Deep Learning–Based Image Steganography Using Encoder–Decoder–Discriminator Architecture

Tejasvi Vasant Hegde *, Tarun R. *, Tanmay Dev D. *, Vinay V. Hegde

Posted Date: 9 February 2026

doi: 10.20944/preprints202602.0598.v1

Keywords: terms—steganography; deep learning; GAN; image hiding; encoder–decoder networks; information security

 doi:10.20944/preprints202602.0598.v1

Article

Deep Learning–Based Image Steganography Using Encoder–Decoder–Discriminator Architecture

Tejasvi Vasant Hegde *, Tarun R. *, Tanmay Dev D. * and Vinay V. Hegde

Dept. of Computer Science and Engineering, RV College of Engineering, Bengaluru, India

* Correspondence: tejasvivhegde.cs23@rvce.edu.in (T.V.H.); tanmaydevd.cs23@rvce.edu.in (T.R.); tarunr.cs23@rvce.edu.in (T.D.D.)

Abstract

Steganography enables covert communication by concealing information within innocuous media. Recent advances in deep learning have opened new avenues for high-capacity and imperceptible image steganography. This paper presents an experimental evaluation of a deep learning–based image-in- image steganography system employing an Encoder–Decoder– Discriminator architecture. The proposed framework embeds a secret image into a cover image to generate a visually indistinguishable stego image while enabling reliable recovery of the hidden content. The system is trained on a dataset of natural images using a composite loss function combining reconstruction, perceptual, structural similarity, and adversarial objectives. Quantitative evaluation using PSNR and SSIM metrics demonstrates strong imperceptibility of stego images and reasonable recovery fidelity of secret images. The results validate the feasibility of GAN-based approaches for practical image steganography and provide insights into their performance trade-offs.

Keywords: terms—steganography; deep learning; GAN; image hiding; encoder–decoder networks; information security

1. Introduction

Secure communication has long been a critical requirement in digital systems. While cryptography protects the content of a message, it does not conceal the existence of communication itself. Steganography addresses this limitation by embedding secret information within benign-looking media such as im- ages, audio, or video [1,2].

Traditional image steganography techniques, including Least Significant Bit (LSB) manipulation, suffer from limited payload capacity and vulnerability to statistical steganalysis [3]. With the advent of deep learning, data-driven approaches have emerged that learn optimal embedding strategies directly from data [4]. Neural networks can model complex spatial dependencies, enabling higher payload capacity while main- taining imperceptibility [5].

In this work, we evaluate a deep learning–based steganog- raphy system that hides an entire image within another image using an Encoder–Decoder–Discriminator architecture. Rather than proposing a new algorithm, this paper focuses on ex- perimentally validating the effectiveness of such architectures through systematic evaluation [6,7].

1.1. Problem Statement

Despite significant advances in image steganography, con- ventional techniques remain constrained by limited embedding capacity and susceptibility to detection through statistical analysis [3]. While deep learning–based steganography ap- proaches have shown promise in overcoming these limitations, many existing implementations emphasize architectural nov- elty without providing comprehensive experimental validation. Moreover, the trade-off between imperceptibility of the stego image and fidelity of the recovered secret image remains insufficiently explored in

 doi:10.20944/preprints202602.0598.v1

practical settings. There is a need for reproducible experimental studies that objectively evaluate encoder–decoder–discriminator architectures using standard- ized quantitative and qualitative metrics.

1.2. Hypothesis

This work is predicated on the hypothesis that a deep learning–based Encoder–Decoder–Discriminator architecture can effectively embed a full-resolution secret image within a cover image while maintaining high visual imperceptibility and achieving acceptable recovery fidelity. Specifically, adver- sarial training via a discriminator is expected to enhance stego image realism, while composite reconstruction and perceptual losses are hypothesized to facilitate meaningful recovery of the hidden secret image.

1.3. Objectives

The primary objectives of this study are as follows:

- To experimentally evaluate the imperceptibility of stego images generated by a deep learning–based steganogra- phy system using objective metrics such as PSNR and SSIM.
- To assess the fidelity of secret image recovery and analyze reconstruction quality under adversarial constraints.
- To investigate the practical trade-offs between visual im- perceptibility and recovery accuracy in encoder–decoder–discriminator architectures.
- To establish a reproducible experimental baseline that can serve as a reference for future research in deep learning–based image steganography.

2. Literature Survey

Early steganography techniques relied primarily on hand- crafted rules such as Least Significant Bit (LSB) substitution and transform-domain embedding methods [8,9]. LSB- based approaches embed secret information by modifying the least significant bits of pixel values, offering simplicity and low computational overhead. Transform-domain techniques, including Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) based methods, embed information in frequency components to improve robustness against basic signal processing operations. While effective in controlled sce- narios, these traditional approaches are inherently limited by low payload capacity and are highly susceptible to statistical steganalysis, as they introduce detectable artifacts into im- age statistics [10]. Consequently, such methods are generally unsuitable for high-security or high-capacity steganographic applications.

With the rise of deep learning, researchers began exploring data-driven steganography techniques that learn embedding strategies directly from image data. Autoencoder-based models were among the earliest neural approaches, framing steganog- raphy as a joint optimization problem of imperceptibility and recoverability. Baluja [4] demonstrated that deep neural networks could successfully hide one image within another while maintaining acceptable visual quality. Subsequently, adversarial learning was introduced to further improve im- perceptibility. Hayes and Danezis [5] proposed adversarial training frameworks where a discriminator attempts to distin- guish stego images from cover images, forcing the generator to produce more realistic outputs.

Encoder–decoder architectures have since become a dom- inant paradigm in deep learning–based steganography. These architectures learn joint latent representations that balance the competing objectives of hiding capacity and reconstruction fidelity. The integration of Generative Adversarial Networks (GANs) further enhances realism by introducing a discrimi- nator that penalizes perceptible deviations from natural image distributions. Zhang et al. [6] demonstrated that adversarial training significantly improves the visual indistinguishability of stego images. Similarly, Hu et al. [11] employed GAN- based frameworks to suppress detectable artifacts and improve resistance to steganalysis.

 doi:10.20944/preprints202602.0598.v1

Several studies have empirically shown that deep neural network–based steganography methods outperform traditional techniques in terms of payload capacity, robustness, and re- sistance to detection [7,12]. These methods leverage hierar- chical feature representations and non-linear transformations to distribute secret information more effectively across the image. However, despite promising results, many existing works emphasize architectural complexity or novelty while providing limited experimental validation. In particular, stan- dardized evaluation protocols, comprehensive metric reporting, and reproducibility remain inconsistent across studies. This gap highlights the need for systematic experimental evalu- ations that objectively assess the performance trade-offs of deep learning–based steganography systems, motivating the reproducible study presented in this paper [13].

3. Methodology

As shown in Figure 1, the proposed steganography framework follows an adversarial Encoder–Decoder–Discriminator archi- tecture.

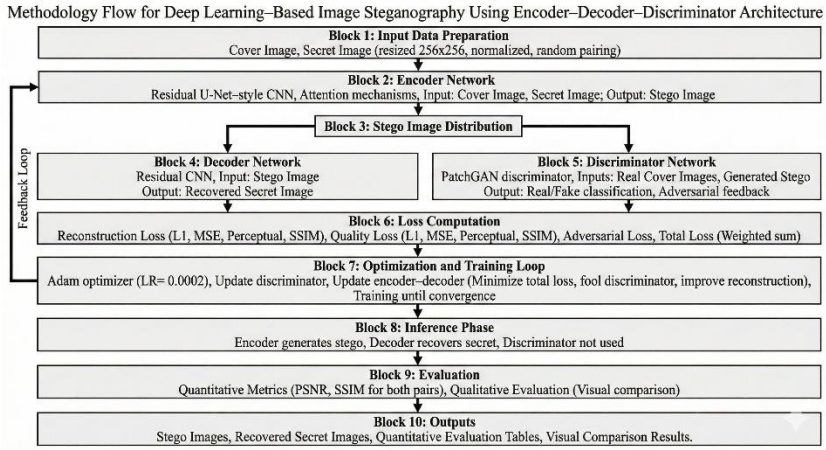

Figure 1. Overall methodology of the proposed deep learning–based image steganography system illustrating the Encoder–Decoder–Discriminator work- flow. The encoder embeds the secret image into the cover image to generate a stego image, the decoder reconstructs the hidden secret image, and the discriminator enforces visual realism through adversarial training.

3.1. System Architecture

The proposed steganography framework is built around an Encoder–Decoder–Discriminator architecture designed to embed a secret image within a cover image while preserving visual imperceptibility and enabling reliable recovery. The system operates in an end-to-end manner, jointly optimizing image hiding and extraction through adversarial training.

The **Encoder** is implemented as a residual U-Net–style convolutional neural network augmented with attention mech- anisms. It accepts a cover image and a secret image as inputs and learns to fuse the secret information into the spatial representation of the cover image. Residual connections facilitate stable gradient propagation, while the U-Net struc- ture enables multi-scale feature integration. Attention modules further guide the network to selectively embed secret features in perceptually less sensitive regions, resulting in a visually indistinguishable stego image.

The **Decoder** is a residual convolutional network tasked with reconstructing the secret image from the stego image. By learning inverse mappings of the embedding process, the decoder extracts hidden information while preserving the structural integrity of the recovered secret image. Residual connections enhance reconstruction stability and improve con- vergence during training.

The **Discriminator** follows a PatchGAN architecture and is trained to distinguish between original cover images and generated stego images. Rather than making a single global decision, the discriminator evaluates local image patches, en- couraging high-frequency realism and suppressing

 doi:10.20944/preprints202602.0598.v1

perceptible artifacts. This adversarial component compels the encoder to generate stego images that conform closely to natural image statistics.

Together, these components form a tightly coupled adver- sarial framework that balances embedding capacity, impercep- tibility, and recovery fidelity.

3.2. Data Preparation and Preprocessing

The model was trained on a dataset of natural images sourced from a randomized subset of images obtained via *picsum.photos*. This dataset provides diverse scene content and texture variations, enabling the model to generalize across different visual patterns. Both cover and secret images were resized to a uniform resolution of 256 × 256 pixels and normalized prior to training.

During each training iteration, a cover image and a secret image are randomly paired to prevent memorization and encourage robust feature learning. Standard data augmentation techniques were avoided to ensure controlled evaluation of embedding and recovery behavior.

3.3. Loss Function Formulation

The training objective is formulated as a weighted combina- tion of multiple loss components designed to jointly optimize imperceptibility and reconstruction accuracy. The total loss function is expressed as a weighted sum of the following components:

- **Reconstruction Loss:** Measures the discrepancy between the original secret image and the recovered image. This term combines L1 loss, Mean Squared Error (MSE), perceptual loss, and Structural Similarity Index Measure (SSIM) loss to capture both pixel-level accuracy and perceptual consistency.
- **Quality Loss:** Enforces visual similarity between the cover image and the generated stego image. Similar to the reconstruction loss, this term integrates L1, MSE, perceptual, and SSIM components to penalize visible distortions introduced during embedding.
- **Adversarial Loss:** Derived from the discriminator, this loss encourages the encoder to produce stego images that are indistinguishable from real cover images, thereby improving visual realism and resistance to detection.

The combined loss formulation enables the system to bal- ance competing objectives and achieve stable convergence during training.

3.4. Optimization Strategy

The network parameters of the encoder, decoder, and dis- criminator are optimized jointly using the Adam optimizer.

A learning rate of 0.0002 is employed for all components, providing a balance between convergence speed and training stability. Training proceeds iteratively, alternating between updating the discriminator and the encoder–decoder networks to maintain adversarial equilibrium.

This optimization strategy ensures effective adversarial learning while preventing mode collapse and excessive dis- tortion in stego images.

4. Implementation Details

The proposed steganography system was implemented using the PyTorch deep learning framework, enabling flexible model design and efficient GPU-accelerated training. The implemen- tation modularizes the encoder, decoder, and discriminator into separate components to facilitate independent experimentation and debugging.

The overall experimental workflow was automated through script-based execution, encompassing data preparation, model training, image hiding, secret extraction, and evaluation. Ded- icated scripts were employed to manage each stage of the pipeline, ensuring consistency across experiments and mini- mizing manual intervention. During training, model parame- ters were

 doi:10.20944/preprints202602.0598.v1

periodically saved as checkpoints to enable recovery, comparative analysis across epochs, and reproducibility of results.

Generated stego images and recovered secret images were systematically stored alongside their corresponding cover and original secret images. This organization enabled direct visual comparison and qualitative assessment of imperceptibility and recovery fidelity. Quantitative evaluation metrics, including PSNR and SSIM, were computed using standardized image quality assessment libraries and recorded for subsequent anal- ysis.

The experimental pipeline was designed to be fully repro- ducible. All preprocessing steps, model configurations, and evaluation procedures were fixed and automated, allowing experiments to be rerun under identical conditions. This re- producibility ensures that the reported results can be inde- pendently verified and provides a reliable baseline for future extensions or comparative studies.

5. Results and Analysis

Visual inspection further supports the quantitative findings. As shown in Figure 2, the stego image is visually indistinguish- able from the original cover image, exhibiting no perceptible artifacts. The recovered secret image retains the overall struc- ture and semantic content of the original secret image, with minor degradation in fine-grained textures, consistent with the reported PSNR and SSIM values.

Figure 2. Qualitative visual comparison of steganography results. From left to right: original cover image, generated stego image (PSNR = 35.74 dB, SSIM = 0.881), original secret image, and recovered secret image (PSNR = 21.34 dB, SSIM = 0.802). The stego image is visually indistinguishable from the cover image, while the recovered secret image preserves overall structure and semantic content with minor texture degradation.

5.1. Quantitative Evaluation

The steganography system was evaluated on a set of 20 unseen test image pairs to assess both imperceptibility and re- covery fidelity. The evaluation focuses on two widely adopted objective image quality metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM).

These metrics jointly capture pixel-level distortion as well as perceptual structural similarity.

Table 1 summarizes the quantitative evaluation results.

Table 1. Quantitative Steganography Evaluation Results.

Metric	Cover vs Stego	Secret vs Recovered
PSNR (dB)	35.71 ± 2.59	21.23 ± 2.72
SSIM	0.8845 ± 0.0532	0.7582 ± 0.1064

The high PSNR value of 35.71 dB and SSIM score of 0.8845 for cover versus stego images indicate strong imper- ceptibility. These values suggest that the embedding process introduces minimal distortion and preserves the perceptual characteristics of the original cover images. The relatively low standard deviation further indicates consistent performance across diverse image content.

 doi:10.20944/preprints202602.0598.v1

For secret image recovery, the PSNR of 21.23 dB and SSIM of 0.7582 demonstrate that the decoder is able to reconstruct the hidden content with reasonable fidelity. While these val- ues are lower than those observed for imperceptibility, they reflect the inherent challenge of recovering high-frequency details when adversarial constraints prioritize visual realism of the stego image. This trade-off between imperceptibility and recovery accuracy is a fundamental characteristic of deep learning–based steganography systems.

5.2. Visual Observations

Quantitative metrics are complemented by qualitative visual inspection to assess perceptual quality and reconstruction behavior.

Imperceptibility: Visual comparison between cover and stego images reveals no perceptible artifacts, distortions, or color inconsistencies under human inspection. The stego im- ages maintain texture continuity and structural coherence, confirming that the encoder successfully distributes the hidden information in a manner aligned with natural image statistics. **Recovery Fidelity:** Recovered secret images retain the over- all structure, semantic content, and major visual features of the original secret images. Minor degradation is observed in fine- grained textures and sharp edges, which is expected due to the adversarial training objective that prioritizes imperceptibility of the stego image. Despite this, the recovered images remain clearly recognizable, demonstrating the practical viability of the decoding process.

6. Discussion and Outcomes

The experimental findings validate the effectiveness of the Encoder–Decoder–Discriminator architecture for image steganography. The discriminator plays a pivotal role in en- forcing imperceptibility by penalizing deviations from natural image distributions, thereby compelling the encoder to gen- erate visually convincing stego images. Simultaneously, the decoder learns to extract hidden information despite these adversarial constraints.

The results highlight a clear and expected trade-off between stego image imperceptibility and secret image recovery fi- delity. While higher imperceptibility is achieved through ad- versarial training and perceptual losses, perfect reconstruction of the secret image remains challenging. This behavior is consistent with prior studies in GAN-based steganography and reflects the competing objectives inherent in the problem domain.

From a practical standpoint, the achieved balance is well- suited for covert communication scenarios where concealment of communication is prioritized over exact reconstruction fidelity. The relatively stable performance across test samples further indicates robustness of the learned embedding strategy.

7. Conclusions

This paper presented a comprehensive experimental evalua- tion of a deep learning–based image steganography system employing an Encoder–Decoder–Discriminator architecture. Extensive quantitative and qualitative analysis demonstrates that the proposed framework achieves strong imperceptibility of stego images while enabling meaningful recovery of hidden secret images.

The reported PSNR and SSIM metrics confirm that ad- versarial training effectively suppresses visible artifacts, and visual inspections corroborate the numerical findings. By emphasizing reproducibility and systematic evaluation rather than architectural novelty, this work establishes a reliable ex- perimental baseline for future research in deep learning–based steganography.

Future work may investigate robustness under common image transformations such as compression and noise, explore adaptive loss weighting strategies to improve recovery fidelity, and develop lightweight architectures suitable for real-time or resource-constrained deployment.

 doi:10.20944/preprints202602.0598.v1

References

1. N. F. Johnson and S. Jajodia, "Exploring steganography: Seeing the unseen," *IEEE Computer*, vol. 31, no. 2, pp. 26–34, 1998.
2. A. Cheddad, J. Condell, K. Curran, and P. Mc Kevitt, "Digital image steganography: Survey and analysis of current methods," *Signal Pro- cessing*, vol. 90, no. 3, pp. 727–752, 2010.
3. J. Fridrich, Steganography in Digital Media: Principles, Algorithms, and Applications. Cambridge University Press, 2009.
4. S. Baluja, "Hiding images in plain sight: Deep steganography," in
5. Advances in Neural Information Processing Systems (NeurIPS), 2017,
6. pp. 2069–2079.
7. J. Hayes and G. Danezis, "Generating steganographic images via adver- sarial training," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 1954–1963.
8. R. Zhang, S. Dong, and J. Liu, "Invisible steganography via generative adversarial networks," *Multimedia Tools and Applications*, vol. 78, no. 7, pp. 8559–8575, 2019.
9. J. Ye, J. Ni, and Y. Yi, "Deep learning hierarchical representations for image steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 9, pp. 2467–2480, 2019.
10. R. Chandramouli and N. Memon, "Analysis of LSB based image steganography techniques," in *Proceedings of ICIP*, 2001.
11. I. J. Cox, M. L. Miller, J. A. Bloom, J. Fridrich, and T. Kalker, *Digital Watermarking and Steganography*. Morgan Kaufmann, 2007.
12. J. Fridrich, M. Goljan, and R. Du, "Detecting LSB steganography in color and grayscale images," *IEEE Multimedia*, vol. 8, no. 4, pp. 22–28, 2001.
13. D. Hu, Q. Wang, and J. Huang, "GAN-based image steganography,"
14. *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1525–1529, 2018.
15. Z. Wu, X. Zhang, and Z. Wang, "A GAN-based image steganography method," *IEEE Access*, vol. 8, pp. 161636–161647, 2020.
16. M. Tancik, B. Mildenhall, and R. Ng, "StegaStamp: Invisible hyperlinks in physical photographs," in *Proceedings of CVPR*, 2020.