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

Executive Summary

Artificial Intelligence (AI) Generative AI (GenAI) Generative AI Image Model (GAIM) Generative AI Image (GAII) Variational Autoencoder (VAE) Generative Adversarial Network (GAN) Diffusion Model (DM) Latent Diffusion Model (LDM) Red, Green, Blue (RGB)

1 Introduction

GAIMs (Generative AI Image Model) 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 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 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 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 GANs, VAEs, and Diffusion Models 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: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models.

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)

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

Diffusion Models were first introduced in 2015 by Sohl-Dickstein et al. [19] and popularised in 2020 by Ho et al. [20] Unlike GANs and VAEs,

diffusion models generate images by iteratively denoising a random noise pattern, producing high quality images with fine details. The challenge in watermarking diffusion models lies in ensuring the watermark does not interfere with the denoising process in order to preserve image quality. 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

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.

Attribution and Traceability. Watermarks can act as unique identifiers for GAIIs. This co

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 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 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 (DCT)

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 (DFT)

Similar to DCT watermarking, DFT embeds watermark information into an images frequency domain by transforming it from the spatial domain to the frequency domain, but using the Discrete Fourier Transform 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 (DWT)

A 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 numerous 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.

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.

3 Methodology

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*, 2021. arXiv: 2102.12092 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2102.12092.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, *High-resolution image synthesis with latent diffusion models*, 2022. arXiv: 2112.10752 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2112.10752.

[3] E. Ferrara, 'Genai against humanity: Nefarious applications of generative artificial intelligence and large language models,' *Journal of Computational Social Science*, vol. 7, no. 1, pp. 549–569, Feb. 2024, ISSN: 2432-2725. DOI: 10.1007/s42001-024-00250-1. [Online]. Available: http://dx.doi.org/10.1007/s42001-024-00250-1.

[4] S. Perlmutter, *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence*, en, Mar. 2023. 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*, en-US, Oct. 2023. 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)*, en, Legislative Body: CONSIL, EP, 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*, en-US, Sep. 2023. 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*, en, Aug. 2023. Accessed: 14th Nov. 2024. [Online]. Available: https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/.

[9] *CompVis/stable-diffusion*, original-date: 2022-08-10T14:36:44Z, Nov. 2024. Accessed: 28th Nov. 2024. [Online]. Available: https://github.com/CompVis/stable-diffusion#reference-sampling-script.

[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,' in *32nd USENIX Security Symposium (USENIX Security 23)*, Anaheim, CA: USENIX Association, Aug. 2023, pp. 2187–2204, ISBN: 978-1-939133-37-3. [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*, 2024. arXiv: 2310.13828 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2310.13828.

[13] A. Kurakin, I. Goodfellow and S. Bengio, *Adversarial examples in the physical world*, 2017. arXiv: 1607.02533 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1607.02533.

[14] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, 2022. arXiv: 1312.6114 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1312.6114.

[15] D. J. Rezende, S. Mohamed and D. Wierstra, *Stochastic backpropagation and approximate inference in deep generative models*, 2014. arXiv: 1401.4082 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1401.4082.

[16] Y. Guo et al., *Freqmark: Invisible image watermarking via frequency based optimization in latent space*, 2024. arXiv: 2410.20824 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2410.20824.

[17] I. J. Goodfellow et al., *Generative adversarial networks*, 2014. arXiv: 1406.2661 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1406.2661.

[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*, 2022. arXiv: 2201.13019 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2201.13019.

[19] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan and S. Ganguli, *Deep unsupervised learning using nonequilibrium thermodynamics*, 2015. arXiv: 1503.03585 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1503.03585.

[20] J. Ho, A. Jain and P. Abbeel, *Denoising diffusion probabilistic models*, 2020. arXiv: 2006.11239 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2006.11239.

[21] M. Bohacek and H. Farid, *Nepotistically trained generative-ai models collapse*, 2023. arXiv: 2311.12202 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2311.12202.

[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 attribution of ai-generated content*, 2024. arXiv: 2404.04254 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2404.04254.

[24] M. G.R and A. Danti, 'A novel hash based least significant bit (2-3-3) image steganography in spatial domain,' *International Journal of Security, Privacy and Trust Management*, vol. 4, no. 1, pp. 11–20, Feb. 2015, ISSN: 2277-5498. DOI: 10.5121/ijsptm.2015.4102. [Online]. Available: http://dx.doi.org/10.5121/ijsptm.2015.4102.

[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. DOI: 10.1147/sj.353.0313.

[26] M. Saqib and S. Naaz, 'Spatial and frequency domain digital image watermarking techniques for copyright protection,' vol. 9, pp. 691–699, Jun. 2017.

[27] G. K. Wallace, 'The jpeg still picture compression standard,' *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[28] A. Bors and I. Pitas, 'Image watermarking using dct domain constraints,' in *Proceedings of 3rd IEEE International Conference on Image Processing*, vol. 3, 1996, 231–234 vol.3. 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: https://doi.org/10.1016/j.ijleo.2015.09.205. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0030402615012863.

[30] A. Abdulrahman and S. Ozturk, 'A novel hybrid dct and dwt based robust watermarking algorithm for color images,' *Multimedia Tools and Applications*, vol. 78, Jun. 2019. DOI: 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, 2003. DOI: 10.1109/TCSVT.2003.815959.

[32] J. Zhu, R. Kaplan, J. Johnson and L. Fei-Fei, *Hidden: Hiding data with deep networks*, 2018. arXiv: 1807.09937 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1807.09937.

[33] Y. Wen, J. Kirchenbauer, J. Geiping and T. Goldstein, 'Tree-rings watermarks: Invisible fingerprints for diffusion images,' in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine, Eds., vol. 36, Curran Associates, Inc., 2023, pp. 58 047–58 063. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/b54d1757c190ba20dbc4f9e4a2f54149-Paper-Conference.pdf.

[34] P. Fernandez, G. Couairon, H. Jégou, M. Douze and T. Furon, *The stable signature: Rooting watermarks in latent diffusion models*, 26th Jul. 2023. DOI: 10.48550/arXiv.2303.15435. arXiv: 2303.15435[cs]. Accessed: 27th Feb. 2025. [Online]. Available: http://arxiv.org/abs/2303.15435.