Department of Computer Science

Submitted in part fulfilment for the degree of BSc.

Identifying images generated by AI using watermarks

Mischa

Version 3.0, 2020-November

Supervisor: Dimitar Kazakov

To all students everywhere

Acknowledgements

I would like to thank my cat Orie for all the help it gave me writing this document.

As usual, my boss was an inspiring source of sagacious advice.

Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence

GenAI Generative AI

GAIM Generative AI Image Model

GAII Generative AI Image

VAE Variational Autoencoder

GAN Generative Adversarial Network

DM Diffusion Model

LDM Latent Diffusion Model

RGB Red, Green, Blue

LSB Least Significant Bit

DCT Discrete Cosine Transform

DFT Discrete Fourier Transform

DWT Discrete Wavelet Transform

LPM Log-Polar Mapping

MSE Mean Squared Error

PSNR Peak Signal-to-Noise-Ratio

SSIM Structural Similarity Index Measure

NCC Normalised Cross-Correlation

Executive Summary

1 Introduction

Generative AI Image Models (GAIMs) such as Midjourney, DALL-E [1], and Stable Diffusion [2] are capable of generating highly realistic images. While these technologies have enabled new possibilities in fields such as content creation, design prototyping, and automation, they also raise significant ethical concerns. A key issue is the potential misuse of these models for generating malicious or misleading content to deceive the public [3]. Current ruling by the US Copyright Office makes Generative AI Images (GAIIs) illegible for copyright protections [4].

In response to these concerns, watermarking has emerged as a promising solution for establishing authenticity, attribution, and traceability in GAIIs. The US White House recommended watermarking as part of its 2023 Executive order on Artificial Intelligence (AI) governance [5]. Similarly, the EU's AI Act 2024 mandates that providers of AI systems generating synthetic content embed invisible watermarks [6]. Leading tech companies have adopted watermarking methods: Microsoft watermarks images created in Bing [7], Google watermarks its Generative AI (GenAI) images using their SynthID [8], and Stability AI watermarks outputs from Stable Diffusion [9].

Digital watermarking involves embedding hidden information within digital content, a practice that dates back to physical media such as watermarked paper and photographs. This same need translated to digital media, particularly to protect intellectual property and prevent unauthorised use. Foundational work by Cox et al. [10] established core principles for digital watermarking, emphasizing resistance to common manipulations such as compression, resizing.

The widespread adoption of GAIMs intensifies concerns about content authenticity, copyright infringement, and attribution. As generative models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models (DMs) advance, the boundaries between human-created and AI-generated content are increasingly blurred. Artists have expressed concerns about unauthorised use of their work in AI training datasets , resulting in AI-generated images mimicking their styles without appropriate attribution or compensation . This has lead to calls for greater transparency and accountability in the creation and use of AI-generated content.

citation needed

citation needed

Adversarial image poisoning methods such as Glaze [11] and Nightshade [12] further showcase the desire for more transparency regarding image training data. These methods both use adversarial perturbations (tiny changes to the input) to change a given models perception of the original image [13].

This project aims to explore the integration of digital watermarking techniques with GAIMs to address issues of authenticity, attribution, traceability and copyright protection. Digital watermarking involves embedding information into digital media, enabling retrieval through specific algorithms while preserving the visual quality of the image. Secondly, the project aims to evaluate the effectiveness of said watermarking implementation through rigorous testing and analysis. Including assessing the robustness of the watermarks against various attacks, such as compression, noise addition, and cropping as well as its capacity for embedding meaningful amounts of data without compromising the visual quality of generated images.

2 Literature Review

2.1 Background

2.1.1 Generative Models in AI Image Generation

Generative images have revolutionised the AI field by enabling the creation of new data that closely resembles the training data. The three primary generative models used in AI image generation being: GANs, VAEs, and DMs.

Variational Autoencoders (VAEs)

VAEs were first defined in 2013 by Kingma et al. [14] and Rezende et al. [15]. VAEs are probabilistic generative models that learn a latent space representation of input data. They consist of an encoder and a decoder, which work together to reconstruct input data from a compressed latent space. Watermarking in VAEs could involve perturbing latent space to insert information [16].

Generative Adversarial Networks

GANs, proposed by Goodfellow et al. [17] in 2014. GANs consist of two neural networks: A generator, and a discriminator. The generator takes an input of random noise and generates an image by reassembling the real data distribution; The discriminator seeks to differentiate between real and generated images. The competing nature of the model helps in improving the quality of images generated over time. However, images generated are sensitive to perturbations [18]. Therefore, adding a watermark could degrade the quality of the image.

Diffusion Models

DMs were first introduced in 2015 by Sohl-Dickstein et al. [19] and popularised in 2020 by Ho et al. [20]. Unlike GANs and VAEs, DMs generate images by iteratively denoising a random noise pattern, producing high quality images with fine details. The challenge in watermarking DMs lies in ensuring the watermark does not interfere with the denoising process in order to preserve image quality. Latent Diffusion Models (LDMs) such as Stable Diffusion, extend the concept of diffusion by performing the denoising process in a latent space rather than directly in pixel space, allowing for more efficient training and inference, as well as improved image quality [2]. In LDMs, a VAE is used to compress the image to into a lower-dimensional latent space, which is then denoised iteratively to generate the final image.

2.1.2 Need for watermarking in AI era

2.1.3 Preventing Training Data Contamination

Lots of GAIMs rely on publicly available data scraped on the internet for training. When models are inadvertently trained on their own or other image models outputs, "model collapse" can happen, where repetitive cycles of training degrade the quality of generated content [21]. Watermarking GAIIs can mitigate this risk by marking synthetic outputs, enabling the dataset curation process to filter out these images for training.

2.1.4 Attribution and Traceability

Watermarks can act as unique identifiers for GAIIs.

Depending on implementation, watermarks could act as unique identifiers to identify the: model/platform, training data, user and time of generation for a given image. Providing greater accountability, especially relating to concerns of images being trained on licensed data.

The Coalition for Content Provenance and Authenticity (C2PA) is a cross industry initiative formed by major tech companies, including Microsoft, Arm, Intel and Adobe, aiming to establish a standard framework for certifying the origin and history of digital content. [22]

finish

Watermarking GAIIs has several critical applications, particularly in the realm of copyright protection, ownership, authenticity, and traceability.

An effective watermarking implementation would provide a means for

finish

A particularly interesting aspect of GAII watermarking is attribution, the ability to trace back the creator of a given image [23]

2.2 Watermarking Techniques

2.2.1 Spatial Domain Watermarking Techniques

Spatial domain watermarking involves embedding watermarks directly into the pixel values. These methods are straightforward, easy to implement, and computationally efficient. However, are typically less resistant to attacks such as compression and transformations.

Least Significant Bit (LSB) Modification

A common technique to embed a watermark information into randomly chosen pixels' Least Significant Bit (LSB). The LSB is changed as to not affect the image quality as it contains less important information. However, it is trivial for an attacker to change all LSB bits to 1 to modify the watermark. To address the problems with LSB watermarking, improvements have been made. One such improvement embeds data not only to the LSB but also higher planes. Moreover, a 2-3-3 embedding technique [24] distributes the watermark across the Red, Green, Blue (RGB) channels of a pixel. This approach results in minimal perceptual distortion while achieving better embedding capacity and robustness.

Patch-based or Block-Based techniques

Proposed by Bender et al. [25] This method involves randomly picking n pairs of image points A, B where the image data in A is darkened, while is brightened in B. This method offers decent robustness in exchange for capacity. [26]

2.2.2 Frequency (or Transform) Domain Watermarking Techniques

These techniques embed watermark information within the frequency domain of an image after a transformation. The transformation spreads the watermark information throughout the image in ways that are less perceptible to the human eye and harder to remove with common attacks.

Discrete Cosine Transform

Discrete Cosine Transform (DCT) watermarking embeds watermark information into an images frequency coefficients after transforming it from the spatial to the frequency domain. This leverages energy compaction, where the majority of an image's visual information is represented by lower-frequency coefficients, while higher-frequency coefficients capture finer image details. A common approach is block-based DCT, where the image is divided into smaller non-overlapping blocks, DCT is then applied to each block. Mid-frequency coefficients are typically chosen, balancing imperceptibility and robustness. Modifying low-frequency coefficients can lead to more noticeable distortions, while high-frequency coefficients are more susceptible to compression and noise attacks. Block-based DCT is particularly suitable for JPEG compression, a prevalent image compression technique which is also block-based [27]. By embedding watermarks in DCT coefficients compatible with JPEG's compression algorithm, the watermark can survive compression without significant degradation [28]. Alternatively, global DCT applies the transformation to the entire image rather than individual blocks. This offers greater robustness against attacks, but is more computationally intensive and less compatible with block-based compression techniques such as JPEG.

The robustness of DCT-based watermarking comes from the ability to embed data in perceptually significant regions of an image, therefore being less likely to be removed by common image processing operations. However, DCT based watermarking methods struggle with maintaining robustness against geometric attacks such as scaling and rotation due to inherently not accounting for spatial transformations [29]. From this hybrid techniques combining DCT with other transformations have arisen [30].

Discrete Fourier Transform

Similar to DCT watermarking, Discrete Fourier Transform (DFT) embeds watermark information into an images frequency domain by transforming

it from the spatial domain to the frequency domain, but using the DFT which decomposes an image into sinusoidal components of varying frequencies, represented as complex-valued coefficients corresponding to magnitude and phase. These coefficients describe the global frequency characteristics of the image, making DFT-based watermarking inherently robust against various image processing operations and certain geometric transformations.

Log-Polar Mapping (LPM) transforms the image into log-polar coordinates before applying DFT. This mapping converts scaling and rotation into linear translations in the frequency domain, enabling efficient watermark extraction after significant geometric transformations [31].

Discrete Wavelet Transform

A Discrete Wavelet Transform (DWT) is any wavelet transform that decomposes a signal into wavelets, offering local analysis in both the time and frequency domains. Unlike DFT, which analyses global frequency count, and DCT which can operate globally or block-based, DWT inherently supports multi-resolution analysis by examining signals at different scales. This dual localisation makes DWT particularly effective for image watermarking, as it can capture coarse and fine image details simultaneously.

2.2.3 State-of-the-Art Watermarking Techniques for AI-Generated Images

HiDDeN

One promising advancement in watermarking is the HiDDeN framework [32]. HiDDeN leverages the sensitivity of deep neural networks to small perturbations in input images to encode information, making it a robust solution for watermarking.

The HiDDeN framework comprises three main components: an encoder, a decoder, and an adversary network. The encoder receives an image and a message string, outputting an encoded image that incorporates the watermark. The decoder attempts to reconstruct the original message from the encoded image, while the adversary network predicts whether a given image contains an encoded watermark, providing adversarial loss to enhance the quality of the encoded images.

The adversarial training enhances the watermark's resilience against nu-

merous attacks. The deep leaning approach allows for a more flexible watermark embedding,

Tree-Ring

Tree-Ring [33] is a pre-generation watermarking method for DMs. The watermark is encoded in Fourier space and is decoded by inverting the diffusion process to receive the noise vector which can be compared against the embedded signal.

Finish description

Stable Signature

Stable Signature [34] is a watermarking approach specifically for LDMs. Stable signature embeds the watermark directly into the latent space of a model, rather than the pixel space, making it more resilient to post-processing transformations.

2.3 Watermarking Optimisations and Enhancements

2.3.1 Perceptual Masking

Perceptual masking exploits the characteristics of human vision by embedding watermarks into regions of an image where the changes will be less noticeable. For example. areas with high texture or edges rather than flat or uniform areas.

2.4 Watermarking Challenges

2.4.1 Attacks on Watermarks

2.5 Metrics for Evaluation

Performance metrics such as Mean Squared Error (MSE), Peak Signal-to-Noise-Ratio (PSNR), Structural Similarity Index Measure (SSIM), Normalised Cross-Correlation (NCC) are commonly used to evaluate the imperceptibility and quality of watermarked images. Performance metrics such as Mean Squared Error (MSE), Peak Signal-to-Noise-Ratio (PSNR), Structural Similarity Index Measure (SSIM), Normalised Cross-Correlation (NCC) are commonly used to evaluate the imperceptibility and quality of watermarked images.

3 Methodology

This chapter outlines the methodology employed to investigate and implement a digital watermarking technique suitable for embedding traceability and attribution information within images generated by AI models, specifically LDMs. The methodology follows a structured approach encompassing model selection, watermarking technique evaluation and selection, implementation planning, and a robust evaluation framework, and informed by the literature review (Chapter 2).

detail the methodology subsections

3.1 Choice of Generative Model

Given the project's focus on contemporary AI image generation and the need for techniques applicable to state-of-the-art models, LDMs were selected as the primary generative architecture for investigation. Specifically, Stable Diffusion [2] is chosen as a representative LDM due to its open-source nature [9], widespread adoption, and the availability of pre-trained models and research focused on watermarking on it [34], [35], [36]. LDMs operate by denoising data in a compressed latent space, offering potential integration points for watermarking that may be more robust than pixel-space methods [34], [37].

3.2 Watermarking Approach Selection

The primary goal is to embed imperceptible watermarks that facilitate traceability and attribution. Based on the literature review, several approaches are viable, particularly those designed for or adaptable to DMs/LDMs:

- **Latent Space Modification:** Techniques like Gaussian Shading [38], Stable Signature [34], Tree-Ring [33], LaWa [37], ZoDiac [36], WMAdapter [39], AquaLoRA [40], and FreqMark [16] propose embedding the watermark within the latent space during or slightly modifying the generation process. This approach is theoretically appealing for robustness as the watermark is integrated early.

- **Deep Learning Steganography:** Methods like HiDDeN [32] use encoder-decoder networks to embed and extract watermarks. While often applied post-generation, they could potentially be adapted or integrated with the LDM's components.
- **Frequency Domain Adaptation:** Traditional methods (DCT, DWT) might be adapted, perhaps applied within the latent space or integrated via learned components, although this is less explored in recent LDM-specific literature compared to direct latent modification.

The final selection of the watermarking technique will be based on the following criteria:

1. **Suitability for LDM Integration:** How readily can the technique be integrated into the Stable Diffusion architecture (e.g., VAE modification, U-Net fine-tuning)?
2. **Robustness Potential:** Theoretical and empirical evidence from literature regarding resistance to common image manipulations (compression, noise, resizing) and potentially model-based attacks.
3. **Capacity for Attribution Data:** Ability to embed a sufficient payload (e.g., unique identifier, timestamp) for traceability purposes [23].
4. **Imperceptibility:** Maintaining high visual quality of the generated images (measured by PSNR, SSIM).
5. **Implementation Feasibility:** Availability of reference implementations or clarity of the proposed algorithm within the project timeframe.

Based on these criteria, Gaussian Shading [38] is selected as the watermarking technique for implementation. Its approach modifies the initial latent sampling process within the LDM architecture, offering direct integration during generation (Criterion 1). A key advantage is its provably performance-lossless nature, meaning it does not require model fine-tuning and aims to preserve the original model's output quality and distribution (Criteria 1, 4, 5). The original paper reports high robustness against common distortions and good capacity (Criteria 2 & 3), making it a strong candidate for embedding attribution data without compromising the user experience or requiring additional training resources.

3.3 Implementation Plan

The implementation will proceed as follows:

1. **Environment Setup:** Configure a Python environment with necessary libraries including PyTorch, Diffusers (for Stable Diffusion), and potentially specific libraries for the chosen watermarking technique (e.g., LIEF if needed, libraries for image processing, evaluation metrics). Utilise available GPU resources for model training/fine-tuning and generation.

2. **Model Acquisition:** Obtain a pre-trained Stable Diffusion model (e.g., v1.5, v2.1, or SDXL) as the base generative model.

3. **Watermark Payload Definition:** Define the structure and content of the watermark message. For traceability and attribution, this could include a unique generation ID, a model identifier, and potentially a timestamp or user identifier (considering ethical implications). Error correction codes (e.g., BCH codes) will be incorporated to enhance robustness during extraction.

4. **Embedding Implementation:** Implement the watermark embedding mechanism based on the selected technique. This might involve:

    - **Latent Sampling Modification:** Following the Gaussian Shading approach [38], the initial latent variable z_T sampling step will be modified. This involves: (i) Diffusing the watermark bits s across the latent dimensions to get s_d. (ii) Randomizing s_d using a stream cipher (e.g., ChaCha20) with a secret key K to get m. (iii) Sampling z_T based on m using the Gaussian quantile function and uniform random sampling (distribution-preserving sampling). This replaces the standard random sampling of $z_T \sim \mathcal{N}(0, I)$.

Finalise watermark payload structure and size.

5. **Extraction Implementation:** Implement the corresponding watermark extraction algorithm. As per Gaussian Shading, this involves: (i) Using the standard VAE encoder to get z'_0 from the input image X'. (ii) Applying DDIM inversion to estimate the initial latent z'_T. (iii) Using the inverse sampling logic (Gaussian CDF) to extract the randomized watermark estimate m' from z'_T. (iv) Decrypting m' with the key K to get s'_d. (v) Applying a reduction/voting mechanism to recover the final watermark estimate s' from the diffused copies in s'_d.

6. **Generation Pipeline:** Integrate the modified latent sampling process into the standard text-to-image generation pipeline of Stable Diffusion. The standard ODE sampler (e.g., DPMSolver) and VAE decoder are used without modification after the initial z_T is sampled using the Gaussian Shading method.

3.4 Evaluation Framework

To assess the effectiveness of the implemented watermarking scheme, a comprehensive evaluation will be conducted, focusing on the core requirements for digital watermarks: imperceptibility, robustness, and capacity.

- **Dataset:** A standard dataset (e.g., MS-COCO captions, or a subset relevant to common AI generation prompts) will be used to generate a diverse set of watermarked images using various text prompts. A corresponding set of non-watermarked images will be generated for comparison.

- **Imperceptibility Assessment:**
  - Quantitative metrics: PSNR and SSIM will be calculated between original (non-watermarked) and watermarked images. Higher values indicate better imperceptibility.
  - Qualitative assessment: Visual inspection of watermarked images to identify any perceptual artifacts.

- **Robustness Testing:** Watermarked images will be subjected to a range of common image processing attacks and distortions:
  - Lossy Compression: JPEG compression at various quality factors (e.g., 90, 70, 50).
  - Noise Addition: Gaussian noise, Salt-and-pepper noise.
  - Geometric Transformations: Resizing, cropping, rotation.
  - Filtering: Gaussian blur.
  -

  The watermark will be extracted after each attack, and the Bit Error Rate (BER) between the original and extracted watermark payload will be calculated. Lower BER indicates higher robustness. Detection accuracy (whether the watermark presence is correctly identified) will also be measured.

- **Capacity Evaluation:** The size of the successfully embedded and extracted watermark payload (in bits) defines the capacity. This will be evaluated in conjunction with imperceptibility and robustness trade-offs.

- **False Positive Rate:** The watermark detector will be run on a set of

Specify dataset and number of images.

Consider adding other relevant attacks from literature/benchmarks, e.g., specific adversarial attacks, combo distortions.

non-watermarked images (both real-world images and AI-generated images from the base model without watermarking) to determine the rate at which it incorrectly detects a watermark. A low false positive rate is crucial for reliable attribution.

- **Computational Cost:** The overhead introduced by the watermarking process (embedding and extraction time) will be measured.

The results will be compared against baseline performance reported in the literature for the chosen or similar watermarking techniques.

3.5 Tools and Environment

- **Programming Language:** Python 3.x
- **Core Libraries:** PyTorch, Hugging Face Diffusers, NumPy, Pandas, Matplotlib, Scikit-image, OpenCV.
- **Watermarking Specific Libraries:** No specific external libraries are strictly required for Gaussian Shading beyond the core ML stack, as the implementation primarily involves modifying the sampling logic within the existing framework using PyTorch and standard libraries (e.g., for stream ciphers if not built-in).
- **Hardware:** Access to GPU resources (e.g., NVIDIA GPU via local machine or cloud service like Google Colab) is required for efficient model execution and potential fine-tuning.

3.6 Ethical Considerations

The development and evaluation will adhere to ethical guidelines. Data used for generation prompts will be sourced appropriately. The watermark payload design will consider privacy implications, particularly if user identifiers are included. The potential for misuse of the watermarking technology itself (e.g., attempts to forge watermarks) is acknowledged, although developing countermeasures against such misuse is beyond the scope of this project's implementation phase. The focus remains on providing a reliable mechanism for traceability and attribution as mandated by emerging regulations [5], [6].

Figure 3.1: A figure containing UoY logo and its caption.

4 Results

5 Conclusion and Future Work

A Some apendix

B Another apendix

Bibliography

[1] A. Ramesh et al. 'Zero-Shot Text-to-Image Generation.' arXiv: 2102.12092 [cs], Accessed: 9th Dec. 2024. [Online]. Available: http://arxiv.org/abs/2102.12092, pre-published.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer. 'High-Resolution Image Synthesis with Latent Diffusion Models.' arXiv: 2112.10752, Accessed: 28th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2112.10752, pre-published.

[3] E. Ferrara, 'GenAI Against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models,' *J Comput Soc Sc*, vol. 7, no. 1, pp. 549–569, Apr. 2024, ISSN: 2432-2717, 2432-2725. DOI: 10.1007/s42001-024-00250-1. arXiv: 2310.00737 [cs]. Accessed: 9th Dec. 2024. [Online]. Available: http://arxiv.org/abs/2310.00737.

[4] 'Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence,' Federal Register, Accessed: 28th Nov. 2024. [Online]. Available: https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence.

[5] T. W. House. 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,' The White House, Accessed: 28th Nov. 2024. [Online]. Available: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[6] *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)*, 13th Jun. 2024. Accessed: 5th Dec. 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj/eng.

[7] 'Bing Preview Release Notes: New experiences powered by Bing Image Creator,' Accessed: 28th Nov. 2024. [Online]. Available: https://blogs.bing.com/search/september-2023/Bing-Preview-Release-Notes-New-Experiences-Powered-by-Bing-Image-Creator.

[8] 'Identifying AI-generated images with SynthID,' Google DeepMind, Accessed: 14th Nov. 2024. [Online]. Available: https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/.

[9] *CompVis/stable-diffusion*, CompVis - Computer Vision and Learning LMU Munich, 28th Nov. 2024. Accessed: 28th Nov. 2024. [Online]. Available: https://github.com/CompVis/stable-diffusion.

[10] I. Cox, M. Miller, J. Bloom and M. Miller, *Digital Watermarking* (The Morgan Kaufmann Series in Multimedia Information and Systems). Morgan Kaufmann, 2001, ISBN: 978-0-08-050459-9. [Online]. Available: https://books.google.co.uk/books?id=uJcEWRv-RRUC.

[11] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka and B. Y. Zhao, 'Glaze: Protecting Artists from Style Mimicry by {Text-to-Image} Models,' presented at the 32nd USENIX Security Symposium (USENIX Security 23), 2023, pp. 2187–2204, ISBN: 978-1-939133-37-3. Accessed: 15th Nov. 2024. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/shan.

[12] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng and B. Y. Zhao. 'Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models.' arXiv: 2310.13828, Accessed: 15th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2310.13828, pre-published.

[13] A. Kurakin, I. Goodfellow and S. Bengio. 'Adversarial examples in the physical world.' arXiv: 1607.02533 [cs], Accessed: 15th Nov. 2024. [Online]. Available: http://arxiv.org/abs/1607.02533, pre-published.

[14] D. P. Kingma and M. Welling. 'Auto-Encoding Variational Bayes.' arXiv: 1312.6114, Accessed: 16th Nov. 2024. [Online]. Available: http://arxiv.org/abs/1312.6114, pre-published.

[15] D. J. Rezende, S. Mohamed and D. Wierstra. 'Stochastic Backpropagation and Approximate Inference in Deep Generative Models.' arXiv: 1401.4082, Accessed: 16th Nov. 2024. [Online]. Available: http://arxiv.org/abs/1401.4082, pre-published.

[16] Y. Guo et al. 'FreqMark: Invisible Image Watermarking via Frequency Based Optimization in Latent Space.' arXiv: 2410.20824, Accessed: 28th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2410.20824, pre-published.

[17] I. J. Goodfellow et al. 'Generative Adversarial Networks.' arXiv: 1406.2661, Accessed: 16th Nov. 2024. [Online]. Available: http://arxiv.org/abs/1406.2661, pre-published.

[18] M. Alfarra, J. C. Pérez, A. Frühstück, P. H. S. Torr, P. Wonka and B. Ghanem. 'On the Robustness of Quality Measures for GANs.' arXiv: 2201.13019, Accessed: 28th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2201.13019, pre-published.

[19] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan and S. Ganguli. 'Deep Unsupervised Learning using Nonequilibrium Thermodynamics.' arXiv: 1503.03585, Accessed: 16th Nov. 2024. [Online]. Available: http://arxiv.org/abs/1503.03585, pre-published.

[20] J. Ho, A. Jain and P. Abbeel. 'Denoising Diffusion Probabilistic Models.' arXiv: 2006.11239, Accessed: 16th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2006.11239, pre-published.

[21] M. Bohacek and H. Farid. 'Nepotistically Trained Generative-AI Models Collapse.' arXiv: 2311.12202, Accessed: 29th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2311.12202, pre-published.

[22] 'Content Credentials : C2PA Technical Specification :: C2PA Specifications,' Accessed: 9th Dec. 2024. [Online]. Available: https://c2pa.org/specifications/specifications/2.0/specs/C2PA_Specification.html.

[23] Z. Jiang, M. Guo, Y. Hu and N. Z. Gong. 'Watermark-based Detection and Attribution of AI-Generated Content.' arXiv: 2404.04254, Accessed: 14th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2404.04254, pre-published.

[24] G. R. Manjula and A. Danti, 'A novel hash based least significant bit (2-3-3) image steganography in spatial domain,' *IJSPTM*, vol. 4, no. 1, pp. 11–20, 28th Feb. 2015, ISSN: 23194103, 22775498. DOI: 10.5121/ijsptm.2015.4102. arXiv: 1503.03674 [cs]. Accessed: 3rd Dec. 2024. [Online]. Available: http://arxiv.org/abs/1503.03674.

[25] W. Bender, D. Gruhl, N. Morimoto and A. Lu, 'Techniques for data hiding,' *IBM Systems Journal*, vol. 35, no. 3.4, pp. 313–336, 1996, ISSN: 0018-8670. DOI: 10.1147/sj.353.0313. Accessed: 3rd Dec. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/5387237.

[26] M. Saqib and S. Naaz, 'Spatial and Frequency Domain Digital Image Watermarking Techniques for Copyright Protection,' vol. 9, pp. 691–699, 1st Jun. 2017.

[27] G. K. Wallace, 'The JPEG still picture compression standard,' *Commun. ACM*, vol. 34, no. 4, pp. 30–44, 1st Apr. 1991, ISSN: 0001-0782. DOI: 10.1145/103085.103089. Accessed: 4th Dec. 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/103085.103089.

[28] A. Bors and I. Pitas, *Image Watermarking Using DCT Domain Constraints*. 16th Sep. 1996, 234 vol.3, 231 pp., ISBN: 978-0-7803-3259-1. DOI: 10.1109/ICIP.1996.560426.

[29] S. Fazli and M. Moeini, 'A robust image watermarking method based on DWT, DCT, and SVD using a new technique for correction of main geometric attacks,' *Optik*, vol. 127, no. 2, pp. 964–972, 2016, ISSN: 0030-4026. DOI: 10.1016/j.ijleo.2015.09.205. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0030402615012863.

[30] A. K. Abdulrahman and S. Ozturk, 'A novel hybrid DCT and DWT based robust watermarking algorithm for color images,' *Multimed Tools Appl*, vol. 78, no. 12, pp. 17 027–17 049, 1st Jun. 2019, ISSN: 1573-7721. DOI: 10.1007/s11042-018-7085-z. Accessed: 10th Apr. 2025. [Online]. Available: https://doi.org/10.1007/s11042-018-7085-z.

[31] D. Zheng, J. Zhao and A. El Saddik, 'RST-invariant digital image watermarking based on log-polar mapping and phase correlation,' *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 8, pp. 753–765, Aug. 2003, ISSN: 1558-2205. DOI: 10.1109/TCSVT.2003.815959. Accessed: 7th Dec. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/1227605.

[32] J. Zhu, R. Kaplan, J. Johnson and L. Fei-Fei. 'HiDDeN: Hiding Data With Deep Networks.' arXiv: 1807.09937, Accessed: 14th Nov. 2024. [Online]. Available: http://arxiv.org/abs/1807.09937, pre-published.

[33] Y. Wen, J. Kirchenbauer, J. Geiping and T. Goldstein, 'Tree-Rings Watermarks: Invisible Fingerprints for Diffusion Images,' *Advances in Neural Information Processing Systems*, vol. 36, pp. 58 047–58 063, 15th Dec. 2023. Accessed: 10th Dec. 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/b54d1757c190ba20dbc4f9e4a2f54149-Abstract-Conference.html.

[34] P. Fernandez, G. Couairon, H. Jégou, M. Douze and T. Furon. 'The Stable Signature: Rooting Watermarks in Latent Diffusion Models.' arXiv: 2303.15435 [cs], Accessed: 27th Feb. 2025. [Online]. Available: http://arxiv.org/abs/2303.15435, pre-published.

[35] Y. Zhao, T. Pang, C. Du, X. Yang, N.-M. Cheung and M. Lin. 'A Recipe for Watermarking Diffusion Models.' arXiv: 2303.10137 [cs], Accessed: 10th Dec. 2024. [Online]. Available: http://arxiv.org/abs/2303.10137, pre-published.

[36] L. Zhang, X. Liu, A. V. Martin, C. X. Bearfield, Y. Brun and H. Guan. 'Attack-Resilient Image Watermarking Using Stable Diffusion.' arXiv: 2401.04247 [cs], Accessed: 3rd Dec. 2024. [Online]. Available: http://arxiv.org/abs/2401.04247, pre-published.

[37] A. Rezaei, M. Akbari, S. R. Alvar, A. Fatemi and Y. Zhang. 'LaWa: Using Latent Space for In-Generation Image Watermarking.' arXiv: 2408.05868 [cs], Accessed: 3rd Dec. 2024. [Online]. Available: http://arxiv.org/abs/2408.05868, pre-published.

[38] Z. Yang, K. Zeng, K. Chen, H. Fang, W. Zhang and N. Yu. 'Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models.' arXiv: 2404.04956 [cs], Accessed: 14th Apr. 2025. [Online]. Available: http://arxiv.org/abs/2404.04956, pre-published.

[39] H. Ci, Y. Song, P. Yang, J. Xie and M. Z. Shou. 'WMAdapter: Adding WaterMark Control to Latent Diffusion Models.' arXiv: 2406.08337 [cs], Accessed: 10th Dec. 2024. [Online]. Available: http://arxiv.org/abs/2406.08337, pre-published.

[40] W. Feng et al. 'AquaLoRA: Toward White-box Protection for Customized Stable Diffusion Models via Watermark LoRA.' arXiv: 2405.11135, Accessed: 28th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2405.11135, pre-published.