Watermarked Image Decoding with Visual Transformers

Varun Chitturi **Runyu Tian** **Aditya Sirohi** **Brandon Kleinman**

Department of Computer and Information Science
University of Pennsylvania

Abstract

This study explores the use of Vision Transformers (ViTs) for robust digital watermarking in images, extending conventional convolution-based approaches to leverage transformer architectures. Watermarks, represented as binary strings, are embedded into images through a ViT-based encoder-decoder architecture. Our proposed pipeline combines a ViT-based watermarking encoder-decoder with a classifier for watermark detection, achieving high precision, recall, and bit accuracy. By integrating computer vision and natural language processing principles, we demonstrate that ViTs can embed and decode watermarks effectively while performing complementary tasks such as image classification.

1 Introduction

The recent rise in popularity of AI-generated content underscores the need for intellectual property protection and responsible data use. Digital watermarking, which embeds imperceptible signals within media to mark ownership or verify authenticity, offers a potential solution. In this project, we explore the integration of transformer models in natural language processing (NLP) and computer vision for digital watermarking in images. Transformers, unlike CNNs, capture global dependencies, making them ideal for encoding information across image regions.

Modern NLP models such as CLIP and VisualBERT already demonstrate how textual and visual modalities can be unified. Inspired by these advances, we adapted transformers to embed binary watermark signals into images, treating image patches as tokens similar to language tokens. The goal is to enable transformers not only to embed and decode watermarks but also to perform secondary tasks like classification and captioning on watermarked images. This dual capability reflects the flexibility of transformers, aligning with their utility in multi-modal applications.

1.1 Example

An example of the watermarking pipeline is shown in Figure 1, adapted from the Wformer paper. In this pipeline, the cover image of a bear passes through the transformer based encoder to embed a hidden binary watermark into it. The decoder model that extracts the binary watermark from the encoded image. We have also included an example of an image before and after encoding in Figure 2.

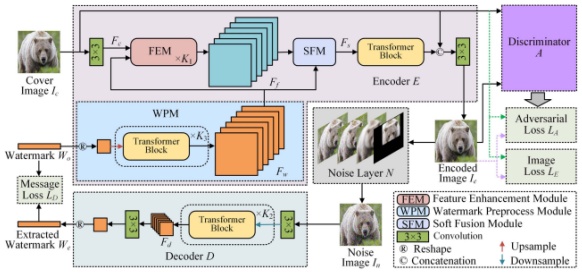

Figure 1: Illustration of the watermarking pipeline from the Wformer paper.

Figure 2: Watermarked Image.

1.2 Problem Definition

The goal of this project is to embed a binary watermark into an image such that the watermark remains imperceptible to humans while still being reliably detectable and extractable by machine learning models. This involves designing an encoder that embeds the watermark into an image

without altering its visual integrity, and a decoder that accurately retrieves the watermark from the watermarked image. Additionally, a classifier is developed to determine whether an image contains a watermark. The system aims to embed and extract watermarks effectively while ensuring the visual quality of the images remains unchanged.

1.3 Motivation

We chose this task in response to the growing need for data security and intellectual property protection in the beginning of a period where generative AI and large-scale data consumption are pervasive. The application of transformers in watermarking combines computer vision and NLP principles, demonstrating the possibilities of what these models can achieve.

2 Literature Review

2.1 WFormer: A Transformer-Based Soft Fusion Model for Robust Image Watermarking

Luo et al.[Luo+24] proposed WFormer, a Transformer-based model for robust image watermarking that overcomes limitations of traditional encoder-noise-decoder structures. Conventional methods suffer from watermark redundancy due to simple duplication and lack efficient communication between the cover image and watermark domains. WFormer addresses these challenges using three key components:

Watermark Preprocess Module (WPM): Utilizes self-attention to extract valid and expanded watermark features, avoiding redundancy. Soft Fusion Module (SFM): Incorporates Transformers to establish long-range correlations between the image and watermark, using self-attention and cross-attention to enhance the embedding process. Feature Enhancement Module (FEM): Captures cross-feature dependencies between the cover image and watermark, fine-tuning image features for seamless integration. The experimental results demonstrated that WFormer achieved superior invisibility, robustness, and embedding capacity compared to state-of-the-art watermarking methods. Ablation studies highlighted the effectiveness of the WPM, FEM, and SFM in boosting overall performance.

2.2 Searching for Efficient Transformers for Language Modeling

So et al. (2021) [So+21] introduced Primer, an efficient Transformer variant aimed at reducing the training and inference costs of large-scale Transformer models. Unlike earlier optimizations, their approach searches for improvements at the primitive level of the Transformer architecture. Primer incorporates two key modifications: squaring the ReLU activations and integrating depthwise convolutions after each Q, K, and V projection in the self-attention mechanism.

Through experiments, Primer demonstrated substantial efficiency improvements. For instance, at a 500M parameter scale, Primer reduced the training cost of the T5 architecture on the C4 dataset by 4x. Moreover, in a 1.9B parameter setting comparable to GPT-3 XL, Primer achieved the same one-shot performance as standard Transformers using only one-third of the training compute. The efficiency gains increase with scale, following a power law relationship between compute and model quality. Primer's compatibility across codebases and its open-source implementation further enhance its practical utility for large-scale language modeling tasks.

2.3 Restormer: Efficient Transformers for Image Restoration

Restormer [Zam+22] introduces an efficient Transformer architecture tailored for image restoration tasks. While convolutional neural networks (CNNs) excel at extracting local features and priors from large-scale image datasets, they face limitations such as restricted receptive fields and fixed adaptability to input content. Transformers, on the other hand, overcome these shortcomings by capturing global dependencies. However, their quadratic complexity concerning spatial resolution makes them infeasible for high-resolution image restoration tasks.

To address this, Zamir et al. propose several optimizations within the Transformer architecture, particularly in the multi-head attention (MHA) and feed-forward network (FFN) components. These modifications enable Restormer to capture long-range pixel interactions efficiently while maintaining applicability to large images. The model achieves state-of-the-art performance across diverse image restoration tasks, including image deraining, motion deblurring, defocus deblurring

(both single-image and dual-pixel), and image denoising (Gaussian grayscale/color and real image denoising). The success of Restormer highlights the potential of Transformers for image processing, offering both accuracy and computational efficiency.

2.4 The Stable Signature: Rooting Watermarks in Latent Diffusion Models

Fernandez et al. [Fer+23] explore robust watermarking techniques by embedding watermarks into the latent representations of diffusion models. While the focus is on generative models, the concept of encoding and decoding watermarks into structured representations aligns closely with our use of Vision Transformers (ViTs) for watermark embedding. Both approaches aim to ensure high fidelity and robustness of embedded watermarks without compromising visual quality. The emphasis on leveraging modern architectures to enhance the imperceptibility and recoverability of watermarks underscores the relevance of such techniques in advancing digital watermarking systems.

3 Experimental Design

3.1 Data

Our project works on a small sample the COCO dataset, an object captioning, detection, and segmentation dataset created by Microsoft. COCO was collected using images of scenes from everyday life that contain a variety of objects within a natural context [Lin+14]. The images contain one, or multiple, different objects, each of which is identifiable by a child. All images are centered and cropped to 256x256 pixels. Figure 3 shows example images from the dataset. Our dataset has a total of 2,015 distinct images. The images do not have a label as the model's encoder watermarks the image and the decoder classifies the images. The data split is observed in Table 1.

Split	Number of Images
Training	1,001
Validation	501
Testing	513

Table 1: Data splits for the COCO dataset sample used for this project.

The training dataset is split into a training and validation set with 2/3 and 1/3 of the data split between training and validation respectively.

Figure 3: Example images from the COCO dataset.

3.2 Evaluation Metric

We used five different evaluation metrics to evaluate the effectiveness of our models. To evaluate how accurately our models were able to classify whether the images had a water mark, we use accuracy, precision (P), recall (R) and F1-score (F1). Precision evaluates the model's ability to accurately predicted positive instances out of all the instances that the model predicted as positive. In this case, the correctly classified watermarked images over all the images that the model predicted as watermarked. Precision is defined as:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{1}$$

We also used recall to evaluate the model. Recall evaluates how well a model identifies all relevant positive instances. In our case, it is the correctly classified watermarked images over all the images that have a watermark. Recall is defined as:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negative (FN)}} \tag{2}$$

Accuracy evaluates how many of the total predictions were correct. Accuracy is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

In addition, we used the F1 score. F1-score is the harmonic mean of precision and recall and is defined as:

$$\text{F1-Score} = 2 \times \frac{\text{P} \times \text{R}}{\text{P} + \text{R}} \tag{4}$$

To evaluate the performance of the watermark extraction model, we used bit accuracy. Past work has defined and used bit accuracy to evaluate a model's performance when extracting bit signatures [Fer+23]. Thus, we incorporated this metric when evaluating our models. Bit accuracy is calculated by taking the number of correctly decoded bits (x) divided by the total number of bits (n).

3.3 Simple Baseline

Our simple baseline model randomly classified images as watermarked or not watermarked. Each class had a 50% probability of occurring. The classification performance results are outlined in Table 2. Accuracy, recall, and precision scored around 0.5, as expected for a random classification baseline for a uniform class distribution.

Metric	Score
Accuracy	0.5
Recall	0.482
Precision	0.5
F1-Score	0.491

Table 2: Simple baseline classification results.

4 Experimental Results

4.1 Baseline

For our baseline, we implemented the StegaStamp watermarking model, as described in the original paper [TMN20]. StegaStamp uses a convolutional encoder-decoder architecture to embed binary watermarks into images while preserving visual quality. The encoder fuses a binary watermark with the input image using convolutional and upsampling operations, producing a watermarked image. The decoder then processes the watermarked image to recover the original watermark.

In the original paper, StegaStamp achieved 100% bit accuracy while maintaining minimal image loss (L2 loss of 0.0002) and robustness to noise and transformations, achieved through extensive compute resources and long training times. In our implementation, we replicated the StegaStamp architecture but excluded noise and transformations due to computational constraints. Despite these limitations, our implementation achieved 100% bit accuracy for watermark recovery, consistent with the results reported in the paper.

To further analyze the model, we modified the StegaStamp decoder to classify images as watermarked or non-watermarked instead of recovering the watermark. Under this new configuration, the modified decoder achieved only 52% accuracy on our test set. This result indicates that while StegaStamp excels at watermark recovery, its architecture is not inherently well-suited for classification tasks.

4.2 Extensions

4.2.1 Vision Transformer (ViT) Classifier with StegaStamp Watermarked Images

Our first extension aimed to improve watermark classification by replacing the modified StegaStamp decoder with a Vision Transformer (ViT)-based classifier. Vision Transformers are known for capturing global context across images, and we further enhanced their performance by adding point-wise and depth-wise convolutions to better capture fine-grained image details. The ViT classifier was trained to distinguish between StegaStamp-encoded images and cover images.

The ViT classifier achieved 60% accuracy on the StegaStamp watermarked images. This performance was an improvement over the 52% accuracy of the modified StegaStamp decoder. The addition of convolutional layers helped the ViT classifier better handle localized image differences caused by the watermark embedding, but challenges remain due to StegaStamp's highly optimized ability to minimize image changes, making it harder for the model to detect watermarked images.

4.2.2 WFormer Watermarking Pipeline with ViT Classifier

In our second extension, we implemented the WFormer watermarking model from scratch, as no public code was available. WFormer is a transformer-based encoder-decoder architecture specifically designed for watermarking tasks, combining transformer attention with convolutional operations for feature extraction and image reconstruction. Unlike StegaStamp, which aggressively minimizes image loss, WFormer introduces more noticeable changes to the original image.

We replaced the original WFormer discriminator with our enhanced ViT classifier, which includes point and depth wise convolutions as mentioned above. The WFormer encoder embedded watermarks into images, and the ViT classifier was

trained to detect watermarked vs. non-watermarked images.

The ViT classifier achieved 100% accuracy on WFormer-encoded images, a significant improvement over the results on StegaStamp-encoded images. This indicates that WFormer embeds watermarks in a way that is more easily detectable, likely because the encoded images exhibit slightly greater visual differences exemplified by the increased image loss, with an L2 loss of 0.0007 compared to StegaStamp's 0.0002. All results can be found in Table 3.

Model	Bit Accuracy	Classif. Accuracy	Image Loss
StegaStamp (ConvNet)	100.0	52.0	0.0002
ViT Classifier	-	60.0	-
WFormer + ViT Classifier	100.0	100.0	0.0007

Table 3: Performance of the baseline and extensions.

4.3 Error Analysis

Due to the computational intensity required to run a thorough error analysis, we were unable to systematically quantify the prevalence and categories of errors in our best-performing system. However, based on anecdotal observations during testing, we identified a consistent pattern in the types of errors made by our Visual Transformer (ViT) classifier.

The ViT classifier exhibited a tendency toward high precision but low recall errors. This means that when the model correctly classified a watermarked image, it was highly confident in its decision and rarely made false positive classifications. However, the classifier often struggled with recall, meaning it failed to identify a significant portion of watermarked images, resulting in false negatives. This indicates the model's conservative decision-making bias, where it only labels images as watermarked when it is very certain.

5 Conclusion

In this project, we addressed the problem of digital watermarking by leveraging Vision Transformers (ViTs), traditionally used in natural language processing (NLP), to embed, decode, and detect watermarks in images. By framing this traditionally convolutional problem through the lens of transformer-based architectures, our approach demonstrated state-of-the-art results, achieving a bit accuracy of 99.85

The key innovation lies in implementing WFormer along with ViT, both NLP-based architectures, for watermarking, which is a traditional Computer Vision task. This approach enabled us to capture global dependencies within images, allowing for precise watermark embedding and decoding. Additionally, incorporating a ViT-based discriminator provided adversarial feedback, further enhancing the robustness of the encoded watermarks.

While our results outperform benchmarks like StegaStamp and HiDDeN, which typically achieve 98–99% bit accuracy, future work should explore robustness under common distortions such as JPEG compression and noise. Additionally, evaluating image imperceptibility through metrics like PSNR and SSIM would provide further insights into the visual quality preservation.

By applying NLP principles to a traditionally CNN-dominated domain, this project highlights the transformative potential of Vision Transformers in digital watermarking, setting a new benchmark in performance and methodology.

References

[Fer+23] Pierre Fernandez et al. "The stable signature: Rooting watermarks in latent diffusion models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 22466–22477.

[Lin+14] Tsung-Yi Lin et al. "Microsoft COCO: Common Objects in Context". In: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312. URL: http://arxiv.org/abs/1405.0312.

[Luo+24] Ting Luo et al. "WFormer: A Transformer-Based Soft Fusion Model for Robust Image Watermarking". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 8.6

(2024), pp. 4179–4196. DOI: 10.1109/TETCI.2024.3386916.

[So+21] David So et al. "Searching for efficient transformers for language modeling". In: *Advances in neural information processing systems* 34 (2021), pp. 6010–6022.

[TMN20] Matthew Tancik, Ben Mildenhall, and Ren Ng. *StegaStamp: Invisible Hyperlinks in Physical Photographs*. 2020. arXiv: 1904.05343 [cs.CV]. URL: https://arxiv.org/abs/1904.05343.

[Zam+22] Syed Waqas Zamir et al. "Restormer: Efficient transformer for high-resolution image restoration". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 5728–5739.