This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Where is the Watermark? Interpretable Watermark Detection at the Block Level

Maria Bulychev, Neil G. Marchant, Benjamin I. P. Rubinstein
University of Melbourne
{m.bulychev, nmarchant, benjamin.rubinstein}@unimelb.edu.au

Abstract

Recent advances in generative AI have enabled the creation of highly realistic digital content, raising concerns around authenticity, ownership, and misuse. While watermarking has become an increasingly important mechanism to trace and protect digital media, most existing image watermarking schemes operate as black boxes, producing global detection scores without offering any insight into how or where the watermark is present. This lack of transparency impacts user trust and makes it difficult to interpret the impact of tampering. In this paper, we present a post-hoc image watermarking method that combines localised embedding with region-level interpretability. Our approach embeds watermark signals in the discrete wavelet transform domain using a statistical block-wise strategy. This allows us to generate detection maps that reveal which regions of an image are likely watermarked or altered. We show that our method achieves strong robustness against common image transformations while remaining sensitive to semantic manipulations. At the same time, the watermark remains highly imperceptible. Compared to prior post-hoc methods, our approach offers more interpretable detection while retaining competitive robustness. For example, our watermarks are robust to cropping up to half the image.

1. Introduction

Artificial intelligence can generate realistic and high-quality digital content, making it increasingly challenging to distinguish between human-created and machine-generated content. Modern diffusion models like Midjourney, Stable Diffusion, and DALL-E have revolutionised content creation with their ability to create an almost limitless range of novel images, producing everything from photo-realistic imagery to artwork of any style and aesthetic [27]. The spread of AI-generated content poses significant risks, including the potential for misinformation, intellectual property theft, and the blurring of lines between authentic and fabricated media [37]. For example, the European Union's AI Act imposes a transparency obligation for generative AI: providers must ensure that AI-generated text, images, video or audio are marked in a machine-readable way and identifiable as artificially generated [35].

Numerous security mechanisms have been developed to provide a means to assert ownership, verify authenticity, and protect intellectual property rights [30, 33]. However, most users share personal photos on social media without protection, unaware of the risks this poses. Unprotected images can be easily downloaded and manipulated, allowing malicious actors to create misleading or defamatory content and deepfakes. To address these issues, the research community has been exploring various solutions, with watermarking emerging as a prominent technique.

Watermarking involves embedding a unique signal into the original source, which, while subtle enough to remain invisible in normal use, can be extracted to validate multimedia content [30]. The process consists of two main phases: embedding and detection. In the embedding phase the watermark information is incorporated into the host data using techniques like modifying pixel values [2], frequency coefficients [8], or other signal properties [12]. Detection involves analysing the potentially watermarked content to verify the presence of the watermark and extract any hidden information, which can be done through statistical analysis [17], correlation [8], or pattern-matching methods [5].

Most image watermarking schemes are treated as black-box systems that output either a binary decision (whether an image is watermarked or not) or a score for the entire input without any explanation [31]. This lack of interpretability can be problematic for users who may wish to understand why a particular input is classified as non-watermarked. For instance, an image that was originally generated with a watermark may no longer be classified as such after certain regions are edited. However, existing schemes do not indicate whether the absence of a watermark is due to localised modifications of a watermarked image or due to the whole image not being watermarked in the first place. Yet, many modern watermarking schemes [34, 39, 42, 43] rely on deep neural networks (DNNs), which behave like black boxes and offer no clear reasoning for their decisions. Although explainability is widely seen as essential for trustworthy AI [9],

current DNN-based detectors lack reliable methods to provide human-understandable explanations primarily because deep neural networks themselves are not inherently interpretable [32].

This context highlights the need for watermarking schemes that offer region-level interpretability, not only detecting whether tampering has occurred, but also where. Targeting this gap, we propose MELB Multidimensional Embedding via Localised Blocking—a watermarking method that combines localised embedding, interpretability, and robustness to common image manipulations. Our main contributions are as follows:

- We introduce a block-wise image watermarking method that embeds watermark signals into the discrete wavelet transform (DWT) domain using a statistical partitioning approach. This design enables robustness against common manipulations such as compression, noise, and blur, while remaining sensitive to semantically meaningful changes in the image structure (*e.g.*, warping or shape modification).
- In contrast to traditional schemes that produce only a global score, our block-wise design allows for the generation of a detection map that highlights the distribution of watermark presence across the image. This enables users to identify which parts of an image have likely been altered and which remain authentic, providing transparency and localised tamper assessment.
- Our watermarks are highly imperceptible, with consistently strong results across standard similarity metrics, mostly exceeding prior approaches. Additionally, our embedding process supports explicit control over the trade-off between invisibility and robustness.
- Our method demonstrates robustness to common image modifications comparable to state-of-the-art post-hoc watermarking methods and shows notably stronger performance under cropping attacks. Even when up to 50% of the image is removed, detection remains reliable, with observed true positive rates of up to 91%.

2. Related Work

Digital Image Watermarking. Traditional watermarking methods [8, 42] embed watermarks post-hoc, working with existing images after their creation. With the development of diffusion models [14], research has expanded to include in-processing watermarking methods [11, 39] that embed the watermark during the generation process. These integrative approaches show promise in embedding signals deep within the image's semantic structure. However, their model-specific nature limits generalisability and applicability to existing images. Our post-hoc approach offers greater flexibility, accommodating AI-generated images from any model as well as natural images.

Post-hoc watermarking techniques [10, 19, 34, 42, 43] can be further distinguished between spatial domain and transform domain methods [3, 13]. Additionally, deep learning-based methods extend beyond these domains by leveraging latent feature spaces for watermark embedding [11, 20]. Spatial domain methods [41] operate directly on pixel values, offering low computational complexity but generally lower robustness to image processing operations [3, 30]. The spatial domain refers to the representation of an image as an array of pixels [41]. By contrast, transform domain methods, such as those using the Discrete Wavelet Transform (DWT) [40], typically exhibit higher robustness to image alterations and compression.

Recent advances in deep neural networks (DNNs) have led to learning-based watermarks that push the boundaries of robustness and imperceptibility. StegaStamp [34] embeds invisible messages into photographs by training with differentiable image perturbations for noise robustness and using a spatial transformer to account for geometric distortions from printing and re-capturing. RivaGAN [42] is a video watermarking method that uses an attention-guided dual-stream architecture to learn optimal watermark placement. The method's robustness is enhanced via adversarial training, where a critic ensures visual quality and a removal network continuously challenges the watermark's integrity. Other recent work [23, 29] explores interpretable and robust watermarking for ownership verification. While these DNN-based approaches offer strong robustness, they lack the ability to provide localised detection, which is a key feature of our proposed method.

Localised Watermarking. Despite the potential for enhanced interpretability and forensic utility, localised image watermarking—where a watermark's presence can be verified for specific image regions—has seen limited research to date. An early example is the hierarchical block-based approach of Utku Celik et al. [36], which introduced fragile watermarks with block-level tamper localisation using least significant bit embedding. More recently, EditGuard [44] introduced a dual watermarking strategy that embeds both localisation and copyright watermarks. While effective for tamper detection, this method is not robust to geometric augmentations such as cropping. Building on this, in concurrent work Sander et al. [25] proposed WAM, a localised watermarking method achieving strong robustness. However, WAM is designed to preserve the watermark even after severe alterations that significantly degrade the image, a feature that can be undesirable for image forensics. Our watermark, in contrast, is intentionally fragile and designed to break under such strong editing, providing a clearer signal that the image has been heavily manipulated. Furthermore, while pre-trained models for WAM are available, its detector is a transformer model with approximately 100 million parameters. Our method is training-free, using a

lightweight detector based on projections in DWT space. In contrast to both EditGuard and WAM, our work thus offers a computationally efficient approach to localised detection that is robust to common manipulations but fragile to severe, perceptually obvious ones. The concept of localised detection has seen limited exploration in other domains as well. For instance, AudioSeal [24] tackles localised watermarking for AI-generated speech, providing sample-level resolution in longer recordings.

Statistical Detection of Randomised Partitions. Kirchenbauer et al. [17] introduced a simple but innovative technique that embeds a watermark into the output of large language models by subtly modifying token probabilities during generation. This is achieved by partitioning the vocabulary at each step evenly into two distinct groups ("Green List" and "Red List"), where the likelihood of Red List tokens appearing in the generated text is reduced, while Green List tokens are boosted. A similar watermark was proposed for tabular data [46] where the watermark is embedded by perturbing specific cells ("key cells"), selecting the perturbation values from the "Green Domains". Our work builds upon these statistical partitioning concepts, adapting them to the image domain. This approach enables us to achieve localised detection, a feature notably absent in most existing image watermarking schemes, while maintaining a balance between imperceptibility and robustness.

3. Preliminaries

Our watermarking framework embeds a robust, yet imperceptible signal into images, allowing owners to identify their intellectual property without compromising visual quality. It also provides an interpretable detection map that visually highlights areas of potential modification (such as through insertions, deletions, or warping). This addresses scenarios on image-sharing platforms like Facebook or Instagram, where users post content without protection, risking unintended alterations or redistribution [45].

The framework is designed to maintain watermark integrity under common image editing operations. This robustness ensures that the connection to the original creator is preserved even when the image undergoes benign modifications or is subject to attempts at removal, aiding in authentication and traceability.

Our treatment of digital image watermarking introduces an additional aspect: localised detection. Unlike traditional watermarking schemes that provide only a global decision for an entire image, our formulation allows for fine-grained, interpretable watermark detection at the local level. We focus on post-hoc watermarking, where watermarks are embedded into any image, regardless of its origin—traditional, photographic, or AI-generated—offering broad utility in diverse scenarios.

3.1. Problem Formulation

An image is a three-dimensional array $I \in \mathcal{V}^{h\times w\times c}$ where h is the height, w is the width, c is the number of channels, and $\mathcal{V}$ is the range of intensity values. For 24-bit colour images, $\mathcal{V} = \{0, 1, \dots 255\}$ and $c = 3$ (*e.g.*, corresponding to RGB). We write $I_{i,j,k}$ to refer to the intensity of the k-th channel at spatial location (i, j), where i indexes rows (top to bottom) and j indexes columns (left to right).

An image watermarking scheme consists of two fundamental algorithms: an embedding algorithm and a detection algorithm. These algorithms are typically operated by a single entity, which we will call the *watermarker*.

Definition 1. The embedding algorithm maps an original image $I \in \mathcal{V}^{h\times w\times c}$ to a watermarked image $I^\star = \mathsf{embed}(I, \mathsf{sk})$ using a secret key sk known only to the watermarker.

Definition 2. The detection algorithm $\mathsf{detect}(\tilde{I}, \mathsf{sk})$ takes a possibly watermarked image $\tilde{I} \in \mathcal{V}^{h\times w\times c}$ and secret key sk. This yields two outputs: a *global detection* score $d \in [0, 1]$ which indicates whether the entire image is watermarked, and a *localised detection map* $V \in [0, 1]^{h\times w}$ quantifying the likelihood that localised regions are watermarked. These scores can be thresholded, so that a positive detection is represented by 1, while a negative detection is represented by 0.

We now summarise the design desiderata for our image watermarking scheme:

- *Imperceptibility*: The watermarked image $I^\star$ should be visually indistinguishable from the original image I. Formally, for some perceptual distance function d and threshold ϵ, we require $d(I, I^\star) \leq \epsilon$.
- *Blindness*: The watermark should be detectable without access to the original image.
- *Localised detection*: Watermark detection should highlight localised factors that influence detection decisions.
- *Soundness*: The false positive rate δ of the detection algorithm should be negligible. For $\mathcal{I}$ a distribution of non-watermarked images, $\Pr_{I\sim\mathcal{I}}[\mathsf{detect}(I, \mathsf{sk}) \neq 0] < \delta$ where we are implicitly applying a threshold to global detection scores.
- *Robustness*: The watermark should be detectable even after common image transformations T: $\mathsf{detect}(T(I^\star), \mathsf{sk}) \approx \mathsf{detect}(I^\star, \mathsf{sk})$.

Threat Model To define the scope of our threat model, we describe the abilities of attackers as follows. 1) We assume that attackers have full access to watermarked images and 2) are aware of the general process of the watermark embedding and detection algorithms, but 3) do not know secret

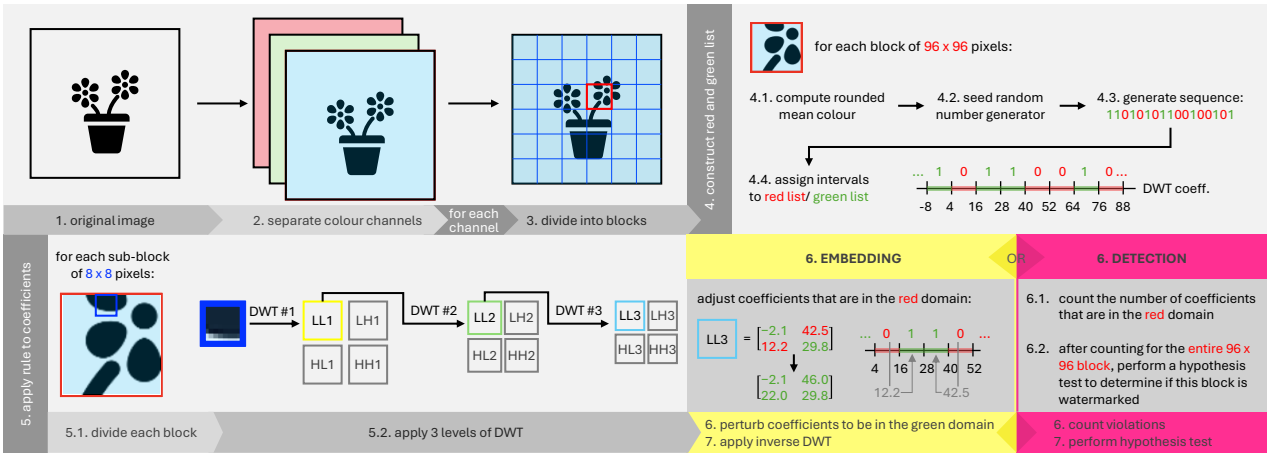


Figure 1. Pipeline for our watermarking method using $l = 12, k = 8, m = 96, d = 3$. The embedding phase partitions the domain of DWT coefficients for each colour channel into allowed and disallowed regions. It then perturbs these coefficients to ensure they all fall within allowed regions. The detection phase reverses this process and counts rule violations. Components highlighted in light grey are identical in both embedding and detection procedures.

information, such as the secret key (see Sec. 4.1). Moreover, we assume that potential attackers 4) aim to modify images while preserving their visual quality. For example, when adding noise, the modifications are expected to remain subtle and imperceptible to the human eye. In cases of blurring or compression, the objective is to maintain the image's core content and semantics. An observer should still be able to recognise and understand the primary subject.

4. MELB: Multidimensional Embedding via Localised Blocking

In this section, we present the algorithms of our watermarking method in detail. The process is illustrated in Fig. 1 and involves two stages: embedding and detection.

4.1. Watermark Embedding

Our watermark embedding method operates in the transform domain, specifically using the Discrete Wavelet Transform (DWT). The key idea is based on the concept of statistical partitioning of DWT coefficients into allowed ("green") and disallowed ("red") regions. Further background on the DWT is provided in App. A.1, and pseudocode is provided in Algorithm 1 of App. B.1.

We first divide the image into non-overlapping blocks of size $m \times m$, ignoring incomplete edge blocks. For each block, we generate a seed using a generic get_seed function (*e.g.*, in our experiments, this function returns the rounded mean intensity of the block per channel). This seed, along with a random secret key sk, is passed as input to a pseudorandom function (PRF), which returns a pseudorandom binary sequence. The resulting binary sequence determines the partitioning of DWT coefficients for the block into red (0) or green (1).

The actual embedding occurs at the sub-block level, using $k \times k$ sub-blocks ($k \leq m$) within each larger block. This reduces the visibility of block-wise artefacts in the watermarked image, as larger blocks could introduce more noticeable discontinuities at boundaries. For each sub-block:

1. We map the image to the transform domain by applying d levels of DWT to the LL band.
2. Using the parent block's red/green partitioning, we project the DWT coefficients onto the allowed (green) region by perturbing any coefficient in a red interval to the centre of the nearest green interval. This perturbation is applied only to the LL (low-frequency) band at level d, as low-frequency components are expected to be more robust to common image manipulations.
3. Finally, we apply the inverse DWT d times to obtain the watermarked sub-block.

This process ensures that in the watermarked image, nearly all DWT coefficients fall within green regions, while in natural, non-watermarked images, approximately 50% of the coefficients would fall in each region. App. B.2 provides a detailed discussion of the parameters, their control mechanisms, and our design choices. The key parameters impacting visibility and robustness are the interval length l and the number of DWT levels d. Increasing l generally improves robustness but may introduce visual artefacts. Higher values of d are typically more advantageous as they enhance both imperceptibility and robustness. Optionally, our method can operate in a mode that skips watermark embedding in low-entropy blocks, where perturbations might be more perceptible (see App. C.1).

Our block-based approach directly supports our goal of localised detection. As we shall see in the next section, the

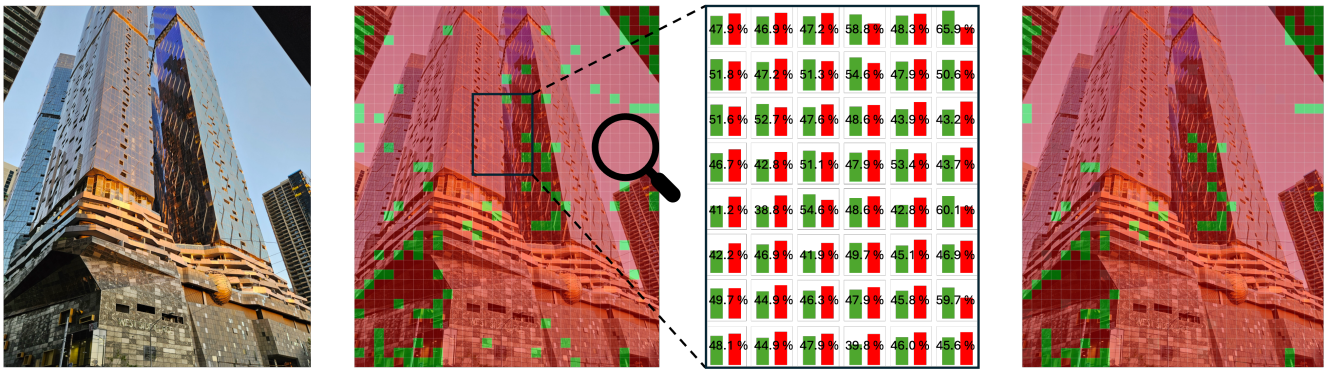


Figure 2. We apply our detection method to a non-watermarked image (left panel) and show the resulting localised detection map (middle-left panel). The middle-right panel illustrates the red/green coefficient distribution for a small part of the image. Most blocks show a balanced red/green ratio (∼50/50), with minor deviations towards the distribution tails. False positives are effectively filtered out, resulting in only 8.3% of the area being classified as watermarked. The right panel shows the post-processed detection map (see Sec. 4.2.1).

detection algorithm outputs scores for each block, which are statistically significant due to the multiple sub-blocks contained within. The resolution of our detection map is controlled by the block size m, allowing for fine-grained localisation of potential modifications.

Remark 1. The use of a PRF with a random secret key sk mitigates forgery risks by preventing unauthorised creation of watermarked images. This approach necessitates a more centralised setup for detection—*e.g.*, where the detection algorithm is made accessible to users through a public API. The PRF can be instantiated with a block cipher like AES or a cryptographic hash function such as SHA3 [17]. While our experiments use a fixed secret key, employing multiple keys can further strengthen the method against brute-force attacks (see Sec. B.4).

4.2. Watermark Detection

To check for the presence of a watermark in an image, we essentially reverse the embedding process. The complete procedure is detailed in Algorithm 4 of App. B.1. Here we describe the key steps. For each colour channel of each block, we:

1. Compute the seed using the get_seed function. This function should produce a stable seed even when the block has undergone minor modification, so that we can recover the seed that was used to embed the potential watermark without access to the original image.
2. Use the seed, along with the PRF and secret key sk, to determine the red/green partitioning of DWT coefficients for the block.
3. Run detection for each $k \times k$ sub-block: we apply d levels of DWT to the LL sub-band, then examine the LL sub-band at the last level, counting how many coefficients fall in the green (allowed) region.

From this block-level analysis, we can proceed to compute either a localised detection map or a global score for the entire image. We describe each of these processes in more detail below.

Localised Detection Map. For each $m \times m$ block, we aggregate the count of coefficients in the green region across all its sub-blocks. We set a detection threshold on this aggregate count based on a statistical hypothesis test with a prescribed significance level (see App. B.3 for details). The resulting detection map is generated at the block level: each $m \times m$ block is coloured green (indicating it is watermarked) if its aggregate green count exceeds the threshold, otherwise it is coloured red, as shown in Figs. 2 and 3.

Global Detection. While the localised detection map provides detailed information about the distribution of the watermark, a global measure is important for making a definitive assessment about whether the entire image is watermarked. To achieve this, we aggregate over the localised detection map, counting the fraction of blocks classified as watermarked (green). If the total count exceeds a predetermined threshold (calibrated to achieve a prescribed false positive rate), we predict that the entire image is watermarked; otherwise, we conclude it is not watermarked.

Remark 2. Since we embed the watermark into blocks throughout the image, we can implement a detection method that identifies watermarks even in cropped images. The method is based on a brute-force strategy to systematically search for the correct starting point of the embedded blocks, as described in Sec. B.6.

4.2.1. Interpreting the Localised Detection Map

The localised detection map provides valuable insights beyond global detection. It can be a helpful aid for image forensics, allowing users to identify specific locations

where a watermarked image may have been altered or where protected (watermarked) content has been inserted into a non-watermarked image.

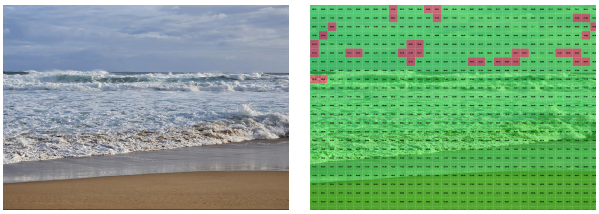
It is important, however, to acknowledge that false positives and false negatives can occur since the hypothesis test suffers some error depending on the chosen significance level. To enhance the visual clarity of our map we apply an optional post-processing step, akin to a low-pass filter: if a green block is entirely surrounded by red blocks, we consider it a likely false positive and recolour it red. Similarly, red blocks completely surrounded by green blocks are recoloured green. It is important to note that this post-processing is only done to improve the utility of the detection map as a visual aid—we do not apply it when computing the global score for the image.

Example: Non-watermarked Image. Fig. 2 displays a non-watermarked image alongside its localised detection map and an analysis of its DWT coefficients. The analysis (zoomed in panel) shows the frequency of rule violations averaged across the three colour channels. We observe that most blocks (over 91%) have balanced distributions of red and green domain coefficients. This balance is expected in natural non-watermarked images where approximately half of the blocks exhibit more than 50% violating coefficients, while the other half show less than 50%.

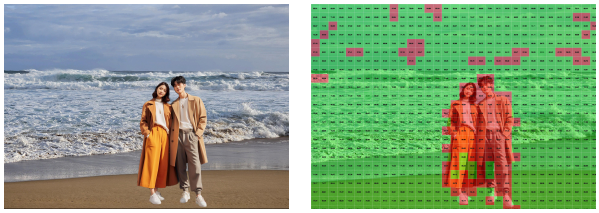
The raw localised detection map (to the right of the image) visually represents the result of our detection process. We see that most of the image's area is red, with only small, scattered green clusters. These green blocks, which make up only 8.3% of the total, and are scattered rather randomly across the image, represent natural statistical variations rather than the presence of a watermark.

It is worth noting that small clusters require careful interpretation. Dense clusters may naturally form in highly monochromatic areas with minimal variation, such as the black shadow in the top right corner of our example. In these regions, blocks tend to have similar block seeds (mean colours) and therefore similar DWT coefficients. If these coefficients happen to fall in the "watermarked" regions of our detector, a green cluster may appear even in non-watermarked images.

Example: Inserted Non-watermarked Content. Conversely, Fig. 3 demonstrates a scenario where non-watermarked content has been incorporated into a watermarked image. The detection map shows a prominent red cluster (corresponding to the inserted couple) within a predominantly green area. This demonstrates the utility of our localised approach for digital forensics and helps highlight the impact of manipulations on the distributed watermark.



(a) Watermarked image and the corresponding localised detection map generated during watermark detection process.



(b) Watermark detection process on an image that has been manipulated post-watermarking.

Figure 3. When editing new content into a watermarked image, the manipulation becomes immediately detectable. While some red squares (false negatives) are visible in the watermarked image (Fig. 3a), they are evenly scattered throughout the image rather than forming dense clusters. In Fig. 3b we inserted two people into the watermarked image from Fig. 3a. The inserted content, which does not contain the watermark, appears as a large red cluster in the detection map, clearly highlighting the area of image modification.

5. Experiments and Analysis

In this section, we evaluate our watermarking scheme's performance, focusing on its detectability, interpretability for image forensics, and robustness against common image manipulations. We begin by identifying a set of image manipulations that preserve visual quality, then assess our method across multiple datasets. We then assess watermark detectability across multiple datasets, examine our scheme's interpretability for image forensics, and evaluate its robustness against common image corruptions. The parameter settings used for our method are detailed in App. C.1.

5.1. Experimental Setup

5.1.1. Datasets

We employ two real-world datasets: MS-COCO [18], where we select the first 5,000 images from the test set and WikiArt [21] with the first 1,000 images. Additionally, we utilise two AI-generated datasets: 3,310 images generated by DALL-E 3 [22], accessed through Hugging Face Datasets [15] and 5,000 images from the DiffusionDB dataset [38]. This selection ensures that experiments cover a broad range of image resolutions. A distribution of image sizes for each dataset is provided in Fig. 11 in App. D.2.

5.1.2. Evaluation Metrics

For image quality assessment, we utilise three common metrics to compare watermarked images with their original counterparts: **PSNR**, where higher values indicate better image quality; **SSIM**, for which higher scores reflect greater structural similarity to the original image and **LPIPS**, where lower scores indicate closer structural similarity.

To evaluate the robustness of watermarking schemes, we employ two key metrics. **Watermark detection rate** (WDR), equivalent to the true positive rate (TPR), measures the proportion of watermarked images correctly identified as watermarked. It ranges from 0 to 1, with higher values indicating better performance. **False positive rate** (FPR) assesses the proportion of non-watermarked images incorrectly identified as watermarked. It also ranges from 0 to 1, with lower values indicating better performance.

5.1.3. Image Manipulations

To assess the robustness of our method, we consider a range of image manipulations, building on the comprehensive evaluation provided by [43]. These include:

- **Brightness** and **contrast** modifications with factor 0.5
- **JPEG compression**, using a quality setting of 50
- Adding **Gaussian noise** with a standard deviation of 0.05
- **Gaussian blur** with a kernel size of 5 and $\sigma = 1$
- **Rotating** the image by 90 degrees
- **VAE-based image compression models**, Bmshj18 [1] and Cheng20 [6], with a quality level of 3.

Our primary focus is on transformations that preserve visual quality. Some of these modifications, particularly the VAE-based compression models (Cheng20 and Bmshj18) and extreme brightness adjustments, can significantly alter the appearance of the image as shown in Fig. 10 of App. D.1. While we include these in our experiments for completeness and comparison with prior work, we consider them outside our primary scope of evaluation.

5.2. Detectability

To demonstrate the effectiveness of our watermark detection across different image types, we analyse the scores generated for both original and watermarked images. In this experiment, we embed the watermark with an interval length of $l = 8$ and conduct hypothesis tests using a 5% significance level for each block. We then calculate the percentage of watermarked blocks per image. The results are presented in Tab. 1. Our findings show a clear distinction between watermarked and non-watermarked images. For watermarked images, the percentage of detected watermarked area exceeds 90%, while for non-watermarked images, this percentage remains below 20%. This provides strong evidence that our watermark is reliably detectable. We additionally compute FPR and ROC AUC, where we observe perfect detection scores of 0% and 1 across all datasets.

5.3. Image Forensics

Our method has potential applications in image forensics owing to its ability to produce a localised detection map. As shown in Sec. 4.2.1, this map can highlight areas that may have been modified or inserted post-watermarking by analysing the distribution and clustering of red and green blocks. The same principle applies to subtle image manipulations, such as shape changes using common warp tools.

In Fig. 4, we present an example image of a woman with applied face modifications such as making the eyes and lips bigger, and slightly changing the shape of the nose. Our localised detection map clearly highlights the correct areas of modification. While our watermark is robust to noise or compression, which apply uniformly across the image and primarily affect higher frequency components, even small warping transformations can break it. This sensitivity occurs because we embed our watermark in the LL band of the DWT transform. The DWT decomposes the image into different frequency subbands, with the LL band representing the large-scale approximation of the image. This band contains the most significant information about the edges and shapes, and is sensitive to geometric transformations.

5.4. Robustness

5.4.1. Cropping

We evaluate our method for different cropping strengths and present the results in Fig. 8. Since we are using the brute-force detection algorithm presented in Sec. B.6, we work with a subsample of 1000 images per dataset. We set the stopping criterion to $p = 0.8$, which yields a FPR of ≈ 0.01 for each dataset. A cropping strength of *e.g.*, 70% means that we randomly remove 30% of the image area. Our results indicate that our method is highly robust against cropping, maintaining a TPR of $> 90\%$ for all datasets when up to 30% of the image area is cropped away. Even under radical cropping conditions where half of the image is removed, our method still achieves high TPR values ranging between 76% and 91%. While the performance on all datasets is strong, the results for WikiArt and DALL-E stand out as particularly robust. This can be attributed to their larger image sizes, providing more blocks for evaluation. A higher number of blocks leads to more stable detection statistics and reduces the influence of random variation.

5.4.2. Other Image Manipulations

In Tab. 2, we compare the performance of our method with other post-hoc watermarking approaches, based on results reported in [43]. Across all similarity metrics, our method achieves high imperceptibility—producing watermarked images that closely resemble the originals.

In terms of robustness, our method performs comparably to existing techniques when evaluated against common image transformations such as JPEG compression, Gaussian



(a) Watermarked image [4] (b) Edited watermarked image (c) Detection map

Figure 4. Our watermark demonstrates high sensitivity to image warping operations, enabling the detection of subtle manipulations.

Dataset	Orig.	Waterm.
COCO	18.62%	92.51%
DALL-E	16.82%	95.24%
DiffDB	18.46%	92.31%
WikiArt	19.49%	93.33%

Table 1. Watermarked area detected in original vs. watermarked images.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Pre-Attack	Rotation	JPEG	Blur	Contrast	Noise	Cheng20	Brightness	Bmshj18
MS-COCO												
MELB	36.07	0.99	0.007	1.000	0.973	0.969	0.991	0.160	0.877	0.020	0.606	0.041
DwtDct [8]	37.88	0.97	0.02	0.790	0.000	0.000	0.156	0.000	0.687	0.000	0.000	0.000
DwtDctSvd [8]	38.06	0.98	0.02	1.000	0.000	0.746	1.000	0.100	0.998	0.032	0.098	0.016
RivaGAN [42]	40.57	0.98	0.04	1.000	0.000	0.984	1.000	0.998	1.000	0.010	0.996	0.010
SSL [10]	41.81	0.98	0.06	1.000	0.952	0.046	1.000	0.996	0.038	0.000	0.992	0.000
CIN [19]	41.77	0.98	0.02	1.000	0.216	0.944	1.000	1.000	1.000	0.666	1.000	0.662
StegaStamp [34]	28.64	0.91	0.13	1.000	0.000	1.000	1.000	0.998	0.998	1.000	0.998	0.998
ZoDiac [43]	29.41	0.92	0.09	0.998	0.538	0.992	0.996	0.998	0.996	0.986	0.998	0.996
DiffDB												
MELB	43.45	0.97	0.03	1.000	0.973	0.961	0.979	0.091	0.882	0.020	0.570	0.032
DwtDct	37.77	0.96	0.02	0.690	0.000	0.000	0.224	0.000	0.574	0.000	0.000	0.000
DwtDctSvd	37.84	0.97	0.02	0.998	0.000	0.812	0.996	0.088	0.982	0.030	0.088	0.014
RivaGAN	40.60	0.98	0.04	0.974	0.000	0.898	0.966	0.932	0.958	0.004	0.932	0.008
SSL	41.84	0.98	0.06	0.998	0.898	0.040	1.000	0.996	0.030	0.000	0.990	0.000
CIN	39.99	0.98	0.02	1.000	0.212	0.942	0.998	1.000	0.998	0.478	1.000	0.662
StegaStamp	28.51	0.90	0.13	1.000	0.000	1.000	0.998	0.998	0.998	0.286	0.998	0.996
ZoDiac	29.18	0.92	0.07	1.000	0.558	0.994	1.000	0.998	0.998	0.988	0.998	0.994
WikiArt												
MELB	30.44	0.99	0.01	1.000	0.980	0.984	0.970	0.241	0.859	0.020	0.470	0.020
DwtDct	38.84	0.97	0.02	0.754	0.000	0.000	0.280	0.000	0.594	0.000	0.000	0.000
DwtDctSvd	39.14	0.98	0.02	1.000	0.000	0.698	1.000	0.094	1.000	0.000	0.096	0.034
RivaGAN	40.44	0.98	0.05	1.000	0.000	0.990	1.000	1.000	1.000	0.072	0.998	0.010
SSL	41.81	0.99	0.06	1.000	0.932	0.082	1.000	0.998	0.108	0.000	0.988	0.000
CIN	41.92	0.98	0.02	1.000	0.228	0.936	1.000	1.000	0.998	0.664	1.000	0.664
StegaStamp	28.45	0.91	0.14	1.000	0.000	1.000	1.000	0.998	1.000	1.000	1.000	1.000
ZoDiac	30.04	0.92	0.10	1.000	0.478	1.000	1.000	1.000	0.998	1.000	1.000	0.994
Colour keys:	Norms:	Best Value	within 2%			WDR:	< 0.70	0.70 - 0.85	0.85 - 0.95	>0.95		

Table 2. Comparison of watermarking schemes according to perceptibility (PSNR, SSIM, LPIPS) and robustness to image manipulations (WDR post-manipulation). A higher WDR post-manipulation means the watermarking scheme is more robust. Grey-shaded manipulations are considered out-of-scope due to their impact on image quality.

noise, and blur. While certain watermarks outperform ours in overall robustness, particularly under extreme perturbations, our method remains sufficiently resilient to modifications that do not substantially degrade image quality. Our approach is less robust to contrast and brightness adjustments. However, this aligns with our threat model, which assumes a malicious actor is unlikely to apply transformations that visibly distort the image. *E.g.*, a brightness factor of 0.5 darkens the image by 50%, resulting in a perceptually significant alteration. Since our method is designed to detect tampering that preserves visual plausibility, we consider this trade-off acceptable within our intended use case.

6. Conclusion

We propose a simple yet effective post-hoc image watermarking method, MELB, that combines localised embedding with region-level interpretability. When embedding, our method operates in the discrete wavelet transform domain, using a block-wise approach that randomly divides the transform coefficients into allowed and disallowed regions, then projects the image blocks onto the nearest allowed regions. For detection, we adopt a one-sided hypothesis test to determine whether an image is watermarked. Our method outputs not only an overall detection score for the whole image but also a detection map, highlighting the areas where the watermark is found, providing users more transparency. Compared to prior post-hoc methods, MELB offers more interpretable detection while retaining competitive robustness against rotation, JPEG compression, blur, and noise, all while introducing minimal visible artefacts. Notably, our method shows strong robustness to cropping, maintaining detectability even when up to half of the image is cropped away. Future work could explore alternative approaches to seed computation beyond our current method of using the rounded mean colour of a block, *e.g.*, investigating various hashing functions.

Acknowledgements

This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative. We also acknowledge the assistance of ChatGPT and GitHub Copilot for scaffolding code in the released supplementary material.

References

[1] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018. 7

[2] Abdullah Bamatraf, Rosziati Ibrahim, Mohd Salleh, and Mohd Najib. A new digital watermarking algorithm using combination of least significant bit (LSB) and inverse bit. *Journal of Computing*, 3(4):1–8, 2011. 1

[3] Mahbuba Begum and Mohammad Shorif Uddin. Digital image watermarking techniques: A review. *Information*, 11(2), 2020. 2

[4] Christopher Campbell. Picture of a woman with red hair. Unsplash Image, 2015. Retrieved January 18, 2025. 8

[5] Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. WavMark: Watermarking for audio generation, 2024. 1

[6] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7939–7948, 2020. 7

[7] Rita Choudhary and Girish Parmar. A robust image watermarking technique using 2-level discrete wavelet transform (DWT). In *2016 2nd International Conference on Communication Control and Intelligent Systems (CCIS)*, pages 120–124, 2016. 13

[8] Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. *Digital Watermarking and Steganography*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2 edition, 2007. 1, 2, 8

[9] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017. 1

[10] Pierre Fernandez, Alexandre Sablayrolles, Teddy Furon, Hervé Jégou, and Matthijs Douze. Watermarking images in self-supervised latent spaces. In *ICASSP 2022-IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3054–3058, 2022. 2, 8

[11] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023. 2

[12] Daniel Gruhl, Anthony Lu, and Walter Bender. Echo hiding. In *Information Hiding*, pages 295–315, Berlin, Heidelberg, 1996. Springer Berlin Heidelberg. 1

[13] Adil Haouzia and Rita Noumeir. Methods for image authentication: a survey. *Multimedia Tools Appl.*, 39(1):1–46, 2008. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 2

[15] Evgeniy Hristoforu. DALL-E 3 images. https://huggingface.co/datasets/ehristoforu/dalle-3-images, 2023. Accessed: January 17, 2025. 6

[16] Nikita Kashyap and G. R. Sinha. Image watermarking using 3-level discrete wavelet transform (DWT). *International Journal of Modern Education and Computer Science*, 4:50–56, 2012. 13

[17] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 17061–17084. PMLR, 2023. 1, 3, 5, 13, 14

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 6

[19] Rui Ma, Mengxi Guo, Yi Hou, Fan Yang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Towards blind watermarking: Combining invertible and non-invertible mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 1532–1542. ACM, 2022. 2, 8

[20] Minzhou Pan, Yi Zeng, Xue Lin, Ning Yu, Cho-Jui Hsieh, Peter Henderson, and Ruoxi Jia. JIGMARK: A black-box approach for enhancing image watermarks against diffusion model edits, 2024. 2

[21] Fred Phillips and Brandy Mackintosh. Wiki Art Gallery, Inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3):593–608, 2011. 6

[22] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 6

[23] Vinu Sankar Sadasivan, Mehrdad Saberi, and Soheil Feizi. IConMark: Robust interpretable concept-based watermark for AI images. *arXiv preprint arXiv:2507.13407*, 2025. 2

[24] Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre Défossez, Teddy Furon, and Tuan Tran. Proactive detection of voice cloning with localized watermarking. In *Proceedings of the 41st International Conference on Machine Learning*, pages 43180–43196. PMLR, 2024. 3

[25] Tom Sander, Pierre Fernandez, Alain Oliviero Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[26] Jordi Serra-Ruiz, Amna Qureshi, and David Megías. Entropy-based semi-fragile watermarking of remote sensing images in the wavelet domain. *Entropy*, 21(9), 2019. 11

[27] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: Protecting artists from style mimicry by Text-to-Image models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204. USENIX Association, 2023. 1

[28] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. 11

[29] Shuo Shao, Yiming Li, Hongwei Yao, Yiling He, Zhan Qin, and Kui Ren. Explanation as a watermark: Towards harmless and multi-bit model ownership verification via watermarking feature attribution. In *Proceedings 2025 Network and Distributed System Security Symposium*. Internet Society, 2025. 2

[30] Sunpreet Sharma, Ju Zou, Gu Fang, Pancham Shukla, and Weidong Cai. A review of image watermarking for identity protection and verification. *Multimedia Tools and Applications*, 83:31829–31891, 2024. 1, 2, 11

[31] Roop Singh, Mukesh Saraswat, Alaknanda Ashok, Himanshu Mittal, Ashish Tripathi, Avinash Chandra Pandey, and Raju Pal. From classical to soft computing based watermarking techniques: A comprehensive review. *Future Generation Computer Systems*, 141:738–754, 2023. 1

[32] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186. Association for Computing Machinery, 2020. 2

[33] Nandhini Subramanian, Omar Elharrouss, Somaya Al-Maadeed, and Ahmed Bouridane. Image steganography: A review of the recent advances. *IEEE Access*, 9:23409–23423, 2021. 1

[34] Matthew Tancik, Ben Mildenhall, and Ren Ng. StegaStamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2126, 2020. 1, 2, 8

[35] European Union. Regulation (eu) 2024/1689: Artificial intelligence act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689, 2024. Accessed: 2025-04-04. 1

[36] M. Utku Celik, G. Sharma, E. Saber, and A. Murat Tekalp. Hierarchical watermarking for secure image authentication with localization. *IEEE Transactions on Image Processing*, 11(6):585–595, 2002. 2

[37] Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1), 2020. 1

[38] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 893–911. Association for Computational Linguistics, 2023. 6

[39] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. In *Advances in Neural Information Processing Systems*, pages 58047–58063, 2023. 1, 2

[40] Xiang-Gen Xia, Charles G. Boncelet, and Gonzalo R. Arce. Wavelet transform based watermark for digital images. *Opt. Express*, 3(12):497–511, 1998. 2

[41] Heng Zhang, Chengyou Wang, and Xiao Zhou. A robust image watermarking scheme based on svd in the spatial domain. *Future Internet*, 9(3), 2017. 2

[42] Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention, 2019. 1, 2, 8

[43] Lijun Zhang, Xiao Liu, Antoni Martin, Cindy Bearfield, Yuriy Brun, and Hui Guan. Attack-resilient image watermarking using stable diffusion. *Advances in Neural Information Processing Systems*, 37:38480–38507, 2024. 1, 2, 7, 8, 15

[44] Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11964–11974, 2024. 2

[45] Y Zhao, B Liu, T Zhu, M Ding, X Yu, and W Zhou. Proactive image manipulation detection via deep semi-fragile watermark. *Neurocomputing*, 585:127593, 2024. 3

[46] Yihao Zheng, Haocheng Xia, Junyuan Pang, Jinfei Liu, Kui Ren, Lingyang Chu, Yang Cao, and Li Xiong. TabularMark: Watermarking tabular datasets for machine learning. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, pages 3570–3584. Association for Computing Machinery, 2024. 3