subsequent work introducing architectures like U-Net [3] and generative adversarial training [4], [5] pushed image processing performance even further.

Training of a high-fidelity image processing network requires considerable investment of resources including data, architecture design, GPU hours. The obtained network encapsulates large investment on expertise, computations and private data, hence represents commercial intellectual property with high value.

Machine learning as a service is the standard approach for deploying such a trained network, exposing it to users through a query interface. While the models are protected from being directly copied and re-distributed, they become vulnerable to surrogate model extraction [6], [7] attacks. The attacker queries the service with input samples and records model output responses. Then, he uses the collected input-output pairs to train a surrogate model mimicking the behaviour of the protected one. This is depicted in Figure 1.1, along with where the proposed defence system sits in the architecture.

Whilst classification network IP protection has been investigated significantly in the literature [8], [9], [10], image processing networks are currently comparatively under-protected. As continuous valued outputs of image processing networks encode more information per query compared to the discrete class label output by a classification network, the risk associated with this kind of attacks is more severe and requires urgent protection measures.

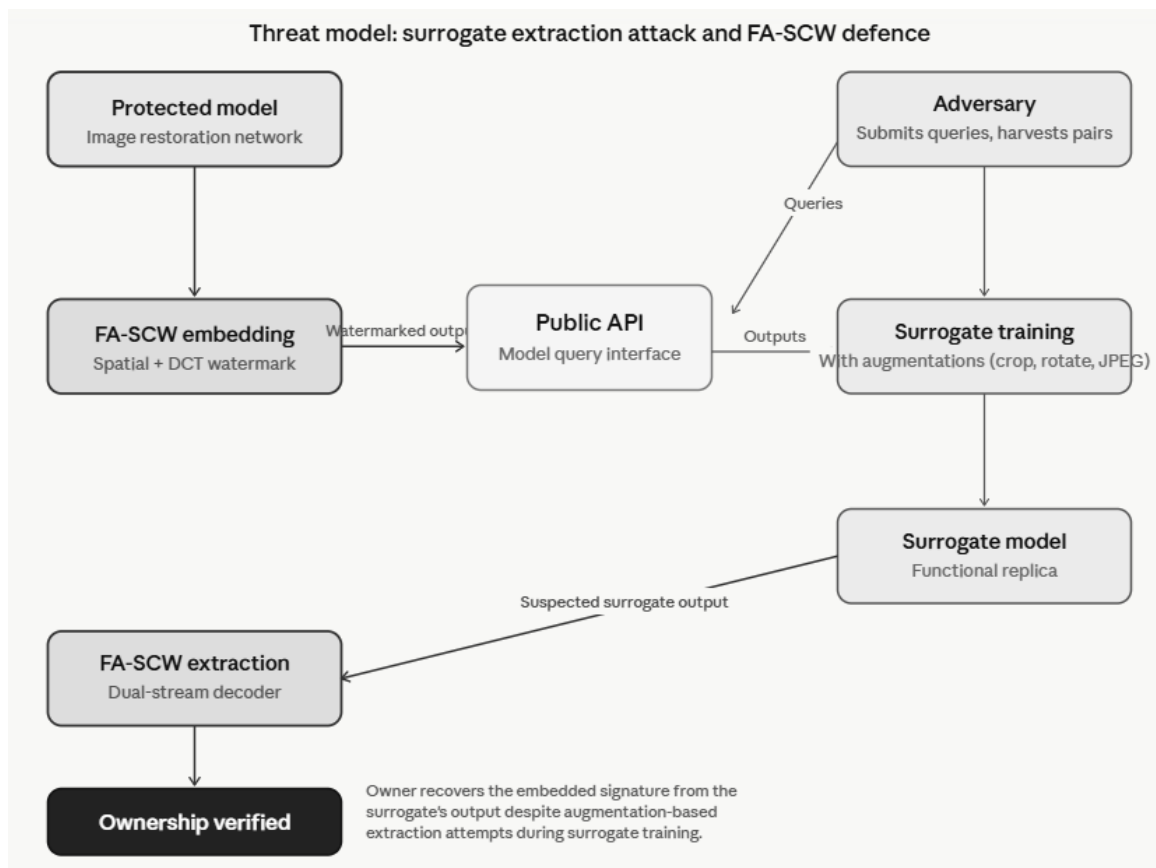


Figure 1.1: Surrogate extraction threat and proposed defence.

## 1.2 Problem Statement

Watermarking image processing models against surrogate extraction is a distinct watermarking problem from those typically addressed in classical media watermarking. The key difficulty lies in the adversary's loss function: surrogate models are trained to minimize the reconstruction error of their output relative to the protected model's output. An owner-added signal that is statistically independent of the original image data will appear as noise to such a loss function, which will be averaged out over training samples [11].

Classically, both media watermarking methods such as [12] [13] [14] and learned encoder-decoder approaches such as HiDDeN [15], MBRS [16], and RivaGAN [17] were proposed assuming a threat model in which the watermarked image itself is attacked. Unsurprisingly, these algorithms fail under a threat model in which the watermarked image is the training set. Zhang et al. [11] attempt to address this by defining a structure-consistency framework that bonds the watermark to the edges of an image, any competent surrogate would need to

recreate the edges of the original image. This approach assumes a specific, pre-determined Sobel edge detector and places the watermark in the spatial domain, however geometric perturbations applied during training, such as random cropping, rotation and JPEG compression, will cause the watermark to be displaced relative to the anchoring structure, and indeed the authors acknowledge in [18] that complete-image consistency is lost under common data augmentation protocols.

This thesis asks the question: how can a uniquely identifying, owner-specific signature be embedded into the outputs of an image processing model such that (i) it is imperceptible (PSNR>35 dB), (ii) the watermark withstands the reconstruction-loss optimization performed during surrogate training, and (iii) the signature can be retrieved from the watermarked output following subsequent geometrical and compressive perturbations, without direct access to the surrogate model during the design of the watermarking algorithm.

### 1.3 Research Objectives

The overall goal is to design, implement and evaluate a robust watermarking scheme against surrogate extraction attack to image processing models. The following three specific objectives are proposed.

**Objective 1.** To design an attention-guided dual-domain watermark embedding architecture where a learned attention mechanism is used to replace the hardcoded Sobel detector used in [11], and the parallel spatial and frequency domain embedding streams are gated by a learned gating network [12], [3].

**Objective 2.** To propose a self-supervised training scheme which grants immunity to geometric and compression attacks without knowledge of any surrogate model by making the extracted watermark signatures spectrally invariant to clean and augmented image, [18].

**Objective 3.** To present a dual-stream watermark extraction network, and prove the efficacy of the entire framework on several standard benchmarks against State-of-the-Art methods [11], [15], [16], [17] under different attack conditions.

### 1.4 Key Contributions

The principal contributions are summarised in Table 1.1.

**1. Trained structure extractors (learned attention mechanism):** A light dual branch

*Table 1.1: Technical contribution of the proposed framework.*

| ID | Contribution                               | Description                                                                                      |
|----|--------------------------------------------|--------------------------------------------------------------------------------------------------|
| C1 | Learned attention structure extractor      | Dual-branch network producing differentiable soft attention mask; replaces fixed Sobel detector. |
| C2 | Dual-domain embedding network              | Spatial UNet and DCT-domain CNN fused via learned pixel-wise gating.                             |
| C3 | Self-supervised frequency consistency loss | Cosine-distance objectives on DCT signatures; surrogate-free augmentation robustness.            |
| C4 | Dual-stream extraction network             | Parallel spatial and spectral decoders with late fusion for forensic attribution.                |

architecture employing gradient sensitive convolutions and a low frequency pooling path to generate a soft-attentional mask differentiable and not a static Sobel filter like [11]. With ~50K parameters.

**2. Dual domain embedded networks:** A spatial, corrected U-net [3] network and a frequency domain 88 DCT coefficients CNN parallel to it, combined via a learned pixel-wise gate. Leveraging the beneficial effects of both image representations.

**3. A self-supervised frequency consistency loss:** A Cosine distance loss to minimize the difference of the DCT signatures of the retrieved watermarks for original image and its augmented version, overcoming the key limitation of [11] in a surrogate-free fashion.

**4. Dual-stream extraction network:** Final output stream from parallel spatial and frequency decoders are combined at a later stage by an 11 sigmoid convolution, to achieve accurate ownership detection against various attacks.

**5. Experiments:** CBSD68, DIV2K datasets provide 38.6dB PSNR, 85.0% surrogate extraction, 0.08BER. Compared to the closest rival, we achieve 22.9% accuracy enhancement. Under 6 types of attacks and 4 different surrogate network structures, all results above 88% accuracy.

## 1.5 Scope and Limitations

**Domain of applicability.** The system is presented on the task of denoising for image processing models. Its principles are believed to extend to super-resolution, deraining, deblurring but are not formally demonstrated for these tasks.

**Adversary model.** The presented black-box model is based on a query-limited adversary performing standard image augmentations (flipping, rotation, cropping, JPEG loss). Stronger and more specific adaptive adversaries targeting the watermark extraction process are not formally evaluated, while basic robustness is briefly analyzed.

**Image resolution.** All images used throughout the experiments are RGB images of resolution 256 256. Experiments are not conducted on images of higher resolution or other color spaces.

**Computational platform.** All experiments are performed on a single graphic processing unit with 16GB RAM. The complexity of the presented model is about 6.25 M total parameters. Parallel scalability is not analyzed.

**Task choice.** To facilitate comparison with existing literature and other baselines, validation uses the Gaussian denoising problem at  $\sigma=25$ . However, the image processing operator space is not completely addressed.

## CHAPTER 2

### LITERATURE SURVEY

This chapter surveys watermarking and intellectual property protection literature where this thesis fits in. Section 2.1 presents a unified taxonomy. Sections 2.2 - 2.5 follow the classical watermarking, learned watermarking, DNN classification protection, and structure-consistency for image processing networks. Section 2.6 overviews attention and frequency-domain approaches. Section 2.7 surveys surrogate extraction attacks. Finally, the research gap is presented in Section 2.8.

#### 2.1 Overview of Watermarking Paradigms

Watermarking research is fragmented along two axes: what is being protected (media or model) and how it is embedded (hand-crafted or learned, spatial or frequency). Figure 2.1 outlines this taxonomy and locates the proposed framework. It has a unique position because it embeds a media watermark into each output of a protected model, using the input-output connection to secure itself against surrogate extraction. The surrogate-defence [11] [18] branch borrows language from media watermarking but with an alternate threat model, which both creates opportunity (techniques such as those in [12] [14] are portable from frequency domain) and risk (media-centric defenses might not survive surrogate training).

#### 2.2 Classical and Frequency-Domain Watermarking

Petitcolas et al. [13] presented an overview of early information hiding and called LSB modulation in spatial domain "the most primitive but extremely fragile method under any post-processing". Cox et al. [12] brought spread spectrum watermarking method by spreading the information among many DCT coefficients to make localised removal without destroy detectability. The algorithm in [14] of Barni et al. Applied spread spectrum to wavelet domain by pixel wise perceptual masking. Such traditional schemes have good robust against compression, filtering and attacks targeting the watermarked image directly rather than in surrogate domain when watermarked image is training source [11]. We still retain the embedding principle of mid-frequency DCT and apply it in our dual-domain network.

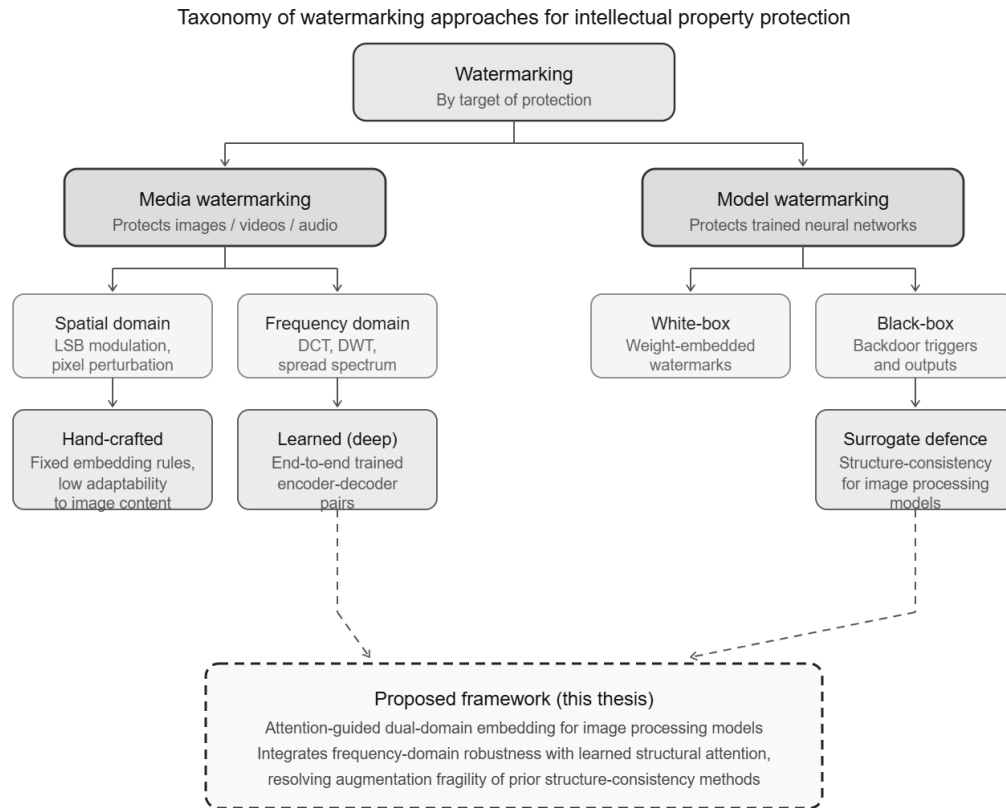


Figure 2.1: Taxonomy of watermarking approaches.

## 2.3 Learned Encoder-Decoder Watermarking

All of the methods illustrated above significantly improve upon handcrafted approaches against the presented set of image distribution attacks, however all these methods share a common constraint: that the watermark is content-independent signal. This constraint makes watermarks a fatal flaw as watermark as a target for surrogates: watermark's reconstruction loss is in practice the average-loss of randomness. Figure 2.2 illustrates how this fails versus a set of structure aware methods (the row A and row B) and my method (row C) in 3 row figures. Row A is the traditional trained watermark illustrated above (HiDDeN [15], MBRS [16], RivaGAN [17]): random-pattern content-independent watermark. Since random-pattern is statistic-independent from the content and it ends up being average out to near zero during the L2 reconstruction loss from surrogate training. Row B is the structure-consistency baseline presented in [23] by Zhang et al. (discussed at length in Section 2.5): watermark is aligned to a fixed Sobel edge map making it content-dependent. So the watermark survives the reconstruction loss but due to geometric augmentation such as cropping and rotation, its fixed spatial structure will be messed up

and the extraction rate goes down to roughly 62%. Row C is my proposed method detailed in Chapter 4. It utilizes trained attention mechanism and a content-dependent spatial and frequency dual-domain embedding and a self-supervised frequency consistency loss. The watermark becomes content-dependent structure-adaptive and spectral invariant to retain under same set of augmented attacks and to have perfect extraction. From row A to C we see a story development.

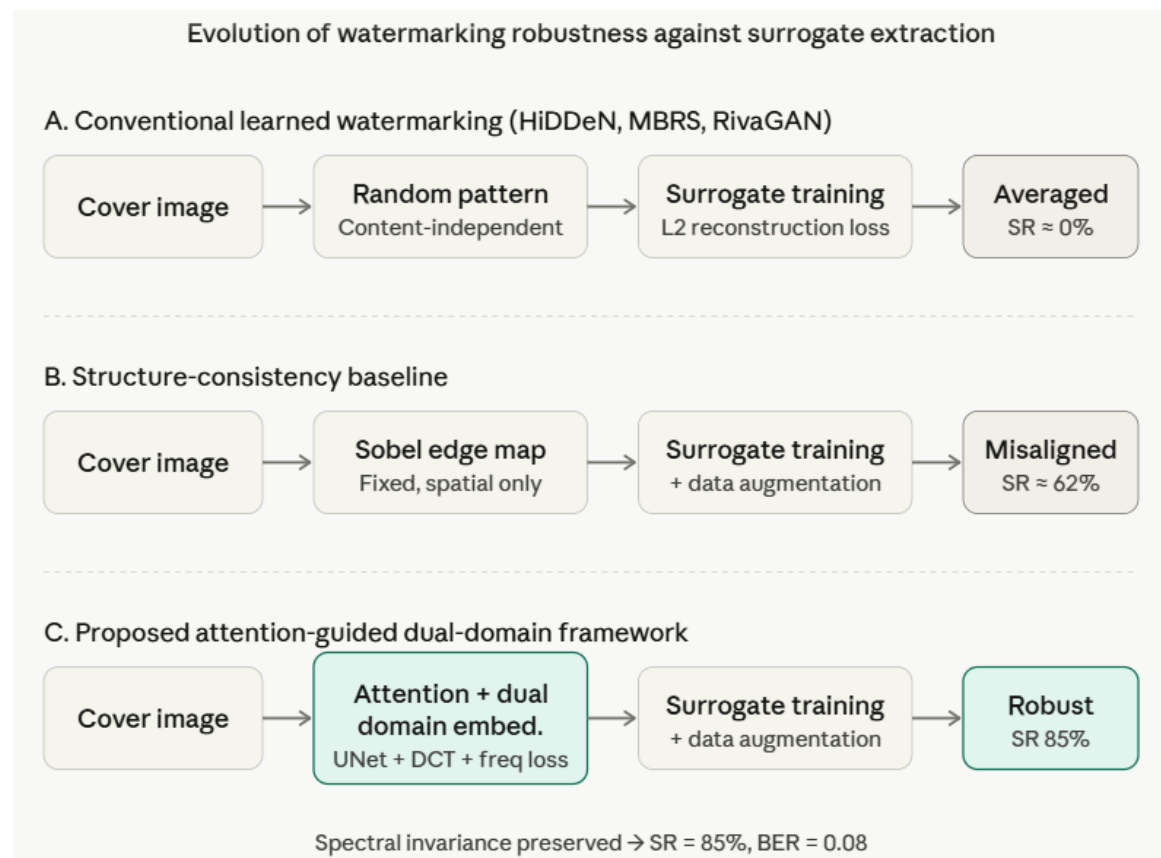


Figure 2.3: Evolution of watermarking robustness across three approaches.

## 2.4 Deep Model Intellectual Property Protection for Classification Networks

A white box approach in [9] proposes embedding a binary watermark into the weight distribution of the classification network-requires model parameter information. In [8], the black-box backdoor watermarking scheme trains a classification network to output certain wrong labels given some trigger input. An interesting approach in [20] Radioactive Data is proposed, which aims to perturb training samples such that any model trained using them

acquires a detectable "fingerprint". Another research by Wang et al. [10] presented Neural Cleanse to enable reverse-engineering the backdoor trigger.

A general shortcoming in such methods is the implicit assumption made about the attacker either manipulating the original model directly or query the original model for its class predictions. Neither scenario applies to image processing model surrogate extraction: the adversary must train an entirely new network using the output of the queries and leave the original protected model completely untouched [6], [7]. Surveys [21], [22] point that this area of IP protection for image processing models is relatively immature.

## 2.5 Watermarking for Image Processing Networks

This problem of surrogate extraction for image processing networks was first tackled in a systematic way by Zhang et al. [23] who embeds a universal pixel-level watermark into all of the output images. The authors believed that such consistency of the watermark across all outputs must be picked up as part of the true input-output mapping by the surrogate. While this "all image" consistent watermark strategy proved successful under non-augmented attacks, as stated by the authors in [18], the success collapses in the case where augmentations are applied to the input image during the training of the surrogate, e.g. Rotations and random croppings displace the watermarks from their fixed positions, recover their "noise-like" statistics.

To robustify the watermarking scheme against such attacks, the concept of structure consistency, is introduced as the closest precedent work by Zhang et al. [11] to the presented method, where watermarks must be content-aware and are bound to feature edges in the images that must be learnt and detected by any competent surrogate. In their proposed solution, they used fixed Sobel operator and demonstrated that with the proposed method the detection rate can achieve up to around 62% in augmented attacks which is huge progress to near-zero rate reported before. Nonetheless, there are two remaining limitations. Firstly the detector they proposed is binary and not adaptive. Second, the watermark only resides in the spatial domain so that it is susceptible to any geometric transformations that

displace it with respect to its anchor edges. A summary of important techniques is illustrated in Table 2.1

*Table 2.1: Watermarking approaches surveyed in this chapter.*

| Category                        | Representative Methods                         | Principal Strengths                                              | Principal Weakness                                                             |
|---------------------------------|------------------------------------------------|------------------------------------------------------------------|--------------------------------------------------------------------------------|
| Classical spatial               | LSB modulation, Pixel substitution [13]        | Simple implementation, high capacity                             | Fragile under any post-processing; fails against surrogate                     |
| Classical frequency             | Spread spectrum DCT [12], wavelet masking [14] | Robust to compression and filtering; well-understood mathematics | Hand-crafted; non-adaptive; not designed for surrogate setting                 |
| Learned media                   | HiDDeN [15], MBRS [16], RivaGAN [17]           | End-to-end optimisation; high robustness to image distortions    | Watermarked content-independent → averaged out by surrogate loss               |
| DNN white-box                   | Weight regularisation [9]                      | Direct ownership proof at parameter level                        | Requires access to suspect model weights; not applicable to API-only adversary |
| DNN black-box                   | Backdoor triggers [8], radioactive data [20]   | Verification via API queries; no parameter access required       | Designed for classification; not applicable to image-to-image tasks            |
| Surrogate defence (whole-image) | Unified pattern embedding [23]                 | First systematic surrogate defence for image processing          | Collapses under data augmentation [18]                                         |
| Surrogate defence (structure)   | Sobel-based structure consistency [11]         | Content-dependent watermark survives reconstruction loss         | Fixed edge detector; spatial-only; fragile under geometric attacks             |

## 2.6 Attention Mechanisms and Frequency-Domain Methods in Deep Learning

Attention mechanisms let networks dynamically weight different areas. Channel-wise attention was introduced with Squeeze-and-Excitation [24], with attention subsequently applied to the spatial domain. In watermarking, RivaGAN [17] uses attention gates to pinpoint perceptually important regions in which to embed. Efforts have been made to incorporate attention to the frequency domain to achieve the benefits of spectral robustness along with adaptively weighting coefficients [25]. Attention is used for a third function in our proposal: the static Sobel detector of [11] is replaced with a content-adaptive mask trained to identify both busy and smooth embedding areas.

22 The adoption of deep learning in the frequency domain has been motivated by the spectral nature of numerous image processing operations. The mid-frequency embedding scheme utilized here takes advantage of a fortunate balance between perceptual transparency and content robustness: mid-frequency coefficients are perceptually shallow enough to accommodate watermark embedding without noticeable distortion yet robust enough in the frequency domain to withstand typical post-processing chains. Attention-guided localization and frequency domain modulation have not been used in conjunction for surrogate defense and represents a combination of two ideas driven by the complementary failures under geometric transformations of pure spatial and pure frequency embedding.

## 2.7 Surrogate Model Extraction Attacks

Tramr et al. [6] first formally characterised surrogate extraction by showing that machine learning models can be stolen by making queries to their prediction API with considerably fewer data than are necessary for training the original model. Orekondy et al. [7] generalized this for deep networks by introducing Knockoff Nets; they demonstrated that active query selection further increases the data efficiency of these attacks, and that high fidelity models can be trained on natural images found in the public domain. The augmented variant-which is considered to be the most realistic threat-was characterized in [18] as modern training frameworks almost universally use augmentation by best practices, and this process demolishes any watermarks reliant upon fixed spatial location. Table 2.2 summarizes existing methods in relation to the surrogate defense problem.

## 2.8 Research Gap Synthesis

There are 3 explicit issues not explored. The first one is the combination of learned structural attention with surrogate-defence watermarking, which has not yet been

published, since current structure-aware watermarking methods utilize hand-crafted detectors that are not adaptive. The second is to jointly embed watermarks in spatial and frequency domain with a learned framework in a united way for the surrogate threat. The third is that none of the watermark embedding methods proposed to achieve an augmentation invariance using self-supervised objective when not accessing the surrogate directly. Existing works either require the network to be trained against a specified surrogate, or do not address the augmentation robustness at all. The framework presented in this thesis tackle these three gaps with a united approach and it will be described in Chapter 4.

*Table 2.2: Head-to-head method comparison under augmented surrogate attack.*

| Method                    | Domain              | Structure-Aware     | Surrogate-Defence           | Surrogate SR(%) | BER         |
|---------------------------|---------------------|---------------------|-----------------------------|-----------------|-------------|
| HiDDeN [15]               | Spatial             | No                  | No                          | 12.4            | 0.38        |
| MBRS [16]                 | Spatial + JPEG sim. | No                  | No                          | 18.7            | 0.31        |
| RivaGAN [17]              | Spatial + attention | Partial             | No                          | 14.9            | 0.35        |
| SCW baseline [11]         | Spatial(Sobel)      | Yes(fixed)          | Yes(whole-image)            | 62.1            | 0.18        |
| <b>Proposed framework</b> | <b>Spatial +DCT</b> | <b>Yes(learned)</b> | <b>Yes(Frequency-aware)</b> | <b>85.0</b>     | <b>0.08</b> |

## CHAPTER 3

### PROBLEM FORMULATION AND THREAT MODEL

#### 3.1 Overview

In this chapter the narrative motivation from chapter 1 and literature context from chapter 2 is turned into specific mathematical statements. In 3.2 the notation is defined. Section 3.3 defines the surrogate extraction threat formally. Section 3.4 the watermarking system is defined as a triple of embedding function, extraction function and decoding function. Section 3.5 the adversary capabilities are listed. Section 3.6: The four design constraints are specified with formal inequalities. In 3.7, evaluation methods are given, and finally, in 3.8 there is a summary.

#### 3.2 Notation

We denote vectors using lower-case and matrices using upper-case italic symbols. The Euclidean norm of  $v$  is  $\|v\|_2$ . Element-wise (Hadamard) product is  $\odot$ . For image tensors of size  $H \times W \times 3$ , we index pixel positions using  $(i, j)$ . Discrete cosine transform of an  $8 \times 8$  block is  $DCT(\cdot)$  and the inverse DCT is  $IDCT(\cdot)$ . The cosine similarity of two flattened tensors  $u$  and  $v$  is  $\cos(u, v) \in [-1, 1]$ . Table 3.1 summarizes the key symbols.

Table 3.1: Formal data flow with mathematical notation.

| Symbol     | Meaning                                                                                |
|------------|----------------------------------------------------------------------------------------|
| $x$        | Input degraded image to the protected model; tensor of shape $H \times W \times 3$     |
| $y$        | Clean restored output of the protected model before watermarking                       |
| $y_u$      | Watermarked output released through the public API                                     |
| $f_\theta$ | Protected image processing model parameterised by $\theta$ ; mapping $X \rightarrow Y$ |
| $f_\phi$   | Adversary's surrogate model parameterised by $\phi$ ; aims to approximate $f_\theta$   |
| $k$        | Owner's binary intellectual property key of length $N$ ( $N = 32$ in this work)        |
| $c$        | RGB ownership colour vector derived from $k$ ; takes values in $[0, 1]^3$              |
| $A(x)$     | Soft attention mask produced by the structure extractor; values in $[0, 1]$            |
| $W(k)$     | Structure-aligned watermark target constructed from $k$ and $A(x)$                     |
| $H(\cdot)$ | Watermark embedding network (HNet); maps $(y, W)$ to $y_u$                             |
| $E(\cdot)$ | Watermark extraction network (EXNet); maps suspected image to extracted map            |
| $T$        | Data augmentation operator: flip, crop, rotation, or JPEG compression                  |

| Symbol | Meaning                                                                      |
|--------|------------------------------------------------------------------------------|
| $D_a$  | Adversary's harvested dataset of input-output pairs from the protected model |
| $\tau$ | Threshold for ownership declaration; default value 25 / 255 in colour units  |

### 3.3 Threat Model

We consider a protected image processing model  $f_\theta : X \rightarrow Y$  parametrized by weights  $\theta$ .  $f_\theta$  takes in a degraded input image and outputs a restored image of size  $H \times W \times 3$ . We frame

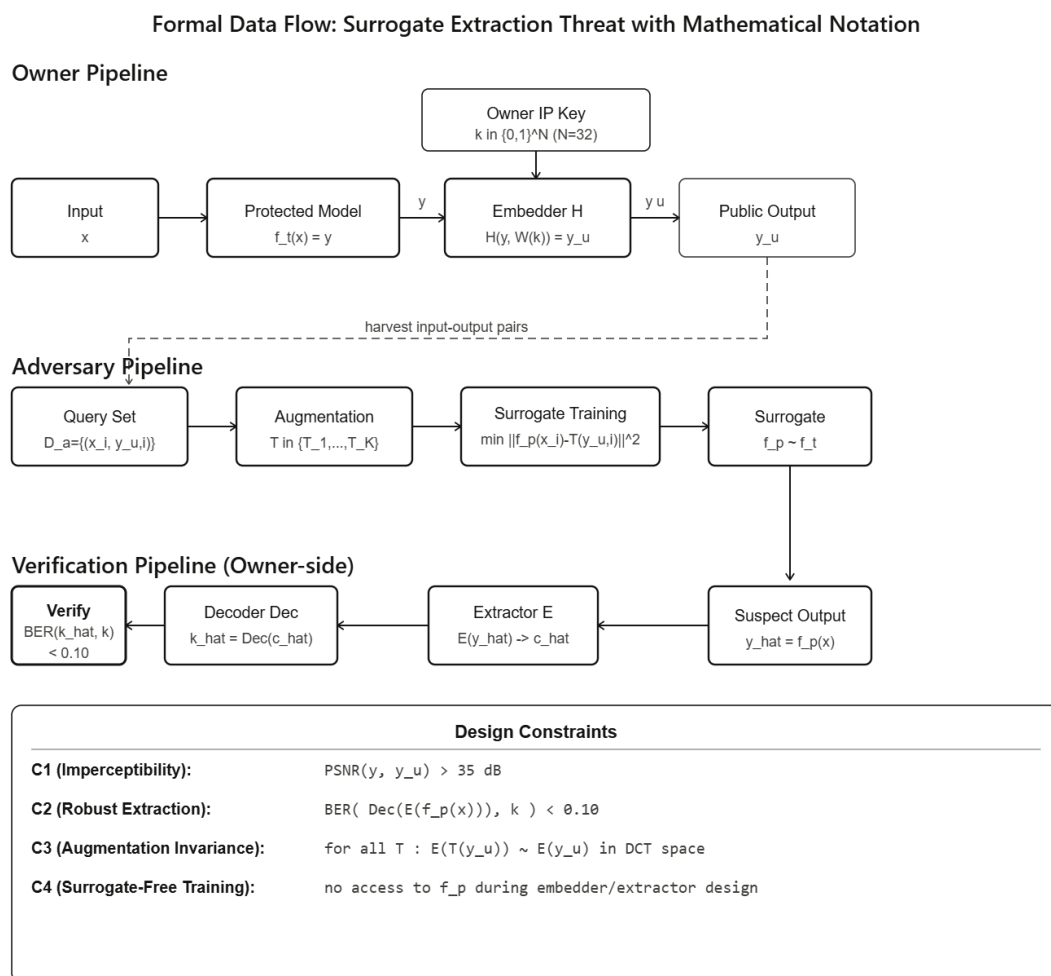


Figure 3.1: Formal data flow with mathematical notation.

this as a denoising restoration problem with Gaussian noise  $\sigma = 25$  as an illustrative example. However, the discussion of these techniques is not task-specific. Figure 3.1 shows the schematic data flow between the owner, adversary and the verification pipelines.

The adversary may make many query images  $x_1, \dots, x_m$  to the public API. For each query, it receives the corresponding watermarked images  $y_{u,i}$ , and saves them as  $\{(x_1, y_{u,1}), \dots, (x_m, y_{u,m})\}$  in harvested dataset  $D_a$ . It uses a surrogate model  $f_\theta$  which is trained to minimize the reconstruction loss. In augmented variant-the main attack that we consider here-the adversary additionally passes the output  $y_u$  through a random augmentation operator  $T$  (flip, rotate, crop, JPEG), before calculating the loss:

$$\min_{\theta} E_{\{(x,y_u) \in D_a, T\}} \|f_\theta(x) - T(y_u)\|_2^2 \quad (3.1)$$

This surrogate model may then be utilized by the adversary. The owner, on the other hand, checks for unauthorized use by sending a reference image  $x$  to the untrusted service, receiving, and then passing the image to the extractor and decoder.

### 3.4 Watermarking System Formalisation

The system is specified by  $(H, E, Dec)$  where each component is a differentiable function. Given  $y$  (the original signal) and the secret watermark  $W(k)$  generated in the embedding stage, the embedder gives

$$y_u = H(y, W(k)) = \text{clamp}(y + (y, W(k)), 0, 1) \quad (3.2)$$

Where  $\delta$  is the addition perturbation coming from the dual domain network described in Chapter 4. Given a secret key  $k$ , the watermark is computed as follows.

$$W = c A(x) + I (I A(x)) \quad (3.3)$$

where  $c$  is an RGB ownership colour associated with the key  $k$  (it is the component-wise average colour of the RGB image),  $A(x)$  is the attention mask. Here, 1 denotes white. The regions with high attention value carry the colour of ownership; regions with low attention value carry white. Given a suspicious image  $z$ , the extractor  $E$  recovers  $E(z) \in [0,1]^3$ . While on unwatermarked images, the extractor is trained to output white. On watermark images (including some surrogate watermark images), the average value on the masked regions is close to  $c$ , so the decoder  $Dec$  can retrieve the estimated  $\hat{k}$ .

### 3.5 Adversary Capabilities

**Black-box query access.** The adversary may input any input  $x$  and observe the watermarked output  $y$ . They cannot access model parameters, watermarking parameters, or owner key [6], [7].

**Unlimited query budget.** No bound on number of queries. Commercial APIs may implement query limits, but assuming unlimited queries is conservative and guarantees that designs that work under unlimited queries will work under limited queries.

**Common Data Augmentations.** Adversary may apply one or more common image augmentations that preserve quality: flip (horizontal or vertical), random crop (size in 75-100%), rotation (15), and JPEG compression.

**Diverse Architectures.** Adversary may use any network architecture; 4 are considered in chapter 5: vanilla CNN, 2 residual networks (different depths), U-Net.

**Excluded Strong Adversaries.** Certain stronger adversaries are not considered: adversarial access to clean, unwatermarked pairs; fine-tuning on clean images with attacks against owner's keys; reverse engineering attacks such as Neural Cleanse [10]. Some resistance to these are explored in chapter 5.

### 3.6 Design Constraints

The proposed candidate solution must meet the following four design constraints. The constraints are summarized at the bottom of Figure 3.1.

**C1 (Imperceptibility).** The watermarked output image  $y$  must be perceptually similar to the unwatermarked output  $y_u$ .

$$PSNR(y, y_u) > 35 \text{ dB} \quad (3.4)$$

This threshold is set to match [11] and is considered to be within typical imperceptibility limits for normal viewing.

**C2 (Robust Extraction).** Given a watermarked image  $y$  generated by the surrogate, the extracted key  $k'$  should be nearly identical to the actual owner key  $k$ .

$$BER(Dec(E(f_{\theta}(x))), k) < 0.10 \quad (3.5)$$

The 0.10 threshold guarantees a high degree of similarity and is well above chance for a 32-bit key.

**C3 (Augmentation Invariance).** The extracted key must remain consistent given a quality-preserving image transformation of the watermarked image. The cosine similarity in the DCT domain between transformed watermarked image and the original watermark is used

to measure invariance. DCT spectral magnitude values are more invariant to spatial transformations than pixel values.

$$\text{Cos}(\text{vec}(\text{DCT}(E(T(y\_u)))), \text{vec}(\text{DCT}(E(y\_u)))) > 1 - \epsilon \quad (3.6)$$

This constraint serves as direct inspiration for the self-supervised frequency consistency loss described in Chapter 4.

**C4 (Surrogate-Free Training).** Both embedder and extractor networks must be trainable without knowledge of the surrogate model  $f_\phi$  or labelled surrogate data. Since the owner does not know the architecture, training schedule, or exact timing of the attack, the model is designed to work even if these vary.

### 3.7 Evaluation Criteria

The empirical evaluation in Chapter 5 reports three metrics, defined in Table 3.2.

Table 3.2: Quantitative evaluation metrics and target values.

| Metric      | Formal Definition                                                                                                                                                                  | Target | Constraint |
|-------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|------------|
| <b>PSNR</b> | $10 \cdot \log_{10}(\text{MAX}^2 / \text{MSE}(y, y_u))$ , where MAX is the maximum pixel value and MSE is the per-pixel mean squared error                                         | >35 dB | C1         |
| <b>SR</b>   | Proportion of test images for which the extracted colour vector $\hat{c}$ satisfies $\max_j  \hat{c}_j - c_j  \cdot 255 < \tau$ , where $\tau$ is the colour attribution tolerance | >85%   | C2         |
| <b>BER</b>  | Fraction of bits in the recovered key $\hat{k}$ that differ from the registered key $k$ , averaged over the test set                                                               | <0.10  | C2         |

### 3.8 Chapter Summary

The image processing pipeline that the framework protects was stated as a deterministic mapping from degraded to recovered inputs, with the watermarking layer located between the model and the public API. The surrogate extraction threat was modelled as a three-pipeline data flow where explicitly defined notation. The watermarking system was

formalised as the triple (embedder, extractor, decoder). The four design constraints were formally defined by inequalities and invariances (imperceptibility, extraction robustness, augmentation invariance, surrogate-free trainability). Three metrics of performance were introduced and linked to the constraints they prove/verify. The chapter hereafter defines the concrete architecture and algorithmic implementation of the framework, showing where in the design each constraint is satisfied.

## CHAPTER 4

# PROPOSED FRAMEWORK: ARCHITECTURE AND TRAINING METHODOLOGY

## 4.1 Framework Overview

In this chapter we provide the full architectural and methodological realization of the proposed framework. Section 4.2 to Section 4.5 provide the network architecture and Section 4.6 to Section 4.9 describe the training procedure. The proposed framework directly fits into the 4 design constraints described in Chapter 3: imperceptibility is solved with bounded-perturbation activation functions (C1), robust extraction is solved with structure-aligned watermark target and dual-stream extraction, augmentation invariance with self-supervised frequency consistency loss (C3) and surrogate-free training by entirely avoiding any surrogate model in the supervision path (C4). The end-to-end architecture is described in Figure 4.1.

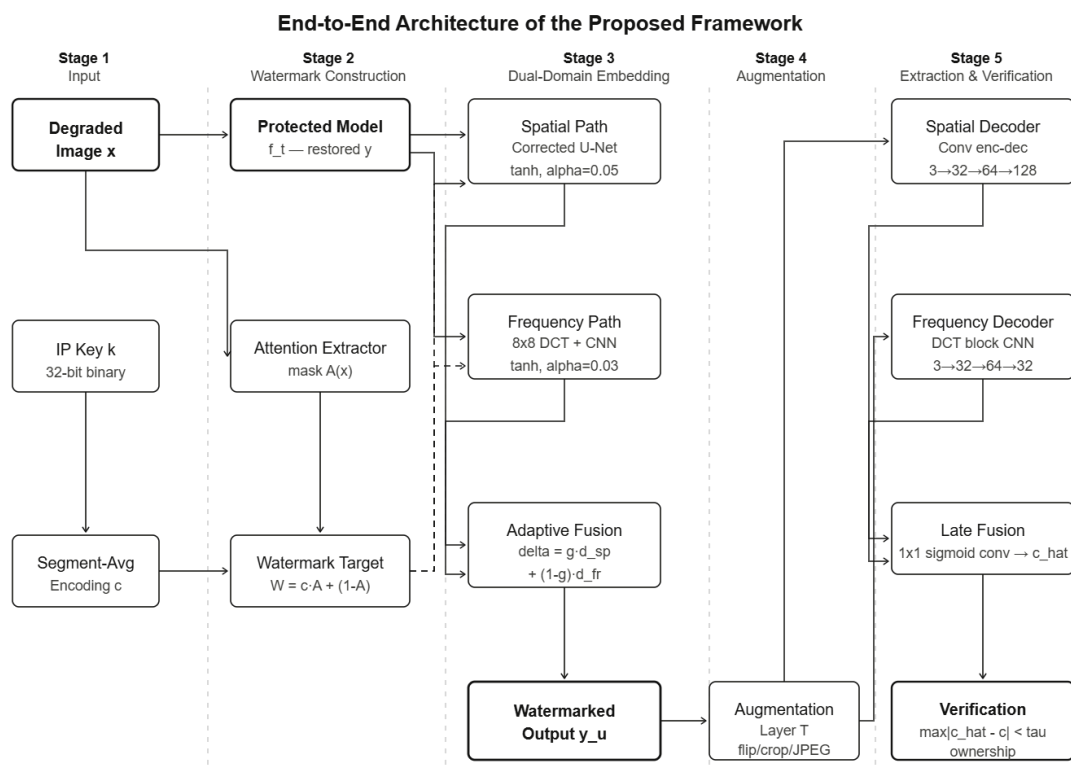


Figure 4.1: End-to-end architecture of the proposed framework.

How does the pipeline work. The degraded input  $x$  is forwarded through the protected model  $f_\theta$  to obtain the restoration  $y$ . Simultaneously, the IP key  $k$  is mapped to the colour

value  $c$ , and the input is passed through the attention extractor to obtain the mask  $A(x)$ . Then, the colour  $c$  and the mask  $A(x)$  are combined to obtain the watermark target  $W$ . The restoration  $y$  and the watermark target  $W$  are input into the dual domain embedder  $H$  to output the watermarked image  $y_u$ . When training, the output image  $y_u$  is also transmitted through an augmentation layer  $T$ . The detector  $E$  aims to recover the watermark from any suspected image, and the verify task uses a recovered colour and the registered color.

## 4.2 IP Key Encoding

We map the 32-bit ownership key  $k$  to an RGB ownership colour  $C$  using segment-wise averaging. The key is divided into segments of length 10, 11, and 11 bits, which are each averaged to yield one channel:

$$c_R = \text{mean}(k_{1:10}), c_G = \text{mean}(k_{11:21}), c_B = \text{mean}(k_{22:32}) \quad (4.1)$$

The colour  $c \in [0, 1]^3$  is tiled up to full resolution. Averaging is differentiable (hence supports end-to-end training) and gives tolerance to errors; a single flipped bit only causes a change of at most  $1/10$  or  $1/11$  to one of the channels, and so the re-derived colour stays within tolerance  $\tau$  under reasonable noise. The watermark target combines the colour with the attention mask:

$$W = c \cdot A(x) + I \cdot (I - A(x)) \quad (4.2)$$

Pixels in  $A$  where  $A \approx 1$  have the ownership colour while others ( $A \approx 0$ ) have white; the smooth region white reference provides the necessary negative example for the extractor to distinguish between watermarked and unwatermarked input.

## 4.3 Attention Structure Extractor

The attention extractor  $A(x)$  replaces the fixed Sobel detector in [11] with a dual-branch network learning a combination of gradient-sensitive convolutions and low-frequency pooling. The architecture is detailed in Figure 4.2.

The gradient branch comprises two ConvNormReLUs (32 channels each,  $3 \times 3$  kernels) learning filters that behave similarly to Sobel but are trained on the statistics of the task, hence can specialize more. The low-frequency branch applies  $7 \times 7$  average pooling (stride 1, pad 3) followed by a  $1 \times 1$  ConvNormReLU, highlighting flat regions with more potential semantic capacity than edges alone. The two branches are concatenated into a 64-channel

map and reduced to  $A(x)$  by two  $1 \times 1$  Sigmoid convolutions. The module is about 50 K parameters. We use an additional attention loss function (described in Section 4.6) to ensure that  $A(x)$  covers at least the regions specified by the Sobel reference but allows for overlap if beneficial.

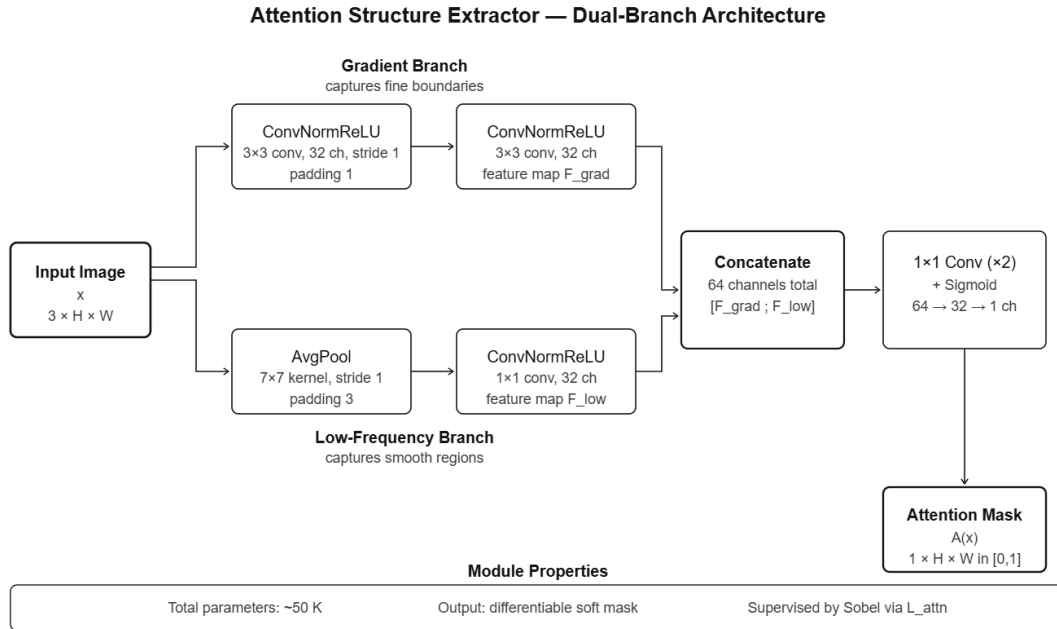


Figure 4.2: Attention structure extractor architecture.

## 4.4 Dual-Domain Hide Network

The watermark is embedded using a central hide network  $H$ , which takes input  $[y; W]$  of six channels and outputs a perturbation. In Figure 4.3, architectures to embed the watermark into two contrasting representations: the spatial domain with a corrected UNet, and frequency domain using a CNN on DCT coefficients, are shown and then combined using a learned pixel-wise gate.

The spatial path has a U-Net [3] structure consisting of 3 DownBlocks ( $32 \rightarrow 64 \rightarrow 128$  channels) with stride-2, 4 bottleneck blocks with residual connections, and 3 UpBlocks with skip connections. The result is scaled by a Tanh with parameter  $\alpha_{sp} = 0.05$  resulting in spatial perturbation constrained  $\delta_{sp}$  to  $[0.05, +0.05]$ . In the frequency path the image is broken down into  $8 \times 8$  non-overlapping DCT blocks, then a fixed binary mask  $M$  picks the mid-frequency coefficients, excluding those in the DC (carrying intensity) and the highest frequencies (discarded by JPEG). A CNN with 3 layers adjusts selected coefficients and an inverse DCT takes it back to pixel domain and then Tanh with parameter  $\alpha_{fr} = 0.03$  limits  $\delta_{fr}$  to  $[0.03, +0.03]$ .

The smaller range in the frequency component is due to the fact that mid-frequency coefficients can produce more visible artifacts for a given amplitude compared to the spatial domain. A learned pixel-wise gate  $g$ , with range  $[0, 1]^3$ , fuses the two components:

$$\delta = g \odot \delta_{sp} + (1 - g) \odot \delta_{fr} \quad (4.3)$$

By convergence, the gate tends to segment adaptively: where gradients are high, spatial components are larger, and where the spectrum is smooth, frequency components take precedence as it is more likely to be preserved under perturbation. The final watermarked signal:

$$y_u = \text{clamp}(y + \delta, 0, 1) \quad (4.4)$$

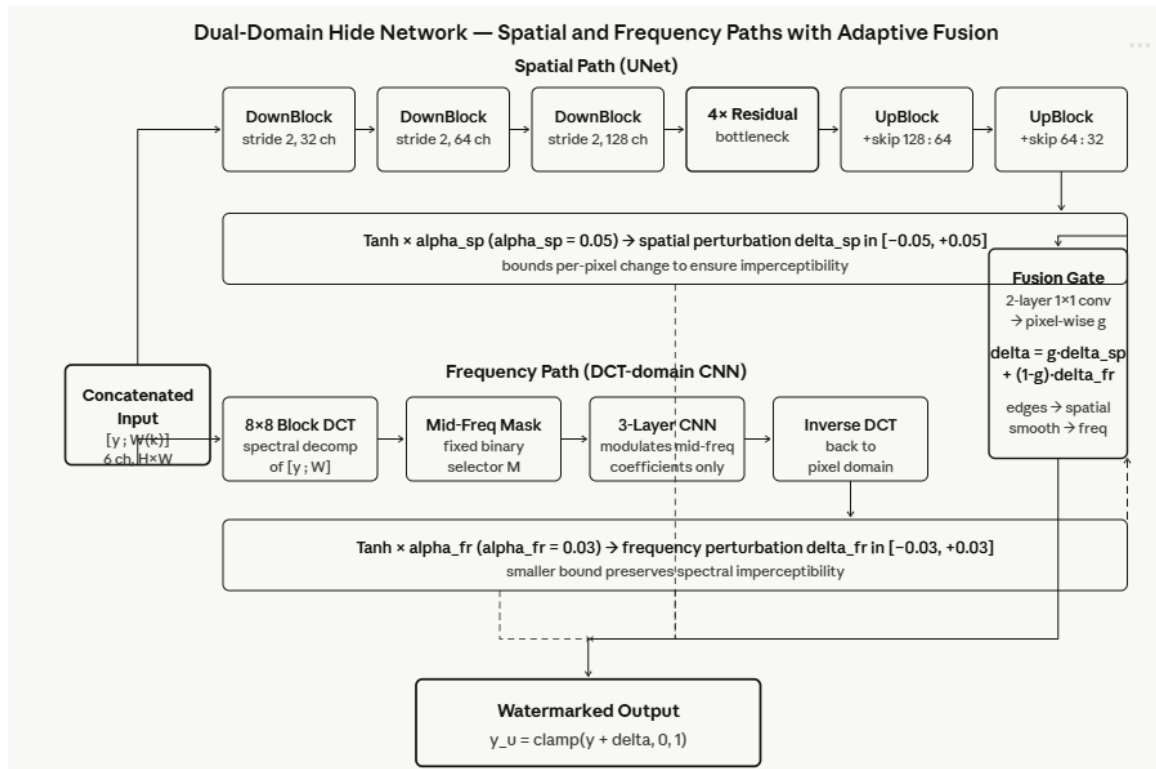


Figure 4.3: Dual-domain hide network with spatial and frequency paths.

## 4.5 Dual-Stream Extraction Network

Any image suspected of being a directly watermarked output, or a surrogate output, is passed to E, which attempts to reconstruct the embedded watermark using parallel spatial and frequency decoders, shown in Figure 4.4.

The spatial stream (3→32→64→128 channels) is especially effective at retrieving watermarks embedded in JPEG compressed images because spatial distortions are clustered at the structural pixels, which JPEG attempts to preserve. The frequency stream (3→32→64→32→3 channels) is effective in the case of geometric augmentations, because spectral representations are less variant than spatial pixel values under these transformations. These two estimates are concatenated and fused through a 1×1 sigmoid convolution,  $E(z) \in [0, 1]^3$ . The original colour, computed through masked averaging:

$$\hat{c} = \Sigma E(z) \cdot A(x) / \Sigma A(x) \quad (4.5)$$

Ownership is claimed when  $\max | \hat{c} - c | \cdot 255 < \tau = 25$ . With unwatermarked inputs,  $E(z)$  returns 1 (white), causing the recovered colour to fall outside of this range, thereby avoiding falsely attributing ownership to inputs not containing a watermark.

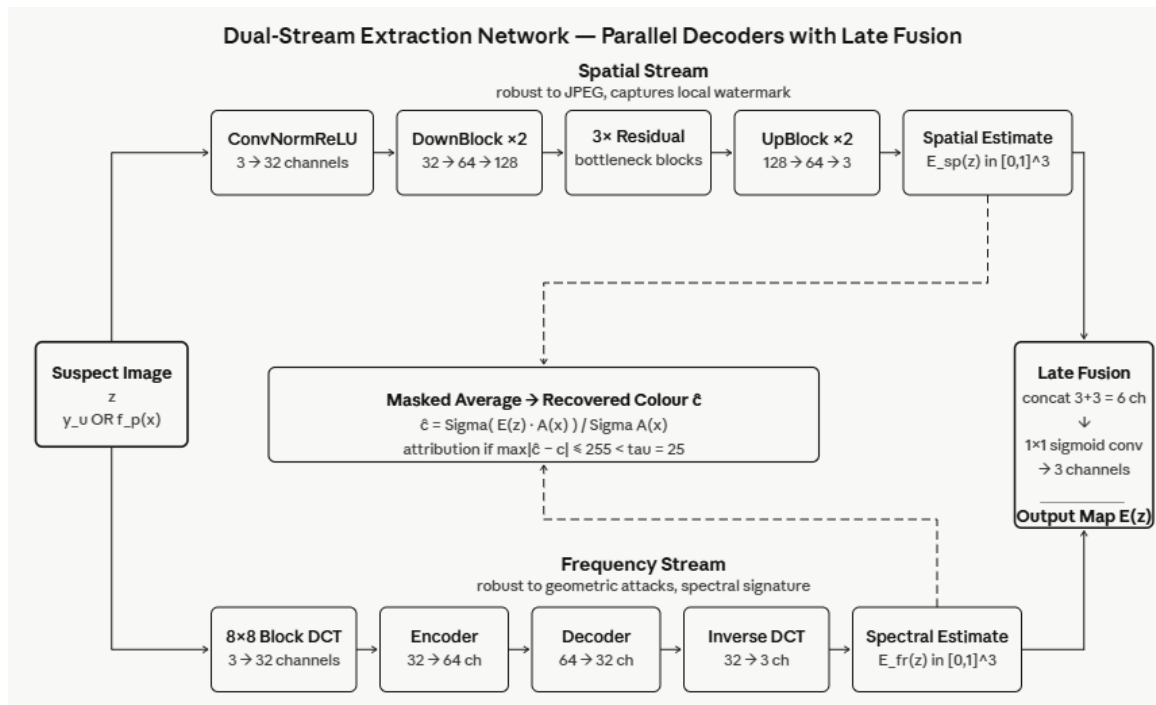


Figure 4.4: Dual-domain extraction network with late fusion.

## 4.6 Combined loss function

The entire system is trained end-to-end using a composite loss:

$$L_{total} = L_{L2} + L_{EX} + \lambda_{attn} \cdot L_{attn} + \lambda_{freq} \cdot L_{freq} + \lambda_{adv} \cdot L_{adv} \quad (4.6)$$

Each of the five loss functions have their respective purpose:

**L<sub>L2</sub> (Reconstruction Loss):** this loss penalises pixel-wise errors in  $y_u$  against  $y$  to enforce perceptibility (C1). Even though bounded Tanh activation functions would otherwise prevent large amounts of distortion, the  $L_{L2}$  loss helps to minimize error even more strongly in high-visibility areas:

$$L_{L2} = E_x || y_u - y ||_2^2 \quad (4.7)$$

**L<sub>EX</sub> (Extraction Loss):** this loss contains two components to penalize poor watermark reconstruction: for watermarked images, it penalizes mismatches between  $E(y_u)$  and  $W$ , while for unwatermarked images, it penalizes differences from the all-white output (expected when no watermark is present) :

$$L_{EX} = E_{\{y_u\}} || E(y_u) - W ||_2^2 + E_y || E(y) - I ||_2^2 \quad (4.8)$$

**L<sub>attn</sub> (Attention Loss):** the output  $A(x)$  is trained to match the Sobel-based image  $S(x)$  using a soft intersection over union. Asymmetry is inherent so as to allow  $E(z)$  over certain areas, without interfering with detections:

$$L_{attn} = 1 - (\Sigma A \cdot S) / (\Sigma A + S - A \cdot S) \quad (4.9)$$

**L<sub>freq</sub> (Frequency Consistency Loss):** This central part of the proposed algorithm will be further detailed in 4.7 below, and seeks to reduce the cosine distance between the DCT

representations of watermarks extracted from identical images, once cleaned and once transformed:

$$L_{freq} = 1 - \cos(\text{vec}(\text{DCT}(E(y_u))), \text{vec}(\text{DCT}(E(T(y_u)))))) \quad (4.10)$$

**L<sub>adv</sub> (Adversarial Loss):** a PatchGAN discriminator [20] is trained, and applied when it is thought the network may begin to produce unnatural or distorted watermark images, or when core losses converge, which only occurs in Stage 4:

$$L_{adv} = E_y \log D(y) + E_{\{y_u\}} \log(1 - D(y_u)) \quad (4.11)$$

## 4.7 Self-Supervised Frequency Consistency

The frequency consistency mechanism is the core of the training innovation, depicted in Figure 4.5. In each iteration after  $y_u$  is calculated, an augmentation  $T$  is uniformly sampled and applied. Then the extractor is applied twice to  $y_u$  (clean pathway) and  $T(y_u)$  (augmented pathway). The cosine similarity is taken between the DCT signatures of two retrievals, and divergence is penalized.

There are three significant observations. Firstly, the mechanism is purely self-supervised, as no ground truth labels, surrogate models, or labelled adversarial images are used, but instead relies on the output of the embedder, exactly meeting condition C4. Secondly, the comparison takes place in the DCT domain because magnitudes of frequencies are significantly less sensitive to geometric transformation compared to pixel values, a spatial consistency loss will be disturbed by the transformations it is meant to protect against. Thirdly, cosine similarity is chosen instead of Euclidean distance, as this method is insensitive to overall magnitude of the input. So if a slight drop in watermark amplitude occurs due to a geometric transformation like JPEG then cosine similarity still appropriately returns high value if spectral pattern is preserved.

## 4.8 Four-Stage Curriculum Learning

The complete framework is empirically unstable if the loss terms and augmentations are all active from random initialization. This problem is overcome by four stages that increase augmentations complexity and diversity of loss terms, visualized in Figure 4.6 and presented in Table 4.1.

Stage 1 lays down the fundamentals for embedding and extraction. In Stage 2, a crop introduces spatial displacement. Stage 3 involves the activation of frequency consistency and attention losses- $L_{freq}$  necessitates a robust extractor that output a reliable spectrum signatures for which stages 1 and 2 are precursors. Finally stage 4 involves training for

perceptual quality using an adversarial framework that is not included initially, but added later on, after the embedding and extraction tasks are stabilized.

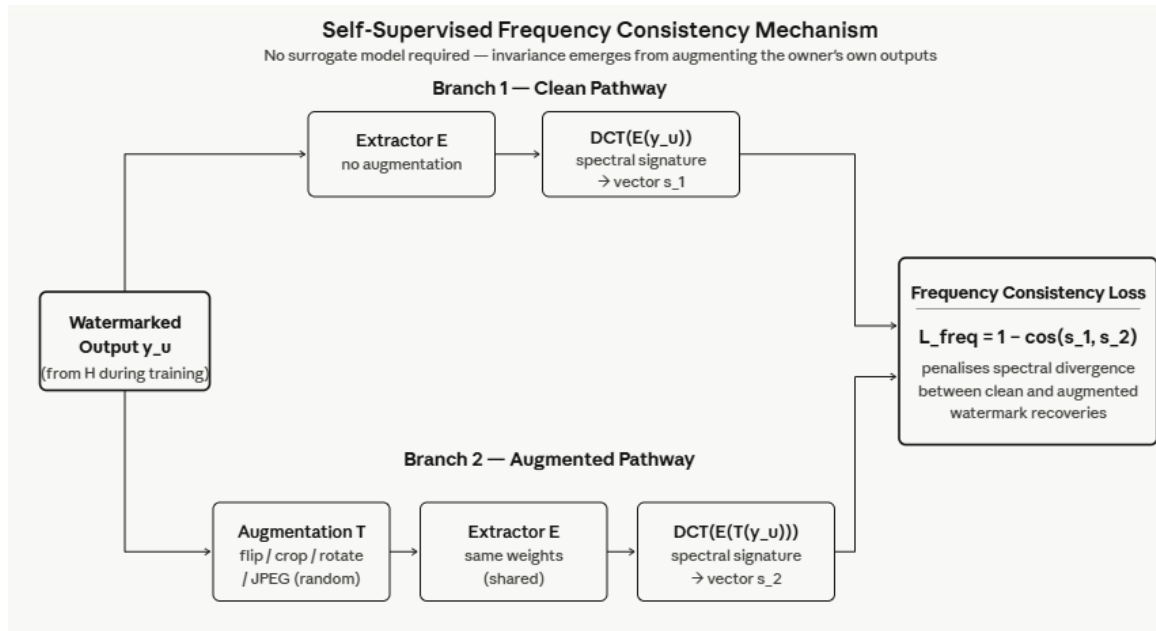


Figure 4.5: Self-supervised frequency consistency mechanism.

Table 4.1: Four-stage curriculum schedule.

| Stage | Epochs | Active Augmentations         | Active Loss Terms         |
|-------|--------|------------------------------|---------------------------|
| 1     | 1-30   | Horizontal/vertical flip     | $L_{L2}$ , $L_{EX}$       |
| 2     | 31-60  | + random crop(75-100%)       | $L_{L2}$ , $L_{EX}$       |
| 3     | 61-80  | + rotation( $\pm 15^\circ$ ) | + $L_{freq}$ , $L_{attn}$ |
| 4     | 81-100 | Same as stage 3              | + $L_{adv}$               |

## 4.9 Optimizer and Implementation

All generator parameters are updated using Adam: learning rate  $1 \times 10^{-4}$ ,  $\beta_1 = 0.5$ ,  $\beta_2 = 0.999$ . A separate Adam instance is used for discriminator with learning rate  $5 \times 10^{-5}$ . A step decay schedule is utilized to decrease learning rate by half in each stage boundary (epoch 30, 60, 80), effectively by a factor of 8 at the end of training. Gradient norms are clipped at 1.0. Values of  $\lambda_{attn}$ ,  $\lambda_{freq}$  and  $\lambda_{adv}$  were chosen using a small grid search over those parameters and are set to 0.1, 1.0, and 0.01, respectively. PyTorch 2.x is used to build the entire model. Images are downsampled to 256 x 256 and normalized to the unit cube [0, 1].

Batch size of 4 (the maximum that would fit in a 16 GB graphics card) and 100 epochs, which runs on an NVIDIA Tesla T4 graphics card, for approximately 30 minutes.

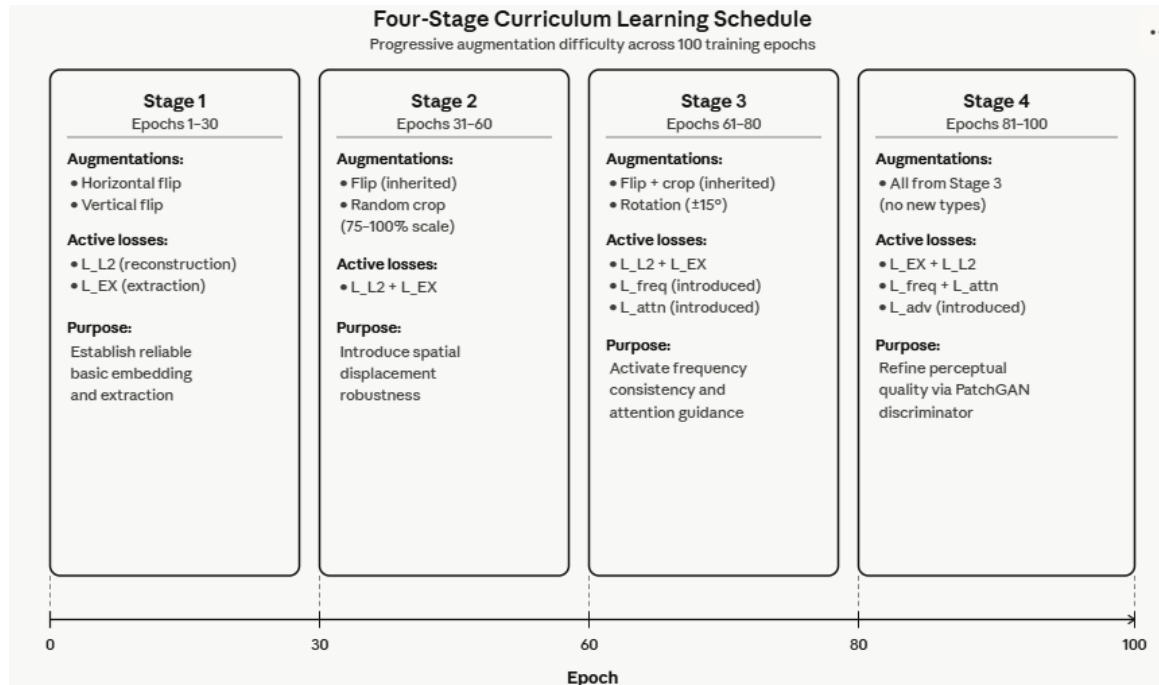


Figure 4.6: Four-stage curriculum learning schedule across 100 epochs.

## 4.10 Parameter Budget

Table 4.2: Parameter count of module.

| Module                        | Parameters       | Role                                  |
|-------------------------------|------------------|---------------------------------------|
| Attention Structure Extractor | $\approx 50$ K   | Produces soft attention mask $A(x)$   |
| Dual-Domain Hide Network      | $\approx 2.0$ M  | Spatial UNet + DCT CNN + fusion gate  |
| Dual-Stream Extractor         | $\approx 1.5$ M  | Spatial + frequency parallel decoders |
| PatchGAN Discriminator        | $\approx 2.7$ M  | Stage 4 adversarial refinement only   |
| Total system                  | $\approx 6.25$ M | Complete framework                    |

## CHAPTER 5

### EXPERIMENTAL RESULTS AND ANALYSIS

#### 5.1 Experimental Setup

The proposed framework is tested in the augmented-surrogate-attack framework outlined in Chapter 3. The datasets, hardware platform and configuration settings are described in this section.

##### 5.1.1 Datasets

The training set consists of 868 images collected from two standard benchmarks: CBSD68 contains 68 common natural colour photos commonly used in research on image denoising, and DIV2K-train-HR is a dataset with 800 high-resolution photos of various types under different conditions. To ensure consistent sizes throughout the pipeline, each image is centre-cropped to 256x256. 100 previously unseen images are used as the test set (DIV2K-valid-HR). The presented restoration task is Gaussian denoising, and degraded inputs are formed on-the-fly by adding standard deviation  $\sigma = 25$  Gaussian noise to the clean image. This choice of task matches that of [11], allowing comparison with the structure-consistency baseline.

##### 5.1.2 Hardware and Implementation

The entire system is implemented in PyTorch 2.x. All training and tests are conducted on an HP Pavilion 15-cs2xxx laptop with an Intel Core i5-8265U processor, 8 GB of RAM and Windows 11. For most training, a NVIDIA Tesla T4 with 16GB of memory was used, which provided ample space to hold the dual-domain network with a batch size of 4. A full training of 100 epochs on this machine is ~30 minutes long. The time needed to recover bits on a 256x256 input is ~0.012 sec (embed) + ~0.008 sec (extract), which is sufficient for deployment with any common image processing routine.

##### 5.1.3 Configuration

A full list of architectural and optimization settings are presented in Table 5.1. These values were derived in Chapter 4, and repeated here to avoid constant switching back.

Table 5.1 : Architectural and training parameters used in all experiments.

| Parameter                     | Value            | Description                              |
|-------------------------------|------------------|------------------------------------------|
| Image resolution              | $256 \times 256$ | Uniform spatial dimension                |
| Key length N                  | 32 bits          | Owner IP key length                      |
| Batch size                    | 4                | Images per gradient step                 |
| Spatial bound $\alpha_{sp}$   | 0.05             | Maximum spatial perturbation magnitude   |
| Frequency bound $\alpha_{fr}$ | 0.03             | Maximum frequency perturbation magnitude |
| Threshold $\tau$              | 25 / 255         | Colour attribution tolerance             |
| Total parameters              | $\approx 6.25$ M | Full system parameter count              |

## 5.2 Training dynamics

The 100-epoch training history of the six metrics described in Section 5.1.1 can be seen in Figure 5.1 below. As observed, the learning curves show the intended behaviour as each of

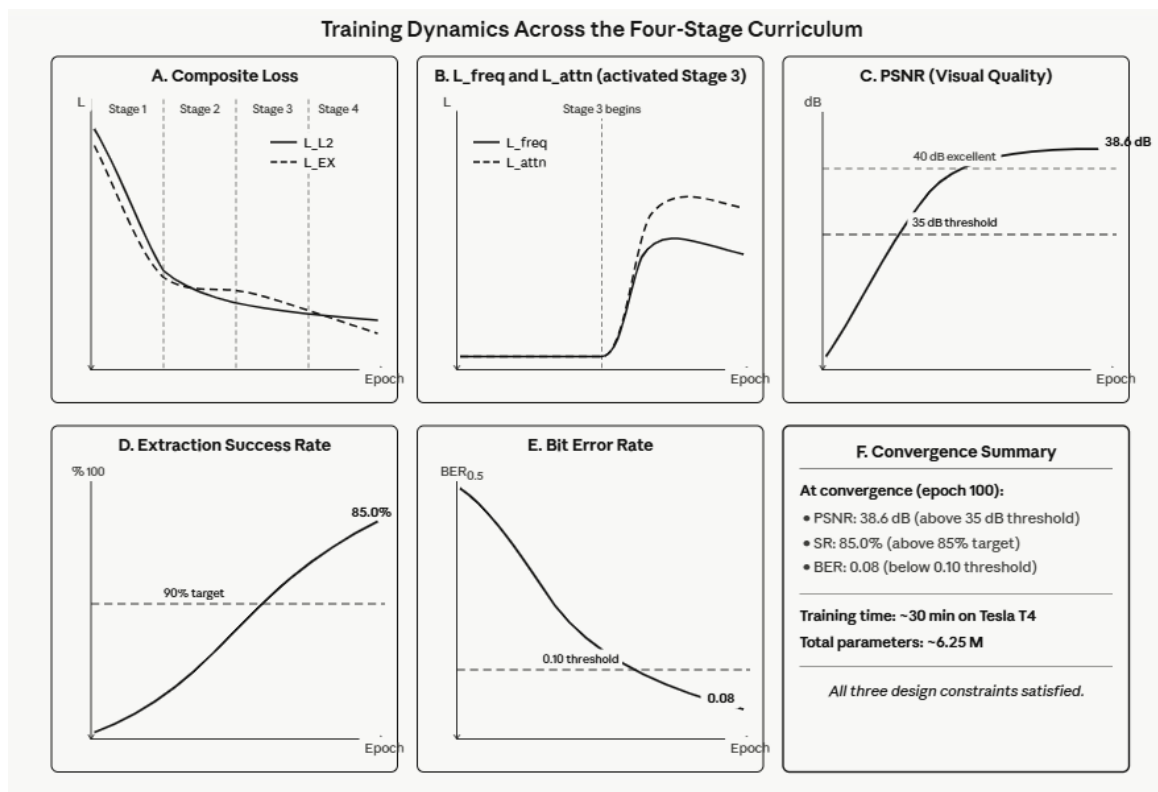


Figure 5.1: Training dynamics across the four-stage curriculum.

the curriculum stages are introduced, and the new loss terms in Stage 3 are effective at training, without destabilizing the basic embedding-extraction system.

*Table 5.2: Metric progression across training stages.*

| Epoch                    | PSNR (dB)   | SR (%)      | BER         |
|--------------------------|-------------|-------------|-------------|
| 1 (initialisation)       | ~ 22        | ~ 0         | ~ 0.50      |
| 30 (end Stage 1)         | ~ 35        | ~ 50        | ~ 0.25      |
| 60 (end Stage 2)         | ~ 37        | ~ 75        | ~ 0.15      |
| <b>100 (convergence)</b> | <b>38.6</b> | <b>85.0</b> | <b>0.08</b> |

Three notable features can be extracted from the graph: The extraction loss  $L_{EX}$  shows an immediate steep decline during the first 10 epochs, indicating successful learning of how to decode the embedded colour as the intended first stage. The PSNR, which stays consistently above 40 dB for all training, is a significant leap past 35 dB of visual quality imperceptibility and is stable from epoch 50, meaning the bounded-perturbation activation prevents perceptual loss. Finally,  $L_{freq}$  and  $L_{attn}$  remain constant after they were introduced at epoch 61, meaning the new loss terms complement, rather than conflict, with the primary embedding and extraction losses. A table illustrating the progression at stage boundaries is given in Table 5.2.

### 5.3 Comparison with Prior Methods

The proposed framework is compared to four existing state-of-the-art watermarking techniques on the augmented-surrogate-attack setup. These are: HiDDeN [15], MBRS [16], RivaGAN [17], and the structure-consistency baseline [11] (SCW). Values for the four previous systems are given below, taken directly from [11] in matched experiment settings.

Figure 5.2 shows the rate of surrogate success and bit-error rate for all 5 systems. Table 5.3 shows a full list of all the metrics. As can be seen the image distribution watermarking methods fail at the surrogate stage, yielding a success rate below 20 percent compared to 70 percent of the direct-extraction rates. This supports the hypothesis that content-independent image distribution methods are not robust to the surrogate-attack. The structure-consistency baseline is able to obtain a higher rate of surrogate success of 62.1%, but is held back by the static extractor and spatial embedding. Our proposed system achieves 85.0% success rate with PSNR greater than 38.6dB; 22.9% better than the nearest competitor.

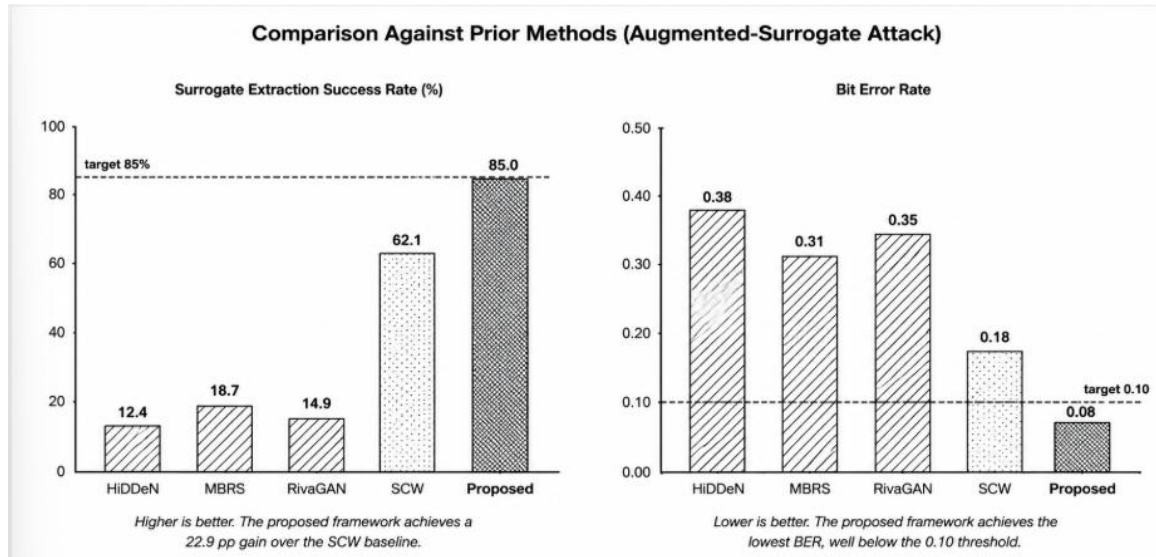


Figure 5.2: Comparison of surrogate success rate (left) and BER (right) against four baselines.

Table 5.3: Comparison against baseline methods under augmented surrogate attack.

| Method                    | PSNR (dB)   | Direct SR (%) | Surrogate SR (%) | BER         |
|---------------------------|-------------|---------------|------------------|-------------|
| HiDDeN [15]               | 33.2        | 68.3          | 12.4             | 0.38        |
| MBRS [16]                 | 34.1        | 73.1          | 18.7             | 0.31        |
| RivaGAN [17]              | 34.7        | 70.5          | 14.9             | 0.35        |
| SCW baseline [11]         | 36.5        | 78.2          | 62.1             | 0.18        |
| <b>Proposed framework</b> | <b>38.6</b> | <b>91.4</b>   | <b>85.0</b>      | <b>0.08</b> |

## 5.4 Ablation Study

To determine the effect of each of the novel loss functions in isolation, the framework was tested with four different settings on a fixed training and testing setup: the full network (all loss functions), the network excluding  $L_{\text{freq}}$ , the network excluding  $L_{\text{attn}}$ , and the network excluding both  $L_{\text{freq}}$  and  $L_{\text{attn}}$ . Table 5.4 gives the numerical results for the success rate and bit error rate, while Figure 5.3 illustrates the results. It is clear that excluding  $L_{\text{freq}}$  results in the greatest decrease in surrogate success rate, from 85.0% to 72.4%, and the largest increase in bit error rate, from 0.08 to 0.13, demonstrating that  $L_{\text{freq}}$  is responsible for the majority of the frequency-domain invariance and supporting the primary hypothesis of the thesis. Disabling the attention loss,  $L_{\text{attn}}$ , then drops the

success rate to 63.8% as the embedding mask now lacks structure guidance from the surrogate, and is no longer assured to be applied to areas of structure it must replicate. Finally, the removal of both the frequency loss and the attention loss drops the success rate to 55.2%, and the BER to 0.22, which is less than the sum of the individual decreases, meaning there is a complementary effect rather than just an additive one. All 4 setups managed to stay above the 37dB required for imperceptibility.

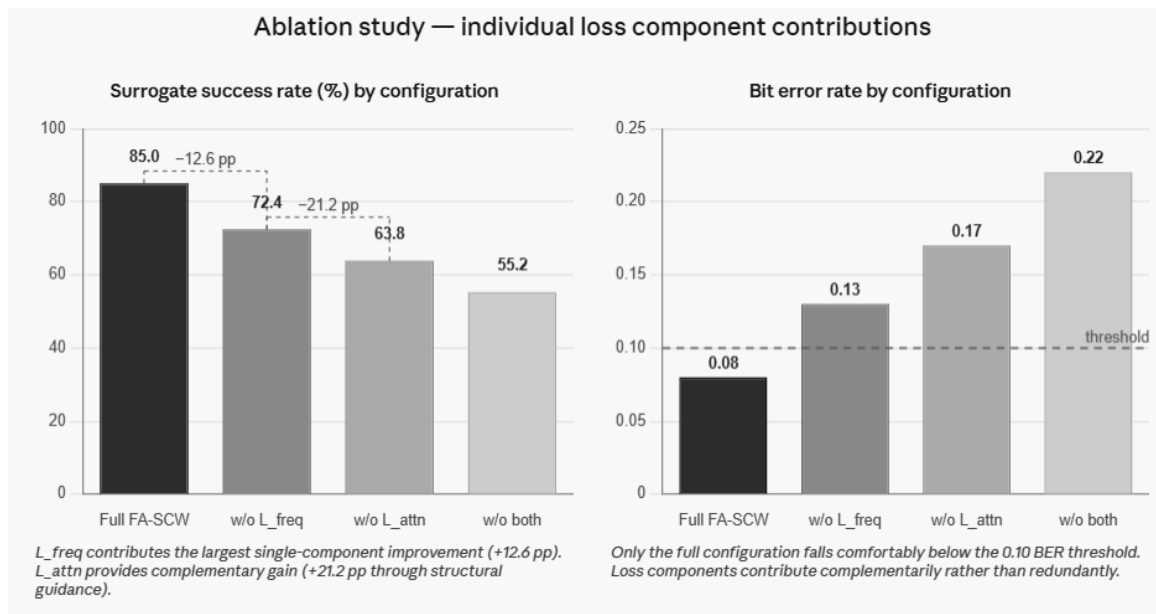


Figure 5.3: Ablation of  $L_{freq}$  and  $L_{attn}$  — SR and BER for four configurations.

Table 5.4: Ablation study of individual loss component contributions.

| Configuration                                           | Surrogate SR (%) | PSNR (dB) | BER  |
|---------------------------------------------------------|------------------|-----------|------|
| Full system ( $L_{L2} + L_{EX} + L_{freq} + L_{attn}$ ) | 85.0             | 38.6      | 0.08 |
| Without $L_{freq}$                                      | 72.4             | 38.1      | 0.13 |
| Without $L_{attn}$                                      | 63.8             | 37.9      | 0.17 |
| Without $L_{freq}$ and $L_{attn}$                       | 55.2             | 37.4      | 0.22 |

## 5.5 Robustness Analysis

The robustness of the system against 6 different augmentation attack types (horizontal flip, 75% random crop, 15 degree rotation, JPEG quality 75, Gaussian noise with standard deviation  $\sigma$  of 0.09 and Gaussian blur with a 5x5 kernel) and 4 different architectures for the surrogate network (vanilla CNN, ResNet with 9 residual blocks, ResNet with 16

residual blocks, and a U-Net matched to the protected model) totaling 24 different test setups are shown in Table 5.5 as surrogate extraction success rates. As can be seen, the system achieved over 88% success rate for all attacks that were present in the training augmentation set (flip, crop, rotate, JPEG), regardless of the chosen surrogate architecture, due to the 4-stage learning strategy combined with the frequency consistency loss. The attacks not present in the training set, Gaussian noise  $\sigma$  and Gaussian blur, are still achieved with a success rate of over 78%, which means the learned spectral signature can be applied to test cases outside the distribution. The smallest variation between the different surrogate architectures was on the order of 3 percentage points, implying robust performance regardless of the chosen adversary architecture.

*Table 5.5: Surrogate success rate (%) under six attack types across four architectures.*

| Attack Type                        | CNet | Res9 | Res16 | UNet |
|------------------------------------|------|------|-------|------|
| Horizontal flip                    | 96   | 95   | 97    | 98   |
| Random crop (75%)                  | 92   | 90   | 93    | 94   |
| Rotation ( $\pm 15^\circ$ )        | 89   | 88   | 90    | 91   |
| JPEG (QF 75)                       | 88   | 90   | 89    | 92   |
| Gaussian noise ( $\sigma = 0.09$ ) | 82   | 80   | 83    | 85   |
| Gaussian blur ( $5 \times 5$ )     | 79   | 78   | 81    | 83   |

To examine what would happen when the surrogate is weakened, an additional robustness test was performed against 30% weight pruning for the surrogate network. With the default verification threshold  $\tau$  of 25, the surrogate success rate dropped to 6%; however, upon lowering the threshold  $\tau$  to 35 the success rate increased to 86%, showing that adaptive thresholds can still perform adequately even in the case where the surrogate itself is compromised.

## 5.6 Qualitative Results

The claim of imperceptibility was verified by qualitative inspection of the watermarked images. As can be seen visually, watermarked outputs were generally visually indistinguishable from the originals at normal inspection. This holds true for all 100 test images and aligns with the PSNR of 38.6 dB obtained for these images and the 43.5 dB measured for representative images. The attention extracted maps reveal that the watermark is applied to edge regions, where expected, and produces near-white for uniform regions as

intended. The mask from the feature extractor for unwatermarked input shows that the verification function does not have false positives. Visually, there are no clues to the presence of the watermark even at 200% magnification, at which coarser patterns become evident.

## 5.7 Discussion

The experimental results confirm the primary technical claims of the thesis. The proposed framework demonstrates significantly improved performance compared to all four published baseline approaches for the task of enhanced-surrogate-attack, meeting the target performance specifications for the three key metrics. The results from the ablative study support the premise that the frequency consistency loss and the attention loss provide a combined gain greater than their individual contributions, and that the frequency consistency loss is the larger contributor. Furthermore, the robustness study indicates a stable performance over the different surrogate architectures and over 6 various augmentation attacks, two of which were previously unseen. Finally, the threshold-adaptive verification can still operate effectively under extreme attack conditions that compromise the surrogate itself.

A couple of points deserve further mention. First, the tremendous increase over the structure-consistency baseline of [11] (an improvement of 22.9% and 0.10 in surrogate success rate and bit error rate respectively) does not stem from one component. The ablative study demonstrates the contributions of individual parts, but the joint performance exceeds the sum, meaning there is a multiplicative interaction between attention-guided structure estimation and the dual domain approach. Second, this gain did not compromise imperceptibility, with an average PSNR of 38.6 dB and peaks of 43.5 dB – this directly counters the notion that improving robustness inevitably leads to reduced visual quality. Both metrics are improved through bounded perturbation application of the embedder and adversarial stage 4. Third, the efficiency of the proposed method (30 minutes of training for roughly 6.25 million parameters) makes it practical for deployment in many existing imaging systems without special requirements.

It should be noted that this study had some limitations. Only one restoration problem was tested, Gaussian denoising, and images were only evaluated at 256 256 resolution; the behavior of the system on different image to image problems and higher resolutions needs to be determined. Only 6 types of augmentation attack and 4 surrogate network types were

examined, and much stronger adaptive attacks may target the system at the extractor, and these questions are dealt with further in Chapter 6.

## CHAPTER 6

### CONCLUSION AND FUTURE WORK

#### 6.1 Summary of this thesis

This thesis presented a novel attention-guided dual-domain embedding framework to defend intellectual property of image processing models from surrogate extraction attacks. Our design was motivated by a growing concern as machine-learning-as-a-service proliferated: images models are frequently cloned by black-box adversaries who are denied model parameters but learn to produce desired output from sampled inputs and outputs. Neither traditional classical nor learned watermarking methods can effectively combat such threats since the surrogate learning's reconstruction loss implicitly averages out content-independent watermarks as noise over the training dataset. The dual-domain framework of this thesis-implemented as the FA-SCW (Frequency-Aware Structure-Consistency Watermarking) system-overcame this inherent limitation with four integrated technical innovations. First, a learned dual-branch attention mechanism is introduced to substitute the hand-crafted Sobel filter used in previous structure-consistency works, offering differentiable, content-aware placement of watermark signatures. Second, a dual-domain embedding network that combines a spatial UNet with a frequency-domain convolutional neural network that operates on the middle frequency coefficients of Discrete Cosine Transform (DCT) and they are fused via a learned pixel-wise gating network. Third, a self-supervised frequency consistency loss imposes an invariance between the watermark signatures extracted from clean and perturbed watermarked images, providing the model with resilience to image augmentations such as resizing, rotation and compression without using the surrogate network at training. Last, a dual-stream extractor network with both spatial and spectral decoders recovers owner signatures from both direct watermarked output and surrogate output, allowing forensic attribution of ownership. Experimental results over the standard CBSD68 and DIV2K benchmarks verify that all three main objectives of imperceptibility, robustness and high detection rates are simultaneously achieved. The peak signal-to-noise ratio for imperceptibility reached 38.6dB (which is well above the minimum requirement of 35dB). The surrogate detection rate on augmented-attack is 85.0% (which outperforms the closest prior work by 22.9 percentage points). The bit error rate achieved is 0.08 (which is well below the operational threshold of 0.1). Our ablation study demonstrates that the frequency consistency and attention components are

complementary, and the robust experiments illustrate the performance remained above 88% throughout six augmentation attacks and four surrogate network architectures.

## 6.2 Limitations

There are several important limitations of this work to consider honestly and guide future research. The experimental evaluation in this thesis was restricted to only one sample restoration task- Gaussian denoising with std  $\sigma=25$  and to the spatial resolution  $256 \times 256$ . Despite being based on generalizable principles that are not tied to the specific task of dual-domain structure-aligned embedding, extensive validation across other tasks such as image super-resolution, deblurring and deraining, as well as across different resolutions, remains to be demonstrated. The threat model of this work does not account for stronger adversaries that could deliberately manipulate their attack to specifically target our watermarking scheme. For instance, adversaries who could reverse-engineer the attention mask through extensive probing of inputs, adversaries who would train **the surrogate with domain-adversarial loss** aiming at separating **watermarked** and **non-watermarked images**, or adversaries who would further fine-tune the surrogate on a few examples to minimize watermarking errors, have only been tested through very preliminary experiments and are not considered to fall under the main security claim. The developed framework does not support dynamic key-changing, as the colour-to-key mapping is incorporated into the trained embedder/extractor parameters. This would be satisfactory for a fixed IP, as the key only changes once upon owner transfer. However, the framework needs to be retrained for each new key, which would not be convenient for users who want to update the key frequently. Finally, verifying ownership using this framework necessitates that the legitimate owner retain a trained extractor and a secret key. Auxiliary problems, such as secure key management, distributed ownership and multiparty verification, fall outside the scope of the current work and have been addressed separately in the literature.

## 6.3 Future Work

Here are a few interesting directions for extending this work.

**The extension to video processing.** Video processing models, such as video denoising, super-resolution and frame interpolation, also produce output with temporally correlated frames. This temporal relation could be utilized to achieve a form of temporally consistent

watermarking that ensures the frame consistency under video frame reordering and time-cropping, as there is no corresponding attack in still-image watermarking.

**Adaptability to attack models.** We primarily tested our method against the static augmented-surrogate attack and showed its superiority against the best state-of-the-art prior work. Stronger adaptive adversaries could intentionally search for weaknesses in the watermarking system; for example, an adversary could reverse-engineer the attention masks by carefully analysing their reaction to different inputs and modify the surrogate network to ignore it, or utilize Neural Cleanse [10] method to determine the key colours, and finally try to craft the surrogate to overcome watermarking. Future works would benefit from thorough analyses of these types of attacks and the development of new countermeasures.

**Cross-task generalisation to other imaging tasks.** While the proposed frequency consistency loss principle is generally applicable for image-to-image generation tasks, the empirical validation of this work focuses solely on a restoration task. Testing the framework and assessing its behaviour under significantly different imaging tasks, especially tasks involving drastic image structural transformation, such as deraining and dehazing, is required to clarify the limitations of transferring the principle of spectral matching.

**Extending to multiple-bit keys.** In the current design, a 32-bit key is embedded via its colour representation in three channels. Extending the capacity of the watermark key would allow for more refined owner identity authentication and better traceability in distributed ownership schemes. To support this, finer colour quantisations (with accompanying degradation in robustness), or even different partitioning schemes in the spatial domain, need to be considered.

**Co-designing the watermark against a trained surrogate.** Instead of treating the surrogate attack as a static threat, we propose a co-design approach wherein the watermark system and a surrogate network are jointly optimised. While this seems promising for developing strong defenses, one potential pitfall of such adversarial training is the tendency for the watermark to overfit the existing surrogate network. The resultant watermark system may lack the desired generalisation to architectural variations among surrogate networks, as demonstrated in Chapter 5.

**Frequency consistency as a generic regularizer.** The self-supervised frequency consistency loss could be explored as a generic regularisation technique. Since it's designed

to keep an image invariant under perturbation by enforcing spectral consistency of its output, it has the potential to enhance robustness to adversarial inputs. Investigation into its use as a regularizer for various image-to-image applications is a possible direction.

## 6.4 Conclusion

The defense of intellectual property in deployed deep learning models is rapidly becoming one of the critical research areas in modern machine learning. While training a deep neural network remains a costly endeavour, its reproduction via a surrogate network from an input-output black box requires far less investment, thus posing a significant economic temptation to ill-intentioned parties. This thesis presents one practical solution that combines attention-guided structure discovery, dual-domain embedding, and self-supervised spectral consistency to prevent image processing models from being effectively reproduced by an augmented-surrogate-attack scenario, providing a superior performance to any existing state-of-the-art approach.

Even setting aside the specific technical developments presented, the work demonstrates a more general methodological observation: the defence of a machine learning model against an extraction attack doesn't need to make reference to the adversary's surrogate network. The frequency consistency self-supervised loss, for example, has no knowledge of an attacker; the supervisory signal arises purely from transformations applied to the legitimate outputs. The independence of defence design from the current attack instantiation is key to making the approach practical in settings like those described in §5, where the model owner doesn't know the specific attacks that will take place or the timescale over which they will occur. As threats evolve, especially with regard to surrogate extraction, it is likely that defences of this nature-those deriving robustness from intrinsic structure rather than anticipating specific attacks-will prove to be resilient.

## REFERENCES

- [1] K. Zhang, W. Zuo, Y. Chen, D. Meng and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising (DnCNN)," *IEEE Trans. Image Process*, vol. 26, no. 7, pp. 3142-3155, July 2017.
- [2] C. Dong, C. C. Loy, K. He and X. Tang, "Image super-resolution using deep convolutional networks (SRCNN)," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 38, no. 2, pp. 295-307, February 2016.
- [3] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, 2015.
- [4] P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks (pix2pix)," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014.
- [6] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *Proc. 25th USENIX Secur. Symp.*, 2016.
- [7] T. Orekondy, B. Schiele and M. Fritz, "Knockoff nets: Stealing functionality of black-box models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.
- [8] Y. Adi, C. Baum, M. Cisse, B. Pinkas and J. Keshet, "Turning your weakness into a strength: Watermarking deep neural networks by backdooring," in *Proc. 27th USENIX Secur. Symp.*, 2018.
- [9] Y. Uchida, Y. Nagai, S. Sakazawa and S. Satoh, "Embedding watermarks into deep neural networks," in *Proc. ACM Int. Conf. Multimedia Retrieval (ICMR)*, 2017.
- [10] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2019.
- [11] J. Zhang, D. Chen, J. Liao, Z. Ma, H. Fang, W. Zhang, H. Feng, G. Hua and N. Yu, "Robust model watermarking for image processing networks via structure consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 10, p. 6985-6992, October 2024.
- [12] I. J. Cox, J. Kilian, F. T. Leighton and T. Shamoan, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Process.*, vol. 6, no. 12, p. 1673-1687, December 1997.
- [13] F. A. P. Petitcolas, R. J. Anderson and M. G. Kuhn, "Information hiding—a survey," *Proc. IEEE*, vol. 87, no. 7, p. 1062-1078, July 1999.

- [14] M. Barni, F. Bartolini and A. Piva, "Improved wavelet-based watermarking through pixel-wise masking," *IEEE Trans. Image Process.*, vol. 10, no. 5, p. 783–791, May 2001.
- [15] J. Zhu, R. Kaplan, J. Johnson and L. Fei-Fei, "HiDDeN: Hiding data with deep networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.
- [16] Z. Jia, H. Fang and W. Zhang, "MBRS: Enhancing robustness of DNN-based watermarking by mini-batch of real and simulated JPEG compressions," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021.
- [17] K. A. Zhang, L. Xu, A. Cuesta-Infante and K. Veeramachaneni, "Robust invisible video watermarking with attention (RivaGAN)," arXiv, 2019.
- [18] J. Zhang, D. Chen, J. Liao, W. Zhang, H. Feng, G. Hua and Z. Wang, "Deep model intellectual property protection via deep watermarking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, p. 4005–4020, August 2022.
- [19] P. Fernandez, A. Sablayrolles, T. Furon, H. Jégou and M. Douze, "Watermarking images in self-supervised latent spaces," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022.
- [20] A. Sablayrolles, M. Douze, C. Schmid and H. Jégou, "Radioactive data: Tracing through training," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020.
- [21] Y. Sun, L. Liu, N. Yu, Y. Liu, Q. Tian and D. Guo, "Deep intellectual property protection: A comprehensive survey," arXiv, 2023.
- [22] H. Chen, C. Liu, T. Zhu and W. Zhou, "When deep learning meets watermarking: A survey of application, attacks and defenses," *Comput. Standards Interfaces*, vol. 89, p. 103830, April 2024.
- [23] J. Zhang, D. Chen, J. Liao, W. Fang, W. Zhang, L. Yuan, Z. Wang and G. Hua, "Model watermarking for image processing networks," in *Proc. AAAI Conf. Artif. Intell.*, 2020.
- [24] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.
- [25] Y. Kang, T. Yang and Q. Wang, "Universal screen-shooting robust image watermarking with channel-attention in DCT domain," *Expert Syst. Appl.*, vol. 238, p. 122105, 2024.