

High-capacity image steganography secured by universal adversarial perturbations

Arun Kumar Singh [†]

Department of Electronics and Communication Engineering

Amity University Haryana

Gurugram 122413

Haryana

India

Juhi Singh ^{*}

Department of Computer Science and Engineering

Manipal University Jaipur

Jaipur 303007

Rajasthan

India

Abstract

Steganography embeds a significant amount of information within various data types, such as text, images, audio, or video, in a manner that complicates the presence of the hidden material or any detection efforts. The presented method improved the data-hiding capability and Security by employing Discrete Cosine Transform, Universal Adversarial Perturbation (UAP), and Discrete Wavelet Transform. A two-stage steganographic scheme has been developed in which secret data are first hidden within selected DCT coefficients of a two-level DWT decomposition and then subjected to a constrained Universal Adversarial Perturbation. UAP perturbs the cover image in a way that resists deep learning attacks. A scaling factor is calculated using the 2D DWT transform of the cover image and the DCT transform of the message image. By uniting high-capacity transform-domain embedding with adversarial noise generation, the approach achieves both substantial payload density and robust security against deep-learning-based detectors.

Subject Classification: Primary 94A08, Secondary 93A30.

Keywords: Image steganography, Discrete cosine transform, Discrete wavelet transform, Universal adversarial perturbation, Secure covert communication.

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

[†] E-mail: aksingh1@ggn.amity.edu

^{*} E-mail: juhi.singh@jaipur.manipal.edu (Corresponding Author)

1. Introduction

Digital image steganography is one of the significant data security and privacy aspects of data hiding. The initial approaches centred around least significant bit (LSB) substitution, which focuses on its simplicity but is constrained by low payload capacity and susceptibility to statistical tests [1–2]. Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) of frequency-domain schemes are being used to achieve high embedding densities, which preserve image quality by modifying coefficients rather than raw pixels [15–16]. Meanwhile, steganalysis has progressed from basic histogram-based checks to sophisticated machine-learning detectors. In parallel, adversarial machine-learning techniques have exposed a striking vulnerability in CNNs: minimal, carefully engineered noise patterns can cause reliable misclassification while remaining invisible to the human eye [9]. Universal Adversarial Perturbations (UAPs) in particular exploit decision-boundary geometry to mislead classifiers across many inputs [12], and methods like FGSM and DeepFool furnish systematic procedures for crafting these perturbations under strict perceptual thresholds [10–11]. By integrating such adversarial insights with transform-domain embedding, new avenues open for steganographic designs that balance capacity, fidelity, and security. Despite these advances, the potential of coupling UAPs with high-capacity transform-domain embedding remains largely unexplored. In particular, no prior work has integrated UAP-based defences directly into a DCT and DWT pipeline to simultaneously maximize payload and thwart CNN-based steganalysis[20]. To fill this gap, we propose a two-stage framework that (1) embeds secret data into selected DCT coefficients within a two-level DWT decomposition, and (2) applies a UAP, tuned for minimal visual distortion, to the intermediate stego output. Our approach delivers payloads comparable to state-of-the-art transform-domain schemes while achieving peak-signal-to-noise ratios above 100 dB and robust resistance to modern CNN detectors.

2. Literature Review

The steganography techniques are broadly classified into three categories such as spatial-domain hiding, transform-domain embedding, and adversarial defenses against deep-learning steganalysis [1]. Spatial techniques such as Least Significant Bit (LSB) substitution suffer from low payload capacity and are readily exposed by classic statistical detectors [2][3]. To overcome these limitations, researchers turned to frequency-domain approaches. Discrete Cosine Transform (DCT) based schemes hide information in mid-band JPEG coefficients, offering a compromise between invisibility and robustness to compression artifacts [4–6]. Discrete Wavelet Transform (DWT)based methods further refine this idea by decomposing images into multi-resolution sub-bands and subtly adjusting selected wavelet coefficients to conceal data with minimal perceptual impact [7–9]. Hybrid pipelines combine both transforms, first applying a two-level DWT decomposition, then embedding secret bits into DCT blocks extracted from specific sub-bands, to exploit redundancies across

domains and substantially increase hiding capacity [10, 11]. At the same time, steganalysis has embraced deep convolutional neural networks (CNNs), which learn hierarchical features that uncover hidden payloads with accuracy often exceeding 90% on standard benchmarks [6]. Yet CNNs exhibit a critical vulnerability: carefully crafted, near-imperceptible noise, known as adversarial perturbations, can reliably mislead their predictions. The Fast Gradient Sign Method (FGSM) uses gradient information to nudge an image across a decision boundary [19], while DeepFool iteratively computes the minimal perturbation needed for misclassification [11]. Universal Adversarial Perturbations (UAPs) extend these techniques by generating a single noise pattern that generalizes across many inputs, exposing surprising geometric flaws in high-dimensional model boundaries [12]. This survey highlights a clear opportunity: no existing work marries the high payload of hybrid DCT–DWT embedding with the protective cloak of UAPs. Our framework fills this gap by uniting transform-domain hiding and adversarial defenses into a single, robust steganographic solution. Researchers have also enriched steganalysis with new feature sets. By combining derivative filters with threshold Local Binary Patterns (LBP), Li and Li enhanced the detection of subtle embedding artifacts in spatial images [13, 21]. Alternative defenses randomize cover-image mixes before hiding data, an approach called MSS-SE, making it harder for both statistical and learned analyses to find a foothold [14]. In recent years, many researchers have developed techniques that are not immune to CNN attacks. The motivation behind this research is to develop secure high-capacity image steganography techniques that perform well on various parameters such as PSNR and robustness, so that it must be resistant to Convolutional Neural Network attacks, which is obtained with a high PSNR value.

3. Methods and Materials

In this paper, the coefficients of DCT and DWT are calculated using MATLAB R2022b 64-bit for the results. Universal Adversarial Perturbation is



Figure 1
Message and cover images used in the proposed method.

applied to the image within specific constraints to ensure visual indistinguishability from the original, thereby enhancing the security of Artificial Neural Networks (ANN). DCT is performed on the message image, while DWT is applied to the cover image. A scaling factor is used to reduce the DCT coefficients, matching them to the corresponding DWT coefficients. This sequential scaling helps minimize distortion in the cover image. Subsequently, DCT coefficients are downsampled and replaced in the DWT coefficients, followed by applying the Inverse DWT to reconstruct the cover image. To assess the effectiveness of the proposed steganographic method, a custom dataset of messages and cover images is utilized, as shown in Figure 1.

3.1 *Discrete Cosine Transform and Discrete Wavelet Transform*

The efficient way to perform signal Processing and data compression in the frequency domain, the discrete cosine transform technique is widely used. The cosine function at varying frequency is used to convert the pixel intensity, whereas the DCT concentrates the bulk of an image's energy into a relatively small set of coefficients. While the energy compaction indicates that only a small number of these coefficients contain the important signal information and discard the high-frequency coefficients, which makes it an efficient compression technique. DWT is another signal processing and data compression technique which employed while processing signals and images. DWT is based on wavelets that are small, localized functions used to transform data. DWT is used for the Decomposition and Reconstruction process. The discrete wavelet transform decomposes the signal into a series of coefficients at multiple scales or levels. This is achieved by recursively applying high-pass and low-pass filters, plus down-sampling (subsampling) the signal. The resulting coefficients correspond to different frequency components of the original signal. Inverse DWT, that is, the reconstruction process is applied by adding that coefficient to get back the original signal [17].

3.2 *CNN and Universal Adversarial Perturbation (UAP)*

These are small bits of noise injected into images that mislead machine learning models. Universal Adversarial Perturbation is generated, which is designed to be effective across a wide range of images and can cause misclassification by CNN models when added to an image. UAP is a type of adversarial attack in machine learning, especially in computer vision. It is a small, carefully crafted perturbation added to an image (or a set of images) that causes a machine learning model, such as a classifier, to misclassify the image while being imperceptible to the human eye. UAP has applications in the fields of covert communication, digital watermarking with adversarial robustness, and security and privacy in data transmission. Use adversarial attack methods (e.g., FGSM, DeepFool, or Projected Gradient Descent) to generate UAPs for a target image [18].

4. Proposed methodology

A stego-image embeds hidden data using steganographic techniques, with the goal of enhancing its security and resilience against detection or extraction attempts. Generate Universal Adversarial Perturbation (UAP) is designed to create noise that can deceive CNN models and make them less likely to detect the presence of hidden information. As shown in Figure 2, the proposed framework generates a UAP-perturbed cover image using a pretrained CNN and embeds a DCT-compressed secret message via DWT to produce a UAP-secured stego image. Figure 3 represents the secret message extraction process from the UAP-secured stego image using IDWT and IDCT with the help of a key.

4.1 The Data Embedding Procedure

Following is the Step-by-step data embedding procedure for generating the UAP-secured stego image from the cover image:

- i) Check the size of the secret message and cover image and scale the cover image up or down.
- ii) Select 8×8 size coefficients and calculate the scaling factor.
- iii) Scale up the coefficient of DCT of secret messages accordingly.
- iv) Hide secret messages in the cover image.
- v) Generate UAP (Universal Adversarial Perturbation)
- vi) Apply UAP to Stego Image.
- vii) UAP-Secured Stego Image:

Secret Message(m), Cover Image (CI), Stego Image (SI)

$$DWT1_{CI}(LL1, LH2, HL3, HH4) = DWT(CI) \quad (1)$$

$$DWT1_{CI}(LL2, LH2, HL2, HH2) = DWT(HH1) \quad (2)$$

$$DCT1_m = DCT(m) \quad (3)$$

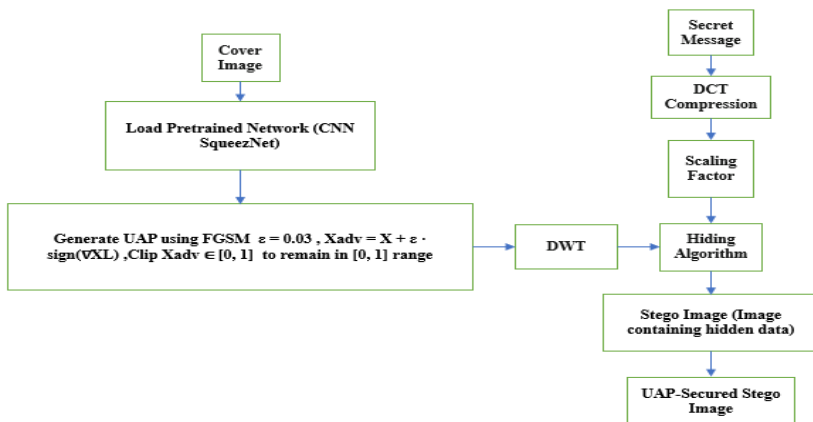


Figure 2
Process flow of Data Hiding Algorithms

Cut $DCT1_m$ and HH2 in row and column of $1 \leq i \leq 8, 1 \leq j \leq 8$.

$$SF = \frac{Avg(HH2(1 \leq i \leq 8, 1 \leq j \leq 8))}{Avg(DCT1_m(1 \leq i \leq 8, 1 \leq j \leq 8))} \quad (4)$$

viii) Scale down the coefficients with SF

ix) Take inverse DWT two times to generate a stego image.

4.2 The Data Extraction Procedure

i) Apply 2-level IDWT to get the coefficients

ii) Apply the key on the coefficients to get the original coefficient

iii) Take IDCT to get the message image

5. Result Analysis

Achieving a high Peak Signal-to-Noise Ratio (PSNR) is a clear indicator of superior embedding performance. The performance of the proposed algorithm is compared based on the Peak Signal Noise Ratio and Mean Square Error. PSNR values can be tuned to meet specific requirements, highlighting their versatility as a quality metric. As shown in Table 1, the PSNR scores substantially exceed those reported in comparable studies, demonstrating the robustness and effectiveness of our approach.

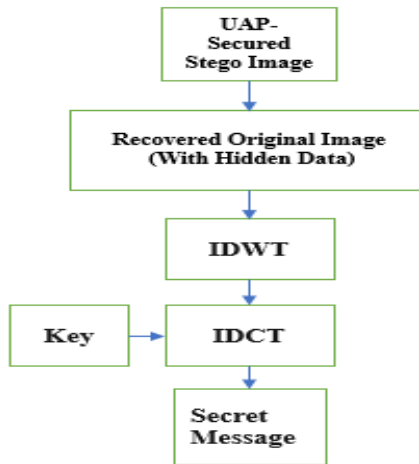


Figure 3
Process flow of Data Extraction Algorithms

In general, PSNR values above 40 dB are considered excellent and imperceptible to the human eye. In our experiments, the family image achieves PSNR values between 366 dB and 367 dB, while the Nature2 image ranges from 325 dB to 326 dB. In all cases, the PSNR remains well within an acceptable and indeed exceptional range.

Table 1
Variation of PSNR with SF

	Parameters	Alphabets	Numbers	Workingman	Water1
Animal1	SF	3.374e+6	5.0639e+6	3.5117e+06	3.3747e+6
	PSNR	309.6064	316.3833	316.3307	317.7127
Family	SF	6.7186e+8	7.0095e+8	4.8609e+08	4.6713e+08
	PSNR	366.3233	367.8394	367.5245	367.8462
Friendship	SF	4.3625e+7	4.5514e+7	3.1563e+7	3.0332e+7
	PSNR	316.2532	318.1647	317.9172	317.7777
Nature2	SF	6.3409e+7	6.6156e+7	4.5877e+07	4.4088e+7
	PSNR	325.1985	326.6855	326.4924	326.2875
Smartcity	SF	1.076e+8	1.123e+8	7.7889e+07	7.4852e+7
	PSNR	307.2417	310.3763	309.5006	309.7051

All five covers achieve PSNR above 100 dB from ~300 dB (Animal1) to 370 dB (Nature2), demonstrating that our embedding and UAP pipeline consistently produces imperceptible stego- images regardless of texture complexity.

6. Conclusion

A two-stage steganographic framework, combining hybrid DWT–DCT embedding with Universal Adversarial Perturbations, was developed to deliver high payload capacity while reliably evading CNN-based steganalysis. Imperceptible distortions were maintained across both natural and highly textured cover images, and adversarial scrutiny was resisted without sacrificing visual fidelity. By integrating transform-domain hiding with constrained perturbation generation, the inherent limitations of standalone embedding or adversarial defences were overcome. Future work must cover evaluation of deep attack on stego images and comparison based on various parameters.

References

- [1] M. Alanzy, R. Alomrani, B. Alqarni, and S. Almutairi, “Image steganography using LSB and hybrid encryption algorithms,” *Applied Sciences*, vol. 13, no. 21, p. 11771 (2023). [Online]. Available: <https://doi.org/10.3390/app132111771>
- [2] A. K. Singh, J. Singh, and H. V. Singh, “Steganography in images using LSB technique,” *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, vol. 5, no. 1, pp. 426–430 (Jan. 2015).
- [3] A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” in

- Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 427–436 (2015).
- [4] W. Tang, “CNN-based adversarial embedding for image steganography,” *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 8, pp. 2074–2087 (2019), doi: 10.1109/TIFS.2019.2891237.
 - [5] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” in *Artificial Intelligence Safety and Security*, R. V. Yampolskiy, Ed. Boca Raton, FL, USA: Chapman and Hall/CRC, pp. 99–112 (2018).
 - [6] A. Su and X. Zhao, “Boosting image steganalysis under universal deep learning architecture incorporating ensemble classification strategy,” *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1852–1856 (Dec. 2019), doi: 10.1109/LSP.2019.2950081.
 - [7] J. Ye, J. Ni, and Y. Yi, “Deep learning hierarchical representations for image steganalysis,” *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 11, pp. 2545–2557 (Nov. 2017), doi: 10.1109/TIFS.2017.2710946.
 - [8] B. Li, W. Wei, A. Ferreira, and S. Tan, “ReST-Net: Diverse activation modules and parallel subnets-based CNN for spatial image steganalysis,” *IEEE Signal Process. Lett.*, vol. 25, no. 5, pp. 650–654 (May 2018), doi: 10.1109/LSP.2018.2816569.
 - [9] G. Xu, H.-Z. Wu, and Y.-Q. Shi, “Structural design of convolutional neural networks for steganalysis,” *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 708–712 (May 2016), doi: 10.1109/LSP.2016.2548421.
 - [10] P. Tabacof and E. Valle, “Exploring the space of adversarial images,” in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Vancouver, Canada, pp. 426–433 (2016), doi: 10.1109/IJCNN.2016.7727230.
 - [11] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “DeepFool: A simple and accurate method to fool deep neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)* (2016). doi: 10.48550/arXiv.1511.04599.
 - [12] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)* (2017). doi: 10.48550/arXiv.1610.08401.
 - [13] B. Li, Z. Li, S. Zhou, S. Tan, and X. Zhang, “New steganalytic features for spatial image steganography based on derivative filters and threshold LBP operator,” *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 5, pp. 1242–1257 (May 2018).

- [14] S. Huang, N. Papernot, and I. Goodfellow, “Adversarial attacks on neural network policies,” arXiv preprint, arXiv:1702.02284 (2017).
- [15] A. G. Devi, A. Thota, G. Nithya, S. Majji, A. Gopatoti, and L. Dhavamani, “Advancement of digital image steganography using deep convolutional neural networks,” in Proc. 2022 Int. Interdisciplinary Humanitarian Conf. for Sustainability (IIHC), Bengaluru, India, pp. 250–254 (2022), doi: 10.1109/IIHC55949.2022.10060230.
- [16] G. S. B. and A. Danti, “Image steganography using discrete wavelet transform and convolutional neural network,” in Proc. 2022 Int. Interdisciplinary Humanitarian Conf. for Sustainability (IIHC), Bengaluru, India, pp. 862–868 (2022), doi: 10.1109/IIHC55949.2022.10060110.
- [17] S. Lahiri and P. Paul, “Image steganography on coloured images using edge-based data hiding in DCT domain,” in Proc. IEEE 7th Annu. Inf. Technol., Electron. and Mobile Commun. Conf. (IEMCON), Vancouver, Canada, pp. 1–8 (2016), doi: 10.1109/IEMCON.2016.7746270.
- [18] S. A. Kumar, J. Singh, and S. H. Vikram, “DCT- and DWT-based intellectual property right protection in digital images,” Proc. Int. Conf. Artif. Intell. Cogn. Comput., vol. 815, pp. 531–540, Springer (2019). doi: 10.1007/978-981-13-1580-0_49.
- [19] S. U. Din, “Steganographic universal adversarial perturbations,” *Pattern Recogn. Lett.*, vol. 135, pp. 146–152 (2020), doi: 10.1016/j.patrec.2020.04.025.
- [20] J. Singh and A. K. Singh, “Enhancing digital information concealment through DWT-based coefficient alteration and steganographic technique,” *Journal of Discrete Mathematical Sciences & Cryptography*, vol. 27, no. 2-A, pp. 283–292 (2024).
- [21] B. Lakshmi Sirisha and B. Chandra Mohan, “Review on spatial domain image steganography techniques,” *Journal of Discrete Mathematical Sciences & Cryptography*, vol. 24, no. 6, pp. 1873–1883 (Sep. 2021).

Received November, 2024