IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2434-2451

International Journal of Engineering

Journal Homepage: www.ije.ir

VidaFormer: A Hybrid Transformer for Image Steganography

V. Yousefi Ramandi, M. Fateh*, M. Rezvani

Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran

P A P E R I N F O

Paper history:
Received 23 July 2026
Received in revised form 15 December 2025
Accepted 04 January 2026

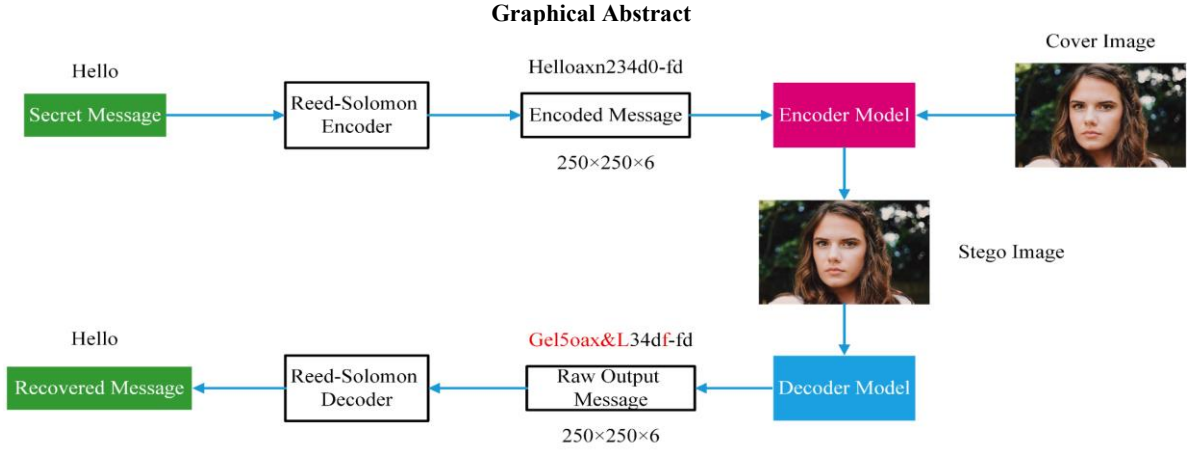
Keywords:
Image Steganography
Information Hiding
Deep Learning
Convolution
Transformers

A B S T R A C T

Image steganography has advanced significantly with the integration of deep learning, particularly through the use of convolutional neural networks (CNNs) for extracting complex features. While CNNs have long been a cornerstone of image analysis due to their efficiency in capturing local dependencies, the emergence of transformer models has revolutionized the field by achieving superior accuracy and effectively capturing global relationships within data. Transformers encounter issues with computational expense and memory consumption, especially when handling high-resolution images. To address these limitations, we propose a hybrid architecture that uses convolutional layers in high-resolution stages and transformer blocks in low-resolution stages. This design balances efficiency and performance by leveraging convolutional layers to capture short-range dependencies and transformers to model long-range relationships. Additionally, by replacing standard convolutions with CoordConv layers, we enhance the model's spatial awareness and feature localization capabilities. We introduce VidaFormer, an innovative steganography framework that integrates deep learning with these architectural enhancements to embed arbitrary binary data into images while maintaining high visual quality in the resulting stego images. VidaFormer outperforms state-of-the-art models, achieving an effective capacity of 4.89 bits-per-pixel on the DIV2K dataset, a 1.7% improvement over the previous best model. This demonstrates VidaFormer's superior performance and scalability for modern image steganography applications.

doi: *10.5829/ije.2026.39.10a.07*

Graphical Abstract



1. INTRODUCTION

Image steganography is the practice of embedding hidden information within digital images, making the concealed data imperceptible to the human eye. This technique is valuable for applications requiring secret data transmission (1) and digital watermarking (2, 3), as it provides a layer of secrecy without raising suspicion.

*Corresponding Author Email: mansoor_fateh@shahroodut.ac.ir (M. Fateh)

Please cite this article as: Yousefi Ramandi V, Fateh M, Rezvani M. VidaFormer: A Hybrid Transformer for Image Steganography. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2434-51.

Image steganography primarily involves two tasks: embedding and extraction. The embedding task, performed by the encoder (see Figure 1), conceals a message within a cover image to produce a stego image, which can then be transmitted through public channels to the receiver. The extraction task, executed by the decoder, retrieves the concealed message from the stego image, ensuring no part of the message is lost or corrupted. These tasks are inherently adversarial: making the stego image indistinguishable from the cover image may compromise the embedded message's integrity. Conversely, relaxing the similarity constraint between the cover and stego images can produce unrealistic stego images, potentially causing noticeable distortions.

Image steganography encompasses several key concepts. Embedding capacity, or simply capacity, measured in bits-per-pixel (BPP), determines how many bits can be hidden per pixel of an image. Recovery accuracy is defined as the fraction of correctly extracted message bits. To correct erroneous bits in the extracted message, error-correction algorithms are employed, which require adding extra data to the original message before embedding. Effective capacity, also measured in bits-per-pixel (BPP), represents the portion of the embedding capacity that corresponds to the original message, excluding additional error-correction data (4). It is determined by embedding capacity and recovery accuracy as discussed in the next the sections.

Transparency aims to make the concealed message harder for attackers to detect and recover. As demonstrated in the experiments, effective capacity and transparency objectives are adversarial: for a given model, increasing effective capacity reduces transparency, and vice versa. Robustness means being resilient to attacks such as noise or image compression (5). A steganographic attack on an image involves a third party degrading it in some way, resulting in a compromised image. A more robust algorithm experiences less reduction in capacity due to such an attack. However, robustness conflicts with capacity and transparency, so increasing robustness typically reduces the other two, and vice versa (6). Traditional methods, such as the least significant bit (LSB) approach, often rely on heuristic processing (7). These methods embed secret data by modifying the least significant bit of each pixel in an image, resulting in minimal visual distortion while covertly embedding the information. Although simple and effective, the LSB method is relatively vulnerable to detection.

Over the past few years, image steganography has transitioned from traditional embedding methods to more advanced deep learning-based approaches. These modern techniques focus on improving two key objectives: embedding information into images in a way that keeps the hidden data undetectable to unauthorized viewers and retrieving longer messages accurately without any loss of data.

Almost all recently proposed methods (8-10) leverage the power of deep learning to increase the capacity of data that an image can hold while maintaining its fidelity. These methods typically use convolutional layers in architectures designed for dense prediction, enabling the embedding and extraction of data at the pixel level. Common architectures employed in these approaches include U-Net (11), CSP-Net (12), and Dense-Net (13).

Although Convolutional Neural Networks (CNNs) have significantly propelled advancements in computer vision, their dependence on local receptive fields poses challenges, especially when it comes to effectively modeling long-range dependencies (14). These challenges can be effectively addressed by incorporating vision transformers, which leverage global attention mechanisms to enhance feature representation and spatial understanding.

Vision transformers (ViTs) (15) are a type of deep learning model that utilizes the self-attention mechanism instead of traditional convolutions to process images. These models divide images into patches and treat them as sequences, similar to natural language processing, allowing them to capture long-range dependencies and global context (16).

Newer transformers like Swin (17) and Swin V2 (18) introduced the concept of shifted windows to overcome the limitations of earlier models. Unlike ViT, which divides images into fixed-size patches and applies global self-attention across the entire image, Swin models use local windows to perform self-attention on smaller, non-overlapping regions. This local attention approach significantly reduces the computational complexity of vision transformer models, making Swin transformer more efficient and practical for various applications.

Although Swin transformers are more efficient than ViT models in terms of computational complexity and scalability, both architectures encounter significant challenges when dealing with high-resolution images. This limitation becomes particularly evident in tasks where preserving full-resolution data is critical, such as image steganography. Vision transformers encode each patch of pixels into an embedding vector, which makes restoring data to full resolution much harder (19, 20). We addressed this challenge in VidaFormer by using transformers only in the bottleneck stage.

To achieve superior results, we had to overcome the issue that standard convolutions lack an understanding of spatial positions (21). Augmenting the processing with positional encoding can improve the performance of convolutional layers. To achieve this, we used CoordConv (22) layers, which enhance traditional convolutional layers by explicitly incorporating spatial coordinates into the input, i.e. adding two additional channels to the input: One encoding the x-coordinates and the other encoding the y-coordinates of each pixel.

This enhancement allows the network to better capture spatial relationships, enabling it to handle tasks requiring precise localization more effectively.

This study presents VidaFormer, an innovative hybrid architecture for image steganography that combines convolutional methods with transformer-based approaches. VidaFormer sets a new benchmark by excelling in hiding capacity and recovery accuracy, all while maintaining the stego image's high fidelity. This approach is founded on a U-Net-inspired architecture, enhanced with CoordConv blocks and Swin V2 transformers, to create a robust and efficient framework for securely embedding and recovering hidden messages in images. To further enhance reliability, Reed-Solomon error correction (23) is utilized, enabling recovery even if portions of the message are corrupted. This combination of techniques addresses the dual challenges of fidelity and data accuracy in steganographic applications.

The proposed architecture marks a notable breakthrough in image steganography, achieving a balance between hiding capability and exceptional performance across various scenarios. By leveraging the strengths of U-Net, CoordConv, and Swin transformer, VidaFormer establishes a powerful framework for secure, high-capacity data hiding, setting a new benchmark for practical and resilient steganographic systems.

Our approach shows a remarkable enhancement in recovery accuracy, driven by the adoption of vision transformers for capturing long-range dependencies and the incorporation of CoordConv, which improves the model's spatial awareness. VidaFormer attains over 90% accuracy at 6 BPP (bits-per-pixel) capacity and achieves effective capacity of 4.89 BPP, substantially surpassing the performance of the current state-of-the-art models. The key contributions of this work are as follows:

(1) Developed a novel architecture integrating two hybrid models, each combining convolutional and transformer stages.
(2) Pioneered the use of CoordConv to provide positional awareness in the image steganography task.
(3) Achieved state-of-the-art accuracy, PSNR, and Reed-Solomon Bit Per Pixel (RS-BPP) on the DIV2K dataset.
(4) Evaluated the transparency of our method using a steganalysis tool to assess its effectiveness in concealing secret messages.
(5) Analyzed the trade-off between capacity and transparency by testing various transparency targets.

This paper is organized into five sections, each addressing key aspects of our research. The first section provides an overview, discussing the principles, challenges, and foundational concepts of our proposed method, while highlighting our main contributions. The second section reviews related works, offering a detailed analysis of prior methods, their strengths, limitations, and relevance to our approach. The third section describes the proposed method in depth, outlining its architecture, key innovations, and the rationale for our design choices. The fourth section presents a comprehensive evaluation of our model through experiments and ablation studies, demonstrating its effectiveness. Finally, the fifth section concludes the paper by summarizing findings and proposing future research directions.

2. RELATED WORK

In this section, different methods of image steganography are introduced. The discussion begins with statistical steganography, encompassing traditional techniques for hiding information within images. Modern approaches are then explored, including CNN-based methods that process images through convolutional layers, GAN-based techniques that generate realistic cover images using adversarial training, and finally vision transformers which use self-attention mechanism to capture long range dependencies inside an image.

2. 1. Statistical Steganography The least significant bit (LSB) method (7, 24-26) is the most commonly utilized technique in this category. This approach identifies bits within an image that can be altered without compromising its visual integrity and substitutes them with the binary representation of the secret message. Numerous variations of LSB-based methods have been proposed. For instance, Singh and Singh (27) embedded the secret message across the R, G, and B planes, while Das and Tuithung (28) utilized Huffman encoding to generate a binary form of the secret message, which is then embedded using LSB. A novel Reversible Data Hiding (RDH) method was proposed by Arham and Nugroho (29), combining difference expansion and modulus function to embed 3-bit data into 2-bit LSB differences of pixel pairs in rectangular blocks. It achieves high payload capacities with excellent visual quality. A novel steganography scheme was proposed by Chhajed and Garg (30) which uses a cover pattern histogram-based decision tree to hide and extract data in binary images, leveraging high-frequency 3x3 pixel block patterns and a secret key for encryption. It achieves high embedding capacity, minimal visual distortion (50-80% of bits embedded without pixel flipping), and enhanced security, outperforming similar methods for steganography and watermarking. A secure spatial steganography algorithm was introduced by Alexan et al. (31) that encrypts a secret message using tan logistic map, transforms it into QR codes, applies DNA coding with a Mersenne Twister key, and embeds

the resulting bit-streams into the LSB channel. Additionally, such as Qu et al. (32) and Wang et al. (33) explored LSB techniques within quantum images rather than the spatial domain.

Alternative methods utilize statistical analysis, including Pixel Value Differencing (PVD), a well-known technique that determines the difference between consecutive pixels to locate appropriate spots for embedding the secret message. Swain (34) introduced a hybrid approach combines LSB and PVD, applying LSB to the first two bits of every 8-bit segment of the secret message and PVD to the remaining six bits. Furthermore, some studies propose coverless solutions for image steganography, where a secret message is used to generate an image inherently designed to conceal the message (35). An errorless robust JPEG steganography method proposed by Qiu et al. (36) ensures message recovery after JPEG recompression by creating a robust set of DCT coefficients through lattice-based embedding. The approach guarantees error-free recovery while maintaining security and robustness against various JPEG compressors and quality factors. A steganography technique for JPEG2000 compressed images proposed by Butora et al. (37) embeds secret data into singular values obtained via singular value decomposition (SVD) in the discrete wavelet transform (DWT) domain, optimized by a genetic algorithm (GA) for scaling factor. It achieves up to 25% embedding capacity with higher PSNR than existing methods.

2. 2. Steganography by CNN These methods often rely on models designed to produce dense prediction outputs, typically using an encoder-decoder architecture. Within this framework, the secret message is concatenated with the cover image and processed to create the stego image. For instance, several researchers (10, 38, 39) employed U-Net as the basis for their encoder-decoder architectures. Rahim and Nadeem (40) presented a dual-path approach where two architectures process the cover image and secret message independently and simultaneously. A different approach was adopted by Wang et al. (41) which uses VGG (42) as the baseline to create a stego image styled after another input image. Alternative paradigms have also been explored; for example, Baluja (43) introduced three networks: one for data preparation, one for embedding the secret message, and one for extracting it. A deep learning-based steganography model was presented by Duan et al. (8) that can embed two secret images within a single carrier image and reconstruct them successfully. A Secret-to-Image Reversible Transformation (SIRT) framework for generative steganography was introduced by Zhou et al. (44), which encodes secret data into synthetic images with high embedding capacity, ensuring both imperceptibility and lossless data recovery while maintaining the fidelity and security of the generated images. A black-box generative steganography method was introduced by Zhang et al. (45) that leverages model volatility to mask steganographic modifications by modeling pixel distributions. The approach achieves high transparency and robustness against feature-based and CNN-based StegAnalyzer.

2. 3. Steganography by GAN These methods are built on Generative Adversarial Networks (GANs) (46), where the adversarial framework encourages the generator to produce high-quality, realistic images. Generative Steganographic Models (47-50) typically add a StegAnalyzer network that is responsible for detecting the presence of the secret message. An innovative approach by Bi et al. (51) integrates image style transfer into steganography, concealing secret information within stylized images while fusing secret and style features at multiple scales for improved concealment. A generic generative steganography framework was achieved by Su et al. (52) by enabling a distribution-preserving secret data modulator and a versatile secret data extractor. A novel approach by Ramandi et al. (49) introduced balancing transparency and capacity during training by dynamically adjusting the weights of the respective losses. A reversible generative steganography method by Tang et al. (53) proposed a distribution-preserving message mapping strategy and a reversible Glow model to encode secret information into latent vectors, achieving high security and accurate extraction.

2. 4. Steganography by Transformers Vision transformers (ViTs) (15) have transformed the field of computer vision by utilizing self-attention (16) mechanisms to effectively capture both local and long-range dependencies in image data, surpassing convolutional networks in tasks such as image classification. The Swin transformer family (17-19) employs a hierarchical structure with shifted windows, allowing for improved scalability and efficiency in dense prediction tasks.

On the steganography front, a transformer-based steganography scheme for enhanced feature extraction was introduced by Wang et al. (54). It also incorporates a recursive permutation image encryption algorithm to strengthen the security of secret images. Zhou et al. (55) used a channel-wise attention transformer model for image steganography that constructs long-range dependencies and optimized data embedding by considering global and local features and a combination of channel self-attention, non-linear enhancement, cross-attention, and global-local aggregation modules to improve data concealment. A transformer-based adversarial network steganography framework proposed by Xiao et al. (56) used attention mechanisms, maintains spatial relationships, and employs a WGAN discriminator to improve imperceptibility.

3. PROPOSED METHOD

3. 1. Overview In this research, we propose the Versatile Dynamic Transformer (VidaFormer) architecture, which achieves high hiding capacity while maintaining high image fidelity, as measured by PSNR (57) and SSIM (58), two primary metrics for image quality assessment (see Section [Evaluation metrics]). It employs MSE targeting, introduced by Ramandi et al. (49), to dynamically adjust the balance between effective capacity and transparency based on a predetermined transparency target.

The overall architecture is presented in Figure 1. For the embedding task, the encoder E takes the secret message M and the cover image C to generate the stego image S, as shown in Equation 1.

$$S = E(C, M) \quad (1)$$

For the extraction task, the decoder D takes the stego image and extracts the recovered message R, as in Equation 2.

$$R = D(S) \quad (2)$$

To enforce , it is crucial that the stego image closely resembles the cover image, making it hard to distinguish between the two.

The encoder has two distinct and competing objectives. On the one hand, it must produce a stego image that closely resembles the cover image, making it difficult for an attacker to detect the presence of a secret message, thereby achieving transparency. Conversely, the encoder must embed the secret message into the stego image for later extraction. To ensure effective recoverability and capacity, the encoder needs to alter pixel values to some degree, which can distort the original cover image and compromise transparency.

The decoder retrieves the original message by decoding the stego image and identifying patterns of hidden content embedded by the encoder. The accuracy of message recovery depends on the level of distortion, which is inversely related to the stego image's transparency. The algorithm's effective capacity is proportional to recovery accuracy; thus, increasing effective capacity reduces transparency and vice versa. Transparency is enforced through a loss function that penalizes pixel value differences between the cover and stego images. Another loss function, based on the difference between the recovered and secret messages, ensures recoverability.

3. 2. Encoder Architecture The encoder, depicted in Figure 2, embeds a binary secret message into a cover image, resulting in the creation of the stego image. Its architecture is inspired by U-Net (11), comprising a feature extraction path (left blocks), a bottleneck (central orange blocks), and a reconstruction path (right blocks). Data undergoes several stages of processing before being reconstructed to generate the desired output.

Initially, the cover image, with three RGB channels, and the secret message, with six channels each representing one bit per pixel, are concatenated to form a single nine-channel feature map. In this configuration, 6 bits are embedded per pixel, resulting in a 6 bits-per-pixel (6-BPP) embedding capacity setting.

In the first stage, three 48-channel convolutional blocks operate at full resolution. Each block, represented by a blue box, consists of three convolutional layers, resulting in a total of nine convolutional layers in this stage. The first convolution in the initial block takes nine input channels and produces 48 output channels, accommodating the channel count difference. The model benefits from extensive processing at full resolution, ensuring comprehensive pixel analysis and effective capture of local relationships between neighboring pixels.

At the beginning of the second stage, a max-pooling layer reduces the feature map's resolution by half. This operation expands the receptive field, enabling the model to capture broader contextual information while lowering computational costs by progressively reducing feature map resolution. The first convolution in the initial block doubles the channel count to 96, enhancing the model's ability to learn high-level

features. Consequently, with W and H representing the input image's width and height, the feature map's shape transitions from (W, H, 48) in the first stage to (W/2, H/2, 96) at the beginning of the second stage. The second stage comprises six convolutional blocks. Feature maps at this stage are more complex, necessitating increased processing, which justifies the greater number of blocks compared to the first stage. The expanded receptive field further enables the model to capture long-range pixel relationships.

At the start of the bottleneck stage, the feature map is reshaped from (W/2, H/2, 96) to (W/4, H/4, 384). As standard in vision transformers (18) (15), this reshaping before the transformer blocks incurs no data loss. Using a 2×2 patch size, the resolution is halved, and the channel count is quadrupled. Positioned between the feature extraction and reconstruction paths, the bottleneck stage (center orange blocks) employs three Swin transformer V2 blocks (18), optimized to capture long-range dependencies and enhance the model's ability to process global contextual information. By incorporating these blocks, the model leverages self-attention mechanisms to detect and encode complex patterns across the cover image, ensuring secure information embedding while preserving visual integrity.

The encoder has completed capturing all necessary information.

A two-stage reconstruction path (right blue blocks) uses this information to generate the embedded stego

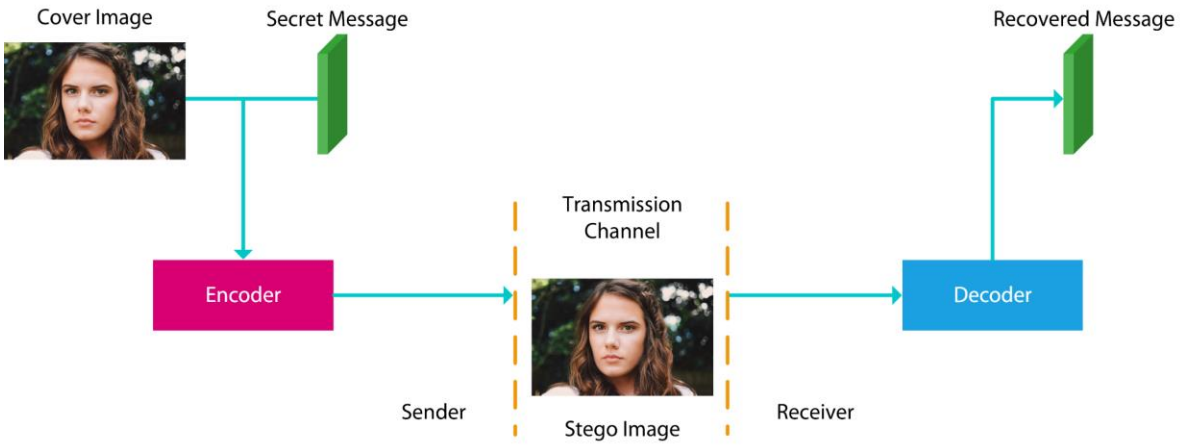


Figure 1. The overall process of image steganography

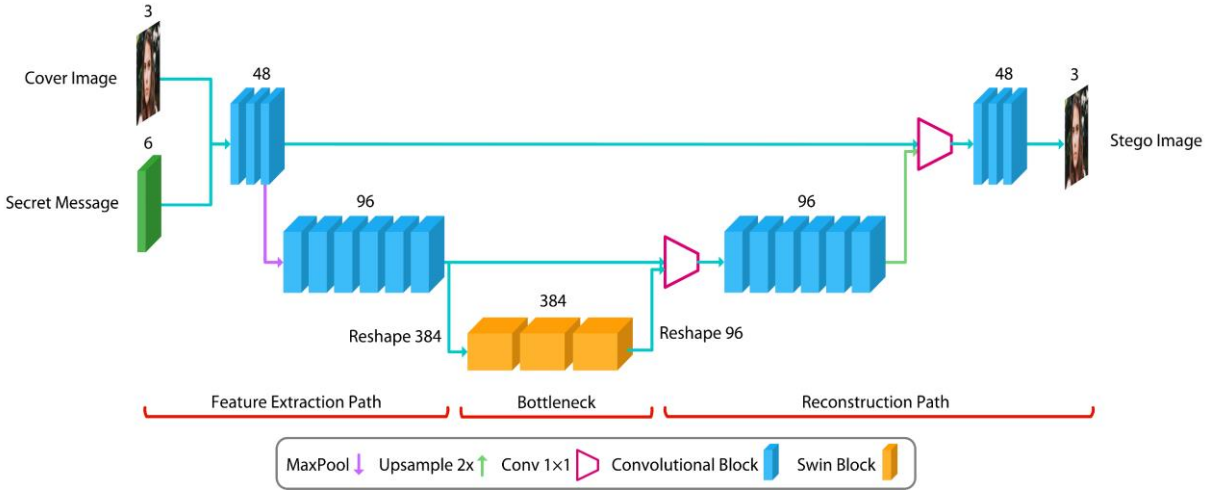


Figure 2. The architecture of the encoder

image. First, the bottleneck stage output is reshaped from (W/4, H/4, 384) to (W/2, H/2, 96) to align with the corresponding feature extraction path output and concatenated with it. The resulting feature map combines high-level information from the bottleneck with low-level information from the feature extraction path. A 1×1 convolution (first magenta trapezoid) projects the feature map to 96 channels. This feature map is then processed through six convolutional blocks, which further refine the integrated low- and high-level features.

The feature map is then unsampled (green arrow) to match the resolution of the first stage of the feature extraction path and concatenated with its output. This concatenated feature map is processed through a 1×1 convolution (second magenta trapezoid) and three additional convolutional blocks to complete processing at full resolution. A final convolution reduces the channel count to three, producing the stego image.

This symmetrical design facilitates the accurate reconstruction of the stego image. The overall architecture is designed to minimize pixel-level distortion in the cover image, ensuring that the resulting stego image closely resembles the original while securely embedding the secret message.

3. 3. Convolutional Blocks Each convolutional block in Figure 3 processes a feature map through a series of convolutional and normalization steps. Initially,

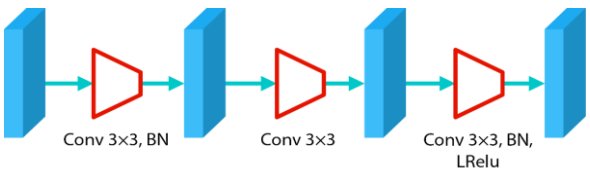


Figure 3. Convolutional block used in VidaFormer encoder and decoder

the input feature map passes through a convolution layer with a 3×3 kernel, followed by batch normalization to stabilize and accelerate convergence. This process is repeated with two additional 3×3 convolutional layers. Finally, batch normalization and a Leaky ReLU activation function are applied to introduce non-linearity while avoiding dead neurons by maintaining small gradients even for negative inputs.

Replacing standard convolutions with CoordConv (22) layers enhances the model's ability to retain spatial localization. CoordConv achieves this by adding explicit coordinate channels (x, y coordinates) to the input of the convolutional layers. This allows the network to retain spatial information and learn spatial dependencies more effectively than traditional convolutions. The added positional awareness enables the network to better understand and preserve the spatial structure of the input data. In steganography, where precise positioning is essential to embed data subtly without disrupting the image's visual integrity, CoordConv proves ideal for enhancing spatial control during the embedding process.

3. 4. Swin Transformer Stage The bottleneck stage in the encoder architecture consists of three Swin V2 (18) transformer blocks with a patch size of 2. Positioned at the network's bottleneck, it processes the feature map with the highest channel count and the lowest spatial resolution. This configuration allows the model to capture intricate, multi-scale dependencies across channels while preserving computational efficiency. It leverages Swin V2's local-global attention capabilities at the most condensed spatial representation.

Each block in this bottleneck stage, inspired by Swin V2, includes two multi-head self-attention modules. The first module applies attention within standard windows, while the second employs a shifted window configuration. This dual approach enhances performance by providing broader contextual awareness while keeping computational complexity linear rather than quadratic to the input size.

The query (Q), key (K), and value (V) matrices are generated through linear projections, transforming the input features as described in the below equations. These matrices are then used to compute multi-head self-attention, as detailed in Equation 6 (16).

$$Q = XW_Q \tag{3}$$

$$K = XW_K \tag{4}$$

$$V = XW_V \tag{5}$$

$$Attention = Softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{6}$$

In the above equations, Q, K, and V are query, key, and value tokens, B is the relative positional bias, and d is the query dimension. This approach is especially effective in the bottleneck, where spatial dimensions are reduced, as it leverages shifted window attention.

3. 5. Decoder Architecture Decoder architecture, illustrated in Figure 4, closely mirrors the encoder, with the primary difference being its input and output. Instead of embedding data, the decoder receives the stego image as input, processes it, and extracts the hidden information, ultimately recovering the binary secret message to align with the original.

3. 6. Network Loss The framework employs two primary loss functions: one for the encoder and one for the decoder as shown in Figure 5. For the encoder, the Mean Squared Error (MSE) loss function, as depicted in Equation 7, minimizes the difference between the cover image and the stego image, thereby ensuring the integrity and visual realism of the stego image. For the decoder, Equation 8 introduces a binary cross-entropy loss, which constrains the recovered message to match the original secret message with high accuracy.

$$L_E = \frac{1}{3\times W\times H}\sum(C - S)^2 \tag{7}$$

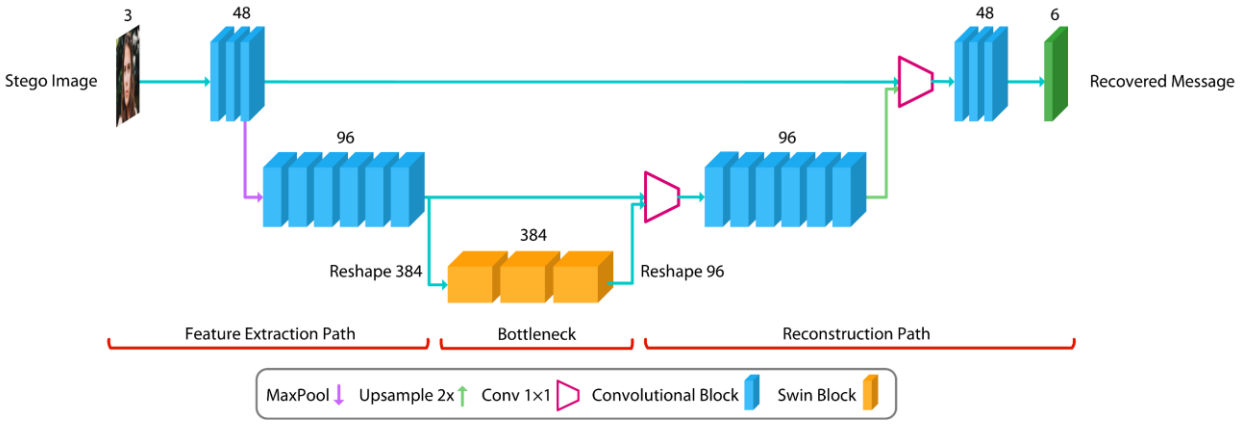


Figure 4. The architecture of the decoder which differs from the encoder in input and output

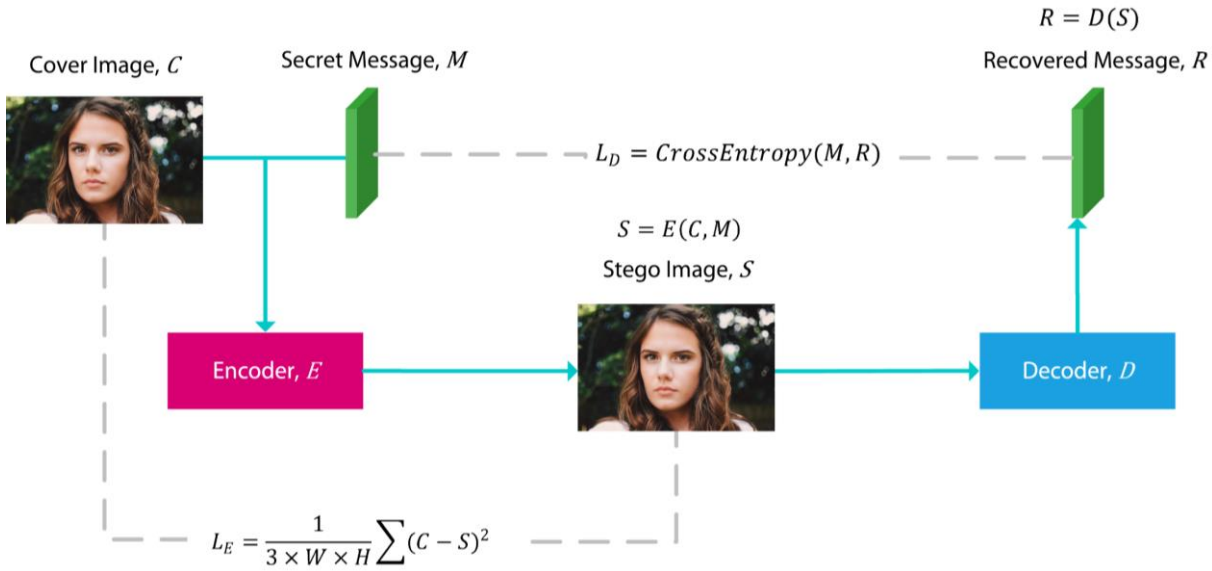


Figure 5. Network loss functions

$$L_D = CrossEntropy(M, R) \tag{8}$$

$$L_{total} = L_E + \theta L_D \tag{9}$$

The overall loss is computed using Equation 9, where θ is adjusted through MSE targeting, as proposed by Ramandi et al. (49). MSE targeting ensures that the stego image maintains sufficient integrity and realism based on the MSE loss while dynamically adjusting θ for each training iteration. This iterative adjustment balances the stego image's integrity and effective capacity.

3. 7. Error Correction Steganographic algorithms aim to extract the original message from the stego image, but the extracted message may contain errors, with some bits differing from the original secret message. However, the algorithm must fully and accurately extract the secret message without compromising its content. To achieve this, we incorporate Reed-Solomon error correction (23). This technique oversamples a polynomial constructed from the secret message, generating extended message that includes error-checking bits. It is particularly effective in steganography, where extracted data may contain errors.

The process is as follows: 1) The original message is processed by the Reed-Solomon algorithm, which adds correction bits to create an extended message. 2) The extended message is input to the encoder to generate the stego image. 3) The extended message, which may contain errors, is extracted from the stego image. 4) The Reed-Solomon algorithm corrects these errors, recovering the original message without any errors.

If the percentage of erroneous bits is below a threshold, the additional bits enable Reed-Solomon error correction to reconstruct the message accurately. Equation 10 specifies the minimum number of extra bits required for error correction, where p is the probability of error per bit, k is the length of the original bit array, and n is the extended length including error correction. Given p and k, we can calculate n to fully recover the message.

$$p.n \leq \frac{n-k}{2} \tag{10}$$

Maximizing k will give effective capacity which is the portion of the embedding capacity that corresponds to the original message, excluding additional error-correction data. We find k_{max} by turning Equation 10 to equality and rearrange as in Equation 11.

$$k_{max} = n(1 - 2p) \tag{11}$$

We denote recovery accuracy as a, defined as the probability of correctly extracting a bit, as opposed to erroneous extraction (p):

$$a = 1 - p \tag{12}$$

By definition, n is the embedding capacity. Combining the above equations will lead to Equation 13.

$$EC = Cap(2a - 1) \tag{13}$$

where EC is effective capacity and Cap is embedding capacity. Therefore, effective capacity can be easily calculated by embedding capacity and recovery accuracy.

4. EXPERIMENTS

The VidaFormer network is implemented using the PyTorch framework. Training and evaluation are

conducted on an NVIDIA RTX 3060 GPU, using images from the DIV2K dataset (59-61). To optimize performance, the embedding capacity is tested across configurations ranging from 1 to 6 bits, with the 6-bit configuration yielding the most effective results. Embedded capacity refers to the number of message bits embedded per pixel of the image (number above the green tensor in Figure 2). The list of hyperparameters is presented in Table 1. Data augmentation techniques, such as horizontal flipping and random cropping, are applied to enhance model generalization. The encoder and decoder are trained using the loss functions described in Network Loss.

4. 1. Evaluation Metrics We use three evaluation metrics to evaluate the quality of the stego image. The first metric is the Mean Squared Error (MSE), defined in Equation 14, which quantifies the difference between the stego image and the cover image.

$$MSE = \frac{1}{W \times H}(y - \hat{y})^2 \quad (14)$$

The second metric is the Peak Signal-to-Noise Ratio (PSNR), described in Equation 15, which evaluates image quality by quantifying pixel noise intensity.

$$PSNR = 10 \times \log_{10} \frac{s^2}{MSE} \quad (15)$$

where s is the maximum possible pixel value of the image.

The third metric is the Structural Similarity Index (SSIM) (57), as shown in Equation 16. SSIM evaluates the perceptual quality of the stego image by comparing its structural features, brightness, and contrast with those of the cover image.

$$SSIM = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)} \quad (16)$$

where $C_1 = 0.0001$ and $C_2 = 0.0009$, and the output is between -1 and 1.

These metrics enable a comprehensive quantitative comparison between the cover and stego images, ensuring the stego image's fidelity and quality.

4. 2. Main Results Table 2 presents the performance of our steganographic method across various embedding capacity, evaluating recovery accuracy, effective capacity, PSNR, and SSIM. As embedding capacity increases, recovery accuracy decreases, from perfect recovery at 1 to 0.907 at 6. This trend underscores the challenge of accurately recovering larger embedded data volumes. Nevertheless, effective capacity consistently improves, reaching 4.89 BPP at a capacity of 6, demonstrating the method's ability to handle higher data volumes.

PSNR values remain stable at lower capacities, peaking at 40.41 for a capacity of 2 but declining to 37.59 at a capacity of 6. SSIM values consistently surpass 0.90 at all capacities, indicating that the structural features of the images are effectively preserved. The highest SSIM of 0.96 is achieved at a capacity of 2. At an embedding capacity of 6, the model performs optimally when prioritizing capacity, achieving effective capacity of 4.89 BPP, 90.7% recovery accuracy, and acceptable image quality (PSNR of 37.59 and SSIM of 0.91). This demonstrates the method's strength in balancing high effective capacity with maintaining good overall performance.

Table 3 compares the performance of the proposed method with other advanced steganography techniques, including VidaGAN, SteganoGAN, AFHS-GAN, and FNNS variants, using recovery accuracy, PSNR, and SSIM as evaluation metrics. The proposed method achieves the highest recovery accuracy of 0.987,

TABLE 1. Hyperparameters used in the training of VidaFormer

Hyperparameter	Value
Epochs	80
Embedding Capacity	1-6
Resolution	256×256
Learning Rate	10^{-4}
Weight Decay	5×10^{-6}

TABLE 2. VidaFormer output for different capacities

Embedding Capacity (BPP)	Recovery Accuracy	Effective Capacity (BPP)	PSNR	SSIM
1	**1.00**	1.00	39.57	0.95
2	0.999	1.99	**40.41**	**0.96**
3	0.998	2.99	40.00	0.92
4	0.987	3.89	39.31	0.92
5	0.918	4.18	39.71	0.92
6	0.907	**4.89**	37.59	0.91

TABLE 3. Comparison of VidaFormer with other state-of-the-art methods applied to the DIV2K dataset (4-BPP)

Method	Recovery Accuracy	PSNR	SSIM
SteganoGAN (50)	0.820	37.49	0.88
FNNS-R (62)	0.891	28.60	0.76
FNNS-D (62)	0.945	25.74	0.65
Secure FNNS (5)	0.912	25.79	0.77
VidaGAN (49)	0.970	37.51	0.90
AFHS-GAN (63)	0.970	**40.14**	**0.94**
Ours	**0.987**	39.31	0.92

surpassing AFHS-GAN and VidaGAN, which both rank second at 0.970. This highlights the method's robust capability to reliably extract hidden data, even outperforming Secure FNNS. For PSNR, a measure of image quality, AFHS-GAN scores 40.14, the highest among all techniques, with the proposed method and VidaGAN following at 39.31 and 37.51, respectively, while FNNS-based methods exhibit significantly lower values. In terms of SSIM, which assesses the preservation of structural details in the image, AFHS-GAN achieves the highest score of 0.94, surpassing the proposed method (0.92) and VidaGAN (0.90). FNNS-based methods perform poorly in this regard, with FNNS-D recording the lowest SSIM of 0.65. Overall, the proposed method competes toe-to-toe with AFHS-GAN, surpassing in effective capacity but falling behind slightly in terms of transparency.

4. 3. Statistical Variance To provide a comprehensive evaluation of VidaFormer's performance and ensure the robustness of the reported metrics, we analyze the statistical variance of key evaluation metrics: recovery accuracy, effective capacity (BPP), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) across multiple runs on the DIV2K dataset. This analysis quantifies the consistency of the model's performance under varying training conditions and highlights its stability for image steganography tasks.

The VidaFormer model was trained and evaluated over 5 independent runs on the DIV2K dataset, using the same experimental setup described in experiment section. For each run, we computed the metrics at embedding capacities of 6 bits-per-pixel (BPP), which maximizes effective capacity. We report the mean and standard deviation of recovery accuracy, effective capacity, PSNR, and SSIM to capture the variability across runs. Table 4 presents the mean and standard deviation of the metrics for the 6-BPP configuration, which is the most challenging due to its high embedding capacity.

The mean recovery accuracy of 0.91 indicates reliable message extraction, with a low standard deviation, suggesting consistent performance across runs. The tight reflects the model's ability to maintain high accuracy despite the high embedding load. The effective capacity averages 4.89 BPP, with a standard deviation of 0.1150 BPP. This minor variability underscores the model's stability in embedding large data volumes while preserving message integrity. Variations are primarily due to differences in initialization and data augmentation effects. The mean PSNR of 37.59 dB, with a standard deviation of 0.4219 dB, indicates stable image quality across runs. The relatively low variance suggests that the model consistently produces stego images with minimal distortion, even under different training conditions. The SSIM mean of 0.91, with a standard deviation of 0.005094, confirms consistent preservation of structural features in stego images. The low variance indicates that VidaFormer reliably maintains perceptual quality, with slight fluctuations attributable to the stochastic nature of training.

TABLE 4. Statistical Variance of Metrics for VidaFormer (6-BPP)

Metric	Mean	Standard Deviation
Recovery Accuracy	0.9071	0.009601
Effective Capacity (BPP)	4.887	0.1150
PSNR (dB)	37.59	0.4219
SSIM	0.9073	0.005094

4. 4. Ablation Studies The ablation studies, presented in Table 5, examine how modifications to the baseline network affect steganographic performance.

Performance is evaluated using recovery accuracy, effective capacity (BPP), PSNR, and SSIM. The baseline model achieves a recovery accuracy of 0.822, an effective capacity of 3.86 BPP, a PSNR of 38.29, and an SSIM of 0.92, serving as a reference for further improvements.

Incorporating CoordConv, which enhances spatial processing, significantly improves performance. Recovery accuracy rises to 0.859, effective capacity increases to 4.31 BPP, PSNR improves to 39.19, and SSIM reaches 0.93, indicating enhanced data handling and image quality.

Using transformer layers in the bottleneck further improves recovery accuracy to 0.907 and effective capacity to 4.89 BPP. However, this entails a slight trade-off, with PSNR decreasing to 37.59 and SSIM to 0.89, reflecting a balance between capacity and image quality.

Compared to AFHS-GAN, which at 6 BPP capacity, achieves an effective capacity of 4.86 BPP, a PSNR of 37.69, and an SSIM of 0.89, the two models offer similar results. This demonstrate that architectural improvements, particularly CoordConv and transformer layers, significantly enhance network performance while maintaining image quality.

TABLE 5. Ablation studies with each main part of the architecture in comparison with VidaGAN (6-BPP)

Network	Recovery Accuracy	Effective Capacity (BPP)	PSNR	SSIM
VidaGAN (49)	0.825	3.9	38.56	0.88
Baseline (conv network)	0.822	3.86	38.29	0.92
+ CoordConv	0.859	4.31	**39.19**	**0.93**
+ transformer	**0.907**	**4.89**	37.59	0.89

4. 5. MSE Adjustment MSE targeting, introduced by VidaGAN (49), balances the fidelity of the cover and stego images by setting a target Mean Squared Error (MSE) value. Figure 6 illustrates how adjusting this MSE value affects recovery accuracy, PSNR, and SSIM when embedding 6 bits-per-pixel (6 BPP).

Increasing the MSE threshold improves recovery accuracy, rising from 0.835 at the lowest threshold (0.00025) to 0.963 at the highest (0.000875). This indicates that higher MSE values enhance the method's ability to reliably recover data. However, this improvement comes at the cost of image quality, as PSNR decreases from 40.038 to 33.142, reflecting increased distortion in the stego images.

SSIM, which measures structural similarity, decreases as the MSE threshold increases, falling from 0.945 to 0.789. Higher thresholds improve data recovery but compromise both structural and visual image quality. Figure 7 illustrates that enhanced recovery accuracy at higher thresholds increases embedding capacity but requires a trade-off in image fidelity.

These findings highlight the flexibility of MSE targeting. Lower thresholds prioritize image quality, yielding high PSNR and SSIM, while higher thresholds emphasize improved recovery accuracy at the cost of visual clarity. This adaptability enables the method to meet diverse application requirements.

4. 6. Computational Cost The model's computational demands are primarily driven by the transformer blocks, which employ self-attention mechanisms that scale quadratically with sequence length, although mitigated here by the shifted window approach in Swin V2 and their placement only in the bottleneck (where spatial dimensions are reduced to W/4× H/4). During training, VidaFormer requires approximately 80 epochs on an NVIDIA RTX 3060 GPU, with each epoch processing the DIV2K dataset taking less than 5 minutes. This results in a total training time of about 6.5 hours for the full model, including data augmentation steps like random cropping and flipping.

Inference times, as detailed in Table 6, are more practical for individual images: embedding completes in 0.2 seconds, benefiting from the efficient convolutional paths, while extraction takes 1.45 seconds, largely due to the Reed-Solomon error correction post-processing. When benchmarked against state-of-the-art methods (Table 7), VidaFormer's times are moderate. It is faster than FNNS (which exceeds 44 seconds) but slower than purely convolutional models like LISO (0.33 seconds) (64). The GPU memory footprint peaks at around 4-6 GB during inference, making it feasible on mid-range hardware but potentially challenging on edge devices without optimization.

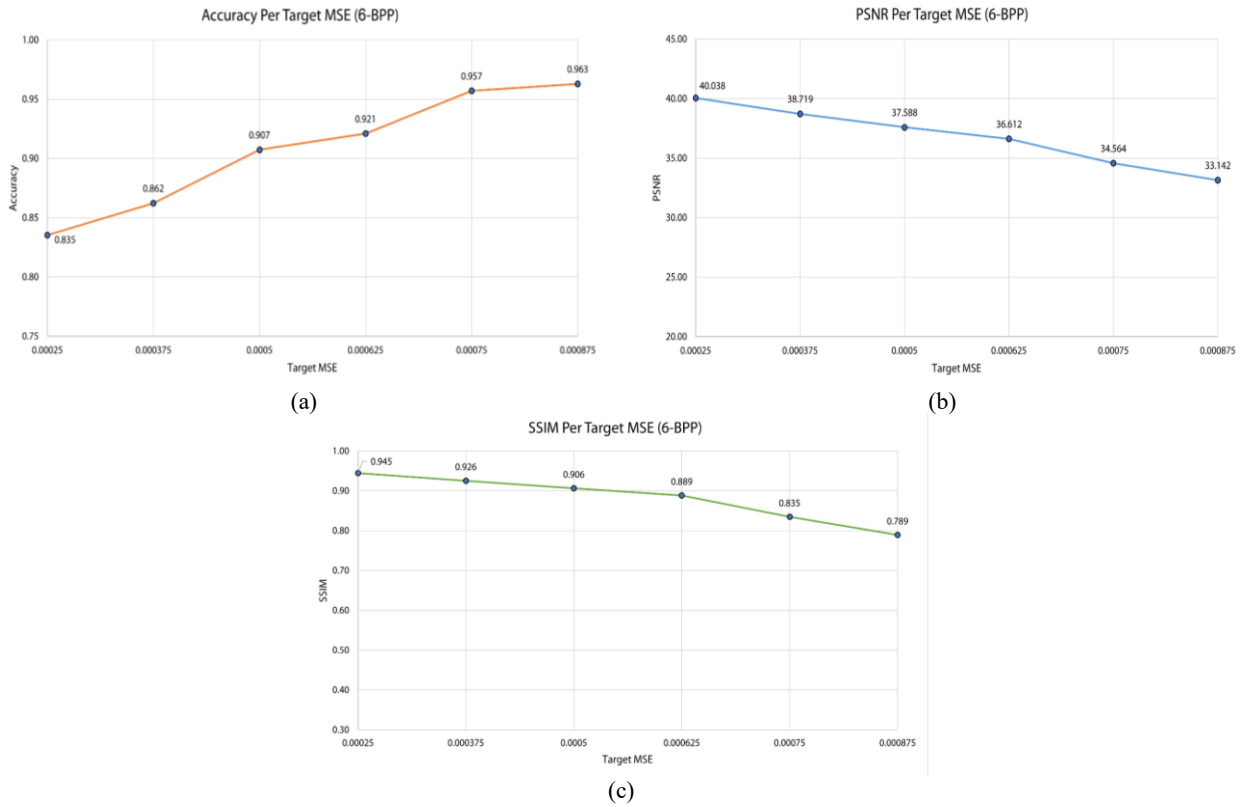
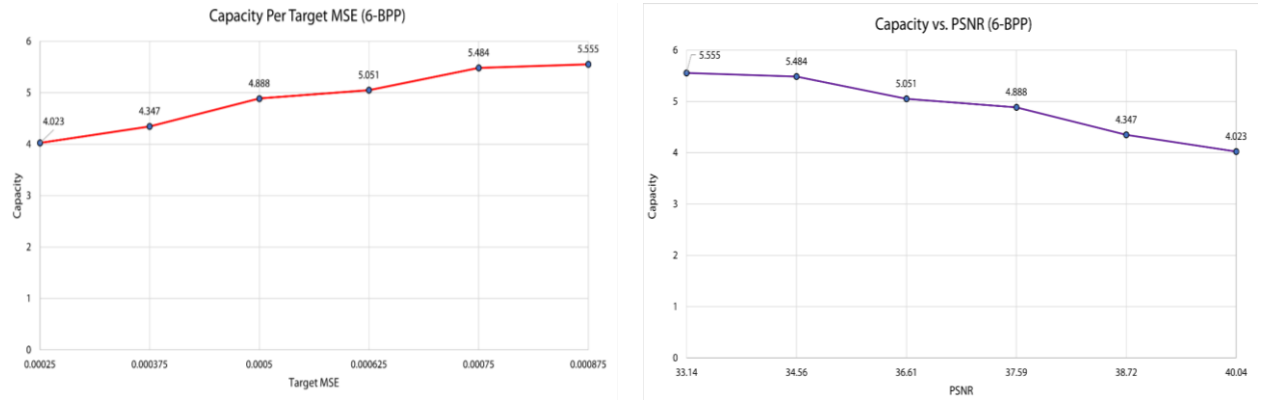


Figure 6. This Figure illustrates the impact of tweaking MSE targeting on three important metrics: Accuracy (a), PSNR (b), and SSIM (c)



(a) (b)

Figure 7. Capacity vs (a) MSE, (b) PSNR

TABLE 6. Computation time for images (4-BPP)

Task	Time Taken (S)
Embedding	0.20
Extraction	1.45
Embedding + Extraction	1.65

TABLE 7. Computation time for different methods (4-BPP)

Method	Time Taken (S)
SteganoGAN (50)	0.9
FNNS-D (62)	44.29
LISO (64)	0.33
VidaGAN (49)	1.28
Ours	1.65

VidaFormer demonstrates good scalability to higher resolutions due to its hybrid design: convolutional layers handle early high-resolution stages with linear complexity, while transformers are confined to down sampled features, avoiding the full quadratic scaling issues of pure ViT models.

Overall, while VidaFormer is computationally more intensive than traditional CNN-based steganography, its design choices enhance scalability compared to full transformer models. Future enhancements, as outlined in future work, such as adopting lighter transformers could further improve its viability for real-time and large-scale deployments, making it a scalable solution for secure data hiding in diverse scenarios.

4. 7. Steganalysis Steganography focuses on concealing messages to evade detection, while steganalysis seeks to uncover hidden content in digital media. GBRAS-Net (65), a recent CNN-based steganalysis tool, performs spatial image steganalysis by incorporating filter banks, depth-wise and separable convolutional layers, and skip connections. It outperforms recent works in detecting steganographic images across multiple adaptive algorithms using BOSSbase 1.01 and BOWS 2 datasets. The approach significantly improves classification accuracy for various steganographic methods and should be able to challenge steganographic algorithms.

In this study, GBRAS-Net was evaluated on 100 masked and hidden images to assess its effectiveness in detecting embedded messages. The results yielded an area under the Receiver Operating Characteristic curve (AUROC) of 0.510 for the secret message detection task, a s shown in Figure 8. This suggests that GBRAS- Net was unable to distinguish between cover and stego images, with predictions only slightly better than random guessing. The classification report and confusion matrix for GBRAS-Net are presented in Tables 8 and 9, respectively, indicating that the steganalysis tool is heavily biased toward classifying images as stego.

To visually assess the effectiveness and embedding characteristics of various steganographic methods, we compare the residuals, the difference between the stego

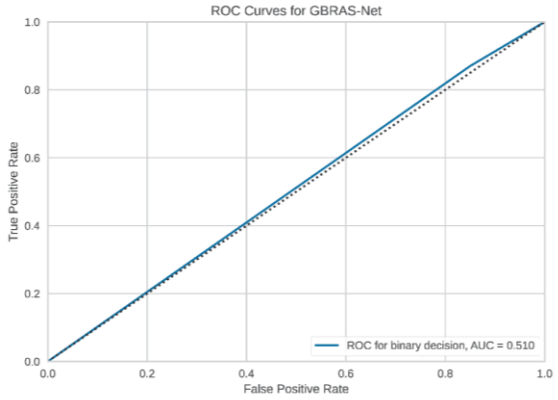


Figure 8. ROC curve for classification between cover and stego images by GBRAS-Net

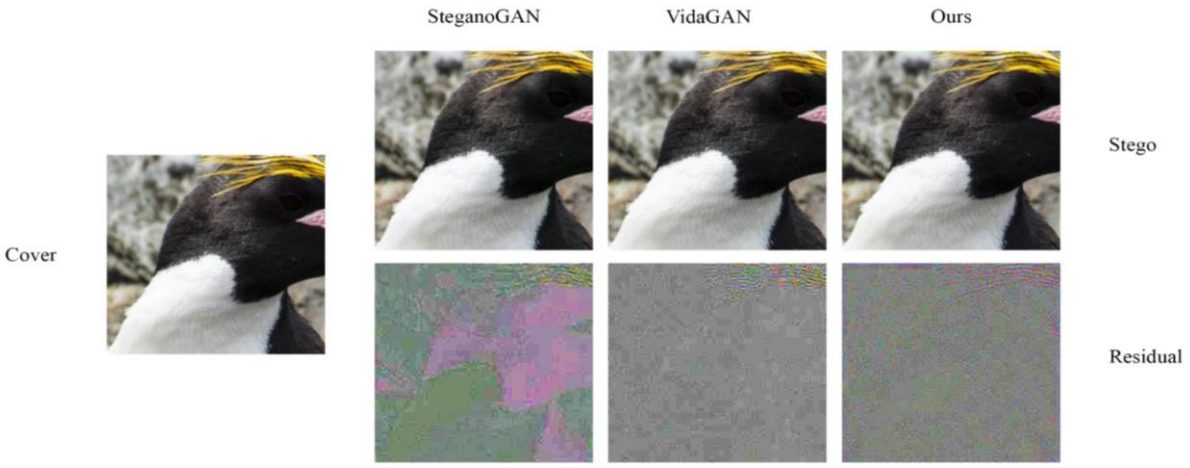


Figure 9. Qualitative results for 3 BPP stegos by different methods. For each method, the residual image is the difference between stego and cover image multiplied by 10.

and cover images, for SteganoGAN (50), VidaGAN (49), and our proposed method. As shown in Figure 9, the residuals are magnified by a factor of 10 to enhance perceptibility and highlight subtle embedding patterns that may not be visible in the original stego images.

In the SteganoGAN residual, we observe large spatial regions where the patterns closely resemble textures from the cover image. This suggests that the data embedding is not uniformly distributed and is influenced by the content of the cover image. Specifically, areas with higher color variation or certain textures seem to carry more embedded information, implying a channel-dependent and image-aware encoding strategy.

In the SteganoGAN residual, we observe large spatial regions where the patterns closely resemble textures from the cover image. This suggests that the data embedding is not uniformly distributed and is influenced by the content of the cover image. Specifically, areas with higher color variation or certain textures seem to carry more embedded information, implying a channel-dependent and image-aware encoding strategy.

The VidaGAN residual, on the other hand, appears predominantly grayscale with less structural variation. This indicates that the embedding is more uniformly spread across the RGB channels, with less dependence on the original content of the image. The uniformity hints at a more balanced, channel-agnostic embedding process.

In contrast, the residual generated by our method displays higher frequency patterns and more diverse coloration, suggesting a finer-grained and denser embedding. These characteristics are consistent with the observed improvement in decoding accuracy. The presence of high-frequency residuals implies that our method is capable of encoding more bits without significantly altering the perceptual quality of the stego image.

TABLE 8. Classification Report for GBRAS-Net

Class	Precision	Recall	F1-Score
Cover	0.536	0.150	0.234
Stego	0.506	0.870	0.640

TABLE 9. Confusion matrix for GBRAS-Net

True Class	Predicted Class	
	Cover	Stego
Cover	15%	85%
Stego	13%	87%

Overall, the qualitative comparison highlights that our method achieves more efficient and content-independent embedding, which translates into higher capacity compared to SteganoGAN and VidaGAN.

4. 8. Lightweight Variant To address the computational complexity of the Swin V2 transformer blocks used in the bottleneck stage of VidaFormer, we explore a lightweight transformer variant, MobileViT (61), as an alternative to demonstrate the trade-off between efficiency and steganographic performance. MobileViT combines convolutional layers with transformer-like blocks designed for resource-constrained environments, offering reduced computational overhead while maintaining competitive feature extraction capabilities. This makes it a suitable candidate for optimizing VidaFormer for scenarios requiring lower latency or deployment on edge devices.

We replaced the three Swin V2 transformer blocks in the bottleneck stage of the VidaFormer encoder (as described in Section 3.4) with MobileViT blocks,

specifically the MobileViT-XS configuration, which balances efficiency and performance. The MobileViT blocks process the feature map at the same resolution (W/4 × H/4, 384 channels) as the original Swin V2 blocks, ensuring compatibility with the existing U-Net-inspired architecture. The MobileViT blocks utilize a combination of depthwise separable convolutions and transformer-like self-attention mechanisms, reducing the computational complexity from the quadratic scaling of Swin V2's shifted window attention to a more efficient linear scaling. The rest of the architecture, including the CoordConv layers in the convolutional blocks and the Reed-Solomon error correction mechanism, remains unchanged.

As presented in Table 10, the MobileViT-based VidaFormer achieves a recovery accuracy of 0.836, lower than the Swin V2-based VidaFormer (0.907), resulting in an effective capacity of 4.03 BPP compared to 4.89 BPP. This indicates a trade-off in steganographic performance, as the lightweight MobileViT blocks are less effective at capturing long-range dependencies compared to Swin V2's shifted window attention. However, the PSNR (37.82 dB) is slightly improved.

In terms of efficiency, the MobileViT-based model significantly reduces computational demands. The training time per epoch drops to 3.2 minutes from 4.8 minutes and the total training time is approximately 4.5 hours compared to 6.5 hours for the Swin V2-based model. Inference times are also improved, with embedding completed in 0.15 seconds and extraction in 1.39 seconds. The GPU memory footprint is reduced to 3.9 GB from 5.8 GB, making the MobileViT-based VidaFormer more suitable for deployment on resource-constrained devices. The integration of MobileViT demonstrates a clear trade-off between efficiency and steganographic performance. The reduction in recovery accuracy and effective capacity (from 4.89 BPP to 4.03 BPP) reflects the limitations of MobileViT's lightweight self-attention mechanism in modeling complex global dependencies compared to Swin V2.

TABLE 10. Performance Comparison of VidaFormer with lightweight variant (6-BPP)

Metric	VidaFormer	VidaFormer (MobileViT-XS)
Recovery Accuracy	0.907	0.836
Effective Capacity (BPP)	4.89	4.03
PSNR (dB)	37.59	37.82
SSIM	0.91	0.91
Epoch Training Time (min)	4.8	3.2
Embedding Time (s)	0.2	0.15
Extraction time (s)	1.45	1.39
GPU Memory (GB)	5.8	3.9

5. CONCLUSION

This study introduces VidaFormer, a new type of hybrid transformer model that sets a high standard for image steganography by balancing capacity, recovery accuracy, and image quality. It combines convolutional layers with CoordConv for better spatial understanding and Swin V2 transformer blocks to capture long-distance connections in images. On the DIV2K dataset, VidaFormer achieves an effective capacity of 4.89 bits-per-pixel (BPP), which is 1.7% better than the previous best model. It uses Reed-Solomon error correction to ensure reliable message recovery and a U-Net-inspired design to keep images clear, with a PSNR of 37.59 and SSIM of 0.91 at 6 BPP. Testing with GBRAS-Net shows it hides messages well, with an AUROC of 0.510, meaning detection is almost like random guessing.

VidaFormer's design tackles the challenges of steganography, such as balancing capacity, transparency, and processing speed, through a well-planned structure. CoordConv helps accurately place hidden data, and transformer layers improve feature detection, overcoming issues with traditional convolutional networks. Tests confirm these parts work well, and MSE targeting allows the model to adapt for different needs, like focusing on image quality or data recovery.

While VidaFormer is a big step forward, it has some issues, like high computational demands and limited dataset variety. Future work will aim to improve efficiency, adaptability, and explore new uses. VidaFormer is a strong foundation for secure, high-capacity data hiding and shows how hybrid deep learning can improve steganography, offering a reliable solution for secure communication in a digital world.

5. 1. Limitations Despite VidaFormer's state-of-the-art performance in image steganography, particularly in achieving high effective capacity and recovery accuracy, several limitations must be acknowledged. First, the integration of transformer blocks in the bottleneck stage, while optimized through a hybrid architecture, introduces computational complexity compared to purely convolutional models. The extraction process, for example, takes approximately 1.45 seconds per image on an NVIDIA RTX 3060 GPU, primarily due to the Reed-Solomon error correction mechanism. This latency may limit its suitability for real-time applications, especially when processing high-resolution images or large datasets.

Second, the evaluation is conducted primarily on the DIV2K dataset, which, while high-quality, is relatively small, with only 100 validation images commonly used in benchmarks. This raises concerns about the model's generalizability to more diverse image sets, such as low-resolution web images or those from real-world scenarios like social media platforms. Furthermore,

while transparency is evaluated using GBRAS-Net steganalysis, achieving an AUROC of 0.510, the tool's bias toward classifying images as stego suggests that additional testing against advanced steganalysis methods (e.g., ensemble detectors or deeper CNN-based analyzers) is needed to fully confirm the model's undetectability.

Finally, the fixed embedding capacity of up to 6 bits-per-pixel (BPP), with an effective capacity of 4.89 BPP, highlights an inherent trade-off between capacity, transparency, and robustness. Prioritizing one of these objectives often compromises the others, and further optimization is needed to achieve a better balance.

5. 2. Future Works To address the identified limitations and further enhance VidaFormer's capabilities, several research directions are proposed. First, improving computational efficiency is a key priority. Exploring more lightweight transformer architectures, such as EfficientFormer, could reduce latency in both embedding and extraction phases, enabling real-time applications on resource-constrained devices like mobile phones or for processing video streams.

Second, expanding the evaluation to include larger and more diverse datasets, such as COCO or ImageNet subsets, will help validate the model's generalizability and reduce the risk of overfitting to the DIV2K dataset.

Third, investigating alternative positional encoding mechanisms beyond CoordConv, such as learnable embeddings or hybrid CNN-transformer integrations at multiple stages of the architecture, could further improve spatial awareness without increasing input channel complexity. Additionally, developing dynamic embedding strategies that allocate capacity based on image content such as embedding more bits in textured regions versus smooth areas could optimize the trade-off between capacity and transparency.

Finally, extending VidaFormer to other domains, such as video steganography, audio data hiding, or multi-modal concealment, presents exciting opportunities. Integrating the model with generative approaches for coverless steganography could enhance security in applications like secure communication over online social networks. These advancements would position VidaFormer as a versatile framework for next-generation steganographic systems.

Acknowledgements

The author appreciates Kharazmi Campus and Faculty of Computer Engineering in Shahrood University of Technology for providing necessary facilities to conduct present research.

Funding

This work has not received any financial support.

Declarations Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable. Competing interests the author declares no financial or organizational conflicts of interest.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the author used ChatGPT exclusively for minor language editing and stylistic refinement to improve clarity and readability. The author carefully reviewed, revised, and approved the final content and takes full responsibility for the accuracy, integrity, and originality of the work.
The author declares that there are no known financial or organizational conflicts of interest that could have influenced the work reported in this paper.

Authors Biosketch

Vida Yousefi Ramandi holds a Ph.D. in Artificial Intelligence from Shahrood University of Technology. Her research focuses on image steganography using deep learning techniques, and she has published several papers in this field. She completed her Ph.D. program with an A+ academic standing and defended her dissertation with distinction.

Mansoor Fateh received the M.S. degree in biomedical engineering and the Ph.D. degree from Tarbiat Modares University, Tehran, Iran. He is currently a Faculty Member with the Faculty of Computer Engineering, Shahrood University of Technology, Iran. His research interests include deep learning and image processing.

Mohsen Rezvani received the PhD degree in computer science from the University of New South Wales, Australia. He is a faculty member in the Faculty of Computer Engineering, Shahrood University of Technology, Iran. His research focuses on computer security, privacy, trust and reputation systems.

REFERENCES

1. You Z, Ying Q, Li S, Qian Z, Zhang X, editors. Image generation network for covert transmission in online social network. Proceedings of the 30th ACM International Conference on Multimedia; 2022. DOI: 10.1145/3503161.3548139.

2. Evsutin O, Melman A, Meshcheryakov R. Digital Steganography and Watermarking for Digital Images: A Review of Current Research Directions. IEEE Access. 2020;8:166589-611. DOI: 10.1109/ACCESS.2020.3022779.
3. Saadati M, Vahidi J, Seydi V, Sheikholharam Mashhadi P. Proposing a New Image Watermarking Method Using Shearlet Transform and Whale Optimization Algorithm. International Journal of Engineering Transactions A: Basics. 2021;34(4):843-53. DOI: 10.5829/ije.2021.34.04a.10.
4. Qi Y, Chen K, Zhao N, Yang Z, Zhang W. Provably Secure Robust Image Steganography via Cross-Modal Error Correction. arXiv e-prints. 2024:arXiv: 2412.12206. DOI: 10.48550/arXiv.2412.12206.
5. Luo Z, Li S, Li G, Qian Z, Zhang X, editors. Securing Fixed Neural Network Steganography. Proceedings of the 31st ACM International Conference on Multimedia; 2023. DOI: 10.1145/3581783.3611920.
6. Fateh M, Mohsen R, Yasser I. A new method of coding for steganography based on LSB matching revisited. Security and Communication Networks. 2021:1-15. DOI: 10.1155/2021/6610678.
7. Mielikainen J. LSB matching revisited. IEEE signal processing letters. 2006;13(5):285-7. DOI: 10.1109/LSP.2006.870357.
8. Duan X, Liu N, Gou M, Wang W, Qin C. SteganoCNN: Image steganography with generalization ability based on convolutional neural network. Entropy. 2020;22(10):1140. DOI: 10.3390/e22101140.
9. Kumar V, Laddha S, Aniket ND. Steganography techniques using convolutional neural networks. J Homepage. 2020;7:66-73. DOI: 10.18280/rces.070304.
10. Wu P, Yang Y, Li X, editors. Image-into-image steganography using deep convolutional network. Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part II 19; 2018: Springer. DOI: 10.1007/978-3-030-00767-6_73.
11. Ronneberger O, Fischer P, Brox T, editors. U-Net: Convolutional Networks for Biomedical Image Segmentation. International Conference on Medical image computing and computer-assisted intervention; 2015. DOI: 10.1007/978-3-319-24574-4_28.
12. Wang C-Y, Liao H-YM, Wu Y-H, Chen P-Y, Hsieh J-W, Yeh I-H, editors. CSPNet: A new backbone that can enhance learning capability of CNN. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops; 2020. DOI: 10.1109/CVPRW50498.2020.00203.
13. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ, editors. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. DOI: 10.1016/j.patcog.2020.107610
14. Luo W, Li Y, Urtasun R, Zemel R. Understanding the effective receptive field in deep convolutional neural networks. Advances in neural information processing systems. 2016;29. DOI: 10.5555/3157382.3157645.
15. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al., editors. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations; 2020. DOI: 10.48550/arXiv.2010.11929.
16. Vaswani A. Attention is all you need. Advances in Neural Information Processing Systems. 2017. DOI: 10.5555/3295222.3295349.
17. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al., editors. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision; 2021. DOI: 10.1109/ICCV48922.2021.00986.
18. Liu Z, Hu H, Lin Y, Yao Z, Xie Z, Wei Y, et al., editors. Swin transformer v2: Scaling up capacity and resolution. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. DOI: 10.1109/CVPR52688.2022.01170.
19. Dong X, Bao J, Chen D, Zhang W, Yu N, Yuan L, et al., editors. Cswin transformer: A general vision transformer backbone with cross-shaped windows. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. DOI: 10.1109/CVPR52688.2022.01181.
20. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. EEE transactions on pattern analysis and machine intelligence. 2022;45(1):87-110. DOI: 10.1109/TPAMI.2022.3152247.
21. Kayhan OS, Gemert JCv, editors. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. DOI: 10.1109/CVPR42600.2020.01428.
22. Liu R, Lehman J, Molino P, Petroski Such F, Frank E, Sergeev A, et al. An intriguing failing of convolutional neural networks and the coordconv solution. Advances in neural information processing systems. 2018;31. DOI: 10.5555/3327546.3327630.
23. Reed IS, Solomon G. Polynomial codes over certain finite fields. Journal of the society for industrial and applied mathematics. 1960;8(2):300-4. DOI: 10.1137/0108018.
24. Fateh M, Rezvani M, Irani Y. A New Method of Coding for Steganography Based on LSB Matching Revisited. Security and Communication Networks. 2021;2021(1):6610678. 10.1155/2021/6610678.
25. Karim SM, Rahman MS, Hossain MI, editors. A new approach for LSB based image steganography using secret key. 14th international conference on computer and information technology (ICCIT 2011); 2011: IEEE. DOI: 10.1109/ICCITechn.2011.6164800.
26. mansoor fateh sr, elahe alipour. Review of Image Steganography Based on LSB and Pixel Classification and Providing a New Method in This Area. Biannual Journal Monadi for Cyberspace Security (AFTA). 2017;5(2):9. http://monadi.isc.org.ir/article-1-82-en.html.
27. Singh A, Singh H, editors. An improved LSB based image steganography technique for RGB images. 2015 IEEE International Conference on electrical, computer and communication technologies (ICECCT); 2015: IEEE. DOI: 10.1109/ICECCT.2015.7226122.
28. Das R, Tuithung T, editors. A novel steganography method for image based on Huffman Encoding. 2012 3rd National Conference on Emerging Trends and Applications in Computer Science; 2012: IEEE. DOI: 10.1109/NCETACS.2012.6203290.
29. Arham A, Nugroho HA. Enhanced reversible data hiding using difference expansion and modulus function with selective bit blocks in images. Cybersecurity. 2024;7(1):61. DOI: 10.1186/s42400-024-00251-7.
30. Chhajed G, Garg B. Novel Scheme for Data Hiding in Binary Images using Cover Pattern Histogram. International Journal of Engineering Transactions B: Applications. 2023;36(11):2124-36. DOI: 10.5829/ije.2023.36.11b.16.
31. Alexan W, Mamdouh E, Aboshousha A, Alsahafi YS, Gabr M, Hosny KM. Stegocrypt: A robust tri-stage spatial steganography algorithm using TLM encryption and DNA coding for securing digital images. IET Image Processing. 2024;18(13):4189-206. DOI: 10.1049/ipr2.13242.
32. Qu Z, Cheng Z, Liu W, Wang X. A novel quantum image steganography algorithm based on exploiting modification direction. Multimedia Tools and Applications. 2019;78:7981-8001. DOI: 10.1007/s11042-018-6476-5.

33. Wang S, Sang J, Song X, Niu X. Least significant qubit (LSQb) information hiding algorithm for quantum image. Measurement. 2015;73:352-9. DOI: 10.1016/j.measurement.2015.05.038.
34. Swain G. Very high capacity image steganography technique using quotient value differencing and LSB substitution. Arabian Journal for Science and Engineering. 2019;44(4):2995-3004. DOI: 10.1007/s13369-018-3372-2.
35. Qiu A, Chen X, Sun X, Wang S, Guo W. Coverless image steganography method based on feature selection. Journal of Information Hiding and Privacy Protection. 2019;1(2):49. DOI: 10.32604/jihpp.2019.05881.
36. Butora J, Puteaux P, Bas P. Errorless robust JPEG steganography using outputs of JPEG coders. IEEE Transactions on Dependable and Secure Computing. 2023. DOI: 10.1109/TDSC.2023.3306379.
37. Bhatia SS, Singh K, Kasana G. Singular Value Decomposition based Steganography Technique for JPEG2000 Compressed Images. International Journal of Engineering Transactions C: Aspects. 2015;28(12):1720-7. DOI: 10.5829/idosi.ije.2015.28.12c.04.
38. Wu P, Yang Y, Li X. Stegnet: Mega image steganography capacity with deep convolutional network. Future Internet. 2018;10(6):54. DOI: 10.3390/fi10060054.
39. Duan X, Jia K, Li B, Guo D, Zhang E, Qin C. Reversible image steganography scheme based on a U-Net structure. IEEE Access. 2019;7:9314-23. DOI: 10.1109/ACCESS.2019.2891247.
40. Rahim R, Nadeem S, editors. End-to-end trained CNN encoder-decoder networks for image steganography. Proceedings of the European conference on computer vision (ECCV) workshops; 2018. DOI: 10.1007/978-3-030-11018-5_64.
41. Wang Z, Gao N, Wang X, Xiang J, Liu G, editors. STNet: A style transformation network for deep image steganography. Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part II 26; 2019: Springer. DOI: 10.1007/978-3-030-36711-4_1.
42. Simonyan K, Zisserman A, editors. Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations (ICLR 2015); 2015: Computational and Biological Learning Society.
43. Baluja S. Hiding images in plain sight: Deep steganography. Advances in neural information processing systems. 2017;30. DOI: 10.5555/3294771.3294968.
44. Zhou Z, Su Y, Li J, Yu K, Wu QJ, Fu Z, et al. Secret-to-image reversible transformation for generative steganography. IEEE Transactions on Dependable and Secure Computing. 2022;20(5):4118-34. DOI: 10.1109/TDSC.2022.3217661.
45. Zhang J, Chen K, Li W, Zhang W, Yu N. Steganography with Generated Images: Leveraging Volatility to Enhance Security. IEEE Transactions on Dependable and Secure Computing. 2023. DOI: 10.1109/TDSC.2023.3341427.
46. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Advances in neural information processing systems. 2014;27. DOI: 10.5555/2969033.2969125.
47. Volkhonskiy D, Borisenko B, Burnaev E. Generative adversarial networks for image steganography. 2016.
48. Volkhonskiy D, Nazarov I, Burnaev E, editors. Steganographic generative adversarial networks. Twelfth international conference on machine vision (ICMV 2019); 2020: SPIE. DOI: 10.1117/12.2559429.
49. Ramandi VY, Fateh M, Rezvani M. VidaGAN: Adaptive GAN for image steganography. IET Image Processing. 2024;18(12):3329-42. DOI: 10.1049/ipr2.13177.
50. Zhang KA, Cuesta-Infante A, Xu L, Veeramachaneni K. SteganoGAN: High Capacity Image Steganography with GANs. ArXiv. 2019;abs/1901.03892. DOI: 10.48550/arXiv.1901.03892.
51. Bi X, Yang X, Wang C, Liu J. High-Capacity Image Steganography Algorithm Based on Image Style Transfer. Security and Communication Networks. 2021;2021(1):4179340. DOI: 10.1155/2021/4179340.
52. Su W, Ni J, Sun Y. StegaStyleGAN: towards generic and practical generative image steganography. Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence: AAAI Press; 2024. p. Article 28. DOI: 10.1609/aaai.v38i1.27776
53. Tang W, Rao Y, Yang Z, Peng F, Cui X, Huang J, et al. Reversible generative steganography with distribution-preserving. Cybersecurity. 2025;8(1):18. DOI: 10.1186/s42400-024-00317-6.
54. Wang Z, Zhou M, Liu B, Li T. Deep Image Steganography Using Transformer and Recursive Permutation. Entropy. 2022;24(7):878. DOI: 10.3390/e24070878.
55. Zhou Y, Luo T, He Z, Jiang G, Xu H, Chang C-C. CAISFormer: Channel-wise attention transformer for image steganography. Neurocomputing. 2024;603:128295. DOI: 10.1016/j.neucom.2024.128295.
56. Xiao C, Peng S, Zhang L, Wang J, Ding D, Zhang J. A transformer-based adversarial network framework for steganography. Expert Systems with Applications. 2025;269:126391. DOI: 10.1016/j.eswa.2025.126391.
57. Zhou W, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing. 2004;13(2):600-12. DOI: 10.1109/TIP.2003.819861.
58. Setiadi DRIM. PSNR vs SSIM: imperceptibility quality assessment for image steganography. Multimedia Tools and Applications. 2021;80(6):8423-44. DOI: 10.1007/s11042-020-10035-z.
59. Agustsson E, Timofte R. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW): IEEE Computer Society; 2017. p. 1122-31. DOI: 10.1109/cvprw.2017.150.
60. Kinga D, Adam JB, editors. A method for stochastic optimization. International conference on learning representations (ICLR); 2015: California;.
61. Mehta, S., & Rastegari, M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. ArXiv. 2021; DOI: 10.48550/arXiv.2110.02178.
62. Kishore V, Chen X, Wang Y, Li B, Weinberger KQ, editors. Fixed neural network steganography: Train the images, not the network. International Conference on Learning Representations; 2021.
63. Zhang S, Li H, Li L, Lu J, Zuo Z. A high-capacity steganography algorithm based on adaptive frequency channel attention networks. Sensors. 2022;22(20):7844. DOI: 10.3390/s22207844.
64. Chen X, Kishore V, Weinberger KQ, editors. Learning iterative neural optimizers for image steganography. The Eleventh International Conference on Learning Representations; 2023. https://openreview.net/forum?id=gLPkzWjdhBN.
65. Reinel T-S, Brayan A-AH, Alejandro B-OM, Alejandro M-R, Daniel A-G, Alejandro A-GJ, et al. GBRAS-Net: A convolutional neural network architecture for spatial image steganalysis. IEEE Access. 2021;9:14340-50. DOI: 10.1109/ACCESS.2021.3052494.

Persian Abstract

چکیده

نهان‌نگاری تصویر با به‌کارگیری یادگیری عمیق و به‌ویژه شبکه‌های عصبی کانولوشنی برای استخراج ویژگی‌های پیچیده، پیشرفت چشمگیری داشته است. در حالی که شبکه‌های کانولوشنی به دلیل کارایی بالا در شناسایی وابستگی‌های محلی، مدت‌هاست به عنوان ابزاری قدرتمند در پردازش تصویر شناخته می‌شوند، ظهور مدل‌های ترنسفورمر تحول بزرگی در این حوزه ایجاد کرده است زیرا این مدل‌ها با دقت بالاتر و توانایی درک روابط سراسری در داده‌ها، عملکرد بهتری ارائه می‌دهند. با این حال، ترنسفورمرها در پردازش تصاویر با وضوح بالا با چالش‌هایی مانند هزینه محاسباتی و مصرف حافظه مواجه هستند. برای رفع این محدودیت‌ها، ما معماری هیبریدی‌ای پیشنهاد می‌دهیم که در مراحل با وضوح بالا از لایه‌های کانولوشنی و در مراحل با وضوح پایین از بلوک‌های ترنسفورمر استفاده می‌کند. این طراحی با بهره‌گیری از لایه‌های کانولوشنی برای استخراج وابستگی‌های کوتاه‌برد و ترنسفورمرها برای مدل‌سازی روابط بلندبرد، تعادلی میان کارایی و عملکرد ایجاد می‌کند. همچنین با جایگزینی کانولوشن‌های استاندارد با کانولوشن مختصات، آگاهی فضایی و قابلیت مکان‌یابی ویژگی‌ها در مدل را بهبود می‌بخشیم. ما چارچوبی نوآورانه برای نهان‌نگاری تصویر به نام VidaFormer معرفی می‌کنیم که یادگیری عمیق را با این بهبودهای معماری ترکیب می‌کند تا داده‌های دودویی دلخواه را در تصاویر جاسازی کند و در عین حال کیفیت بصری بالای تصاویر نهان را حفظ نماید. روش پیشنهادی با دستیابی به ظرفیت مؤثر ٤/٨٩ بیت بر پیکسل روی دادگان DIV2K که ٪۱/۷ بهتر از بهترین مدل قبلی است، عملکرد بهتری نسبت به مدل‌های پیشرفته فعلی دارد. این نتایج نشان‌دهنده عملکرد و مقیاس‌پذیری برتر VidaFormer برای کاربردهای مدرن نهان‌نگاری تصویر است. کد منبع در نشانی https://github.com/vidayousefi/vidaformer-ije در دسترس است.