Journal of Soft Computing and Computer Applications

Manuscript 1022

Review of Video Steganography by using Deep Learning Methods: Datasets, Techniques, and Evaluations

Noor Fahem Sahib

Soukaena Hassan Hashem

Ekhlas Falih Naser

JOURNAL OF SOFT COMPUTING AND COMPUTER APPLICATIONS 2025;2:1022

REVIEW

Review of Video Steganography by using Deep Learning Methods: Datasets, Techniques, and Evaluations

Noor Fahem Sahib [a,*], **Soukaena Hassan Hashem** [b], **Ekhlas Falih Naser** [c]

[a] University of Technology – Iraq, College of Computer Science, Al-Sina'a St., Al-Wehda District, 10066 Baghdad, Iraq
[b] University of Technology – Iraq, College of Computer Science, Department of Computer Security and Cybersecurity, Al-Sina'a St., Al-Wehda District, 10066 Baghdad, Iraq
[c] University of Technology – Iraq, College of Computer Science, Department of Information Systems, Al-Sina'a St., Al-Wehda District, 10066 Baghdad, Iraq

ABSTRACT

The growing prevalence of cyber threats, including fraud and attacks, has intensified the demand for secure methods of safeguarding confidential information exchanged between users. As telecommunications increasingly rely on multimedia data, video steganography has become a prominent technique to address these concerns. By embedding sensitive data within video files, this approach enhances protection against unauthorized access and common internet-based attacks, offering a robust layer of security in an era of escalating digital risks. With the introduction of Deep Learning (DL) steganography methods recently, video steganography can be defined as a rapidly developing subject within information security. This study provides a thorough analysis of the many DL-based video steganography methods that have been covered in the literature. Those methods usually have two modules: a decoder and an encoder. Those techniques have assisted in resolving issues with visual quality as well as capacity, which have been previously evident in older techniques. The study also presented a description of the main databases used in DL-based video steganography, where the University of Central Florida 101 (UCF101) database is the most used in this field. Mean Squared Error (MSE), Peak-Signal Noise-Ratio (PSNR), Average Pixel Difference (APD), and Structural Similarity Index (SSIM) have been the top-rated evaluation metrics utilized to assess the efficacy of the video steganography method. All things considered, this paper bridged the gap between two research domains, video steganography and DL, and was a useful tool for academics studying the subject. Previous studies have demonstrated that

Received 18 March 2025; revised 11 July 2025; accepted 8 August 2025.
Available online 26 December 2025

* Corresponding author.
E-mail addresses: cs.24.08@grad.uotechnology.edu.iq (N. F. Sahib), soukaena.h.hashem@uotechnology.edu.iq (S. H. Hashem), ekhlas.f.naser@uotechnology.edu.iq (E. F. Naser).

https://doi.org/10.70403/3008-1084.1022

Reversible Neural Networks (RNNs) and Generative Adversarial Networks (GANs), which rely on Convolutional Neural Networks (CNNs), have given high results.

Keywords: Video steganography, Deep learning, Data hiding, Convolutional neural networks, Generative adversarial networks

1. Introduction

Data hiding, which embeds messages inside cover content and is commonly used in industries including medical systems, access control, copyright protection, and law enforcement, is a subset of video steganography [1]. Video steganography makes it possible to hide messages within video content while hiding the transmission method because the human visual system is less sensitive to slight changes in digital media, especially digital video. Because of two major factors, such method has become more and more important recently: first, security concerns in the information domain have increased because of the quick advancement regarding computer applications; second, video is an electronic medium that is well suited for steganography because of the emergence of powerful digital tools for sharing and transmitting video and because it has a larger data capacity than other multimedia formats. To protect secret data from online hacking attempts and frequent attacks, video steganography has gained popularity [2].

The carrier video could be in compressed or raw format, and the hidden information (secret) could be any kind of media, such as audio, text, video, images, or binary files. A thorough classification of video steganography techniques according to different standards. The cover video format, which could be either compressed or raw, is the major criterion for classification [3]. Videos in the raw domain are further divided into transform as well as spatial domains. Methods like Least Significant Bit (LSB) substitution and other important approaches are included in the spatial domain [4]. Traditional steganography techniques hide data within images or video files by subtly altering redundant bits. However, these approaches often lack robustness against compression, noise, or advanced detection algorithms. Additionally, their limited payload capacity and susceptibility to statistical analysis make them unreliable for modern security needs [5]. On the contrary, the transform domain converts the cover video before embedding the secret data using the Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT). Compressed cover videos are used in video steganography in the compressed domain since they take up less storage space than raw videos. Either following or throughout video compression, the embedding process occurs. Intra-prediction modes, motion vectors, DCT/Discrete Sine Transform (DST), and entropy coding modules are often employed methods in this field [6]. Three major problems impact conventional techniques for concealing data in video files [7]:

- Fixed Encoding Techniques: Static approaches increase the danger of discovery and decrease security because they are simpler to decode.
- Low Embedding Capacity: Significant quality loss is frequently caused by a limited hidden data capacity.
- Ignoring Structural Patterns: The efficiency of embedding is decreased when inherent patterns in videos or images are not used.

To increase imperceptibility and capacity for secure steganography, such difficulties underscore the necessity for advanced techniques, like DL, such as Generative Adversarial

Networks (GANs) and Convolutional Neural Networks (CNNs). A decoder and an encoder are usually the two main parts of DL-based steganography. A steganography carrier is the result of the encoder embedding secret information into a carrier. The hidden information is subsequently extracted from the steganography carrier by the decoder. This approach is frequently called an end-to-end model [8].

End-to-end DL-based steganography, as opposed to conventional methods, does away with the necessity of human feature extraction, enabling Neural Networks (NNs) to effectively incorporate secret data and automatically learn features. Video steganography models showed outstanding visual quality and a large payload capacity. According to experimental findings, DL-based steganography performs better than traditional techniques in terms of security and efficacy [9].

The contribution of this paper is to examine the literature on DL-based video steganography and explain the power of DL in this field. Also, this paper presents insights into future research in video steganography with highlighting significant advancements and difficulties.

The paper is organized as follows: Section two illustrates some literature review that is related to the paper's topic. Section Three explains the datasets, deep learning techniques, and the evaluation metrics that have been used with video steganography. The challenges that video steganography presents are presented in Section Four. Finally, Section Five gives the conclusion that has been drawn from this paper.

2. Literature review

Over time, steganography has changed significantly. The earliest methods were simple and involved hiding secret messages in another medium. Those straightforward techniques were more vulnerable to detection, though. More secure steganography methods were therefore required. As digital signal processing has advanced, many advanced digital steganography techniques have been created.

Abdulmunem et al. [8] concentrate on creating a steganography technique that uses CNN to effectively hide data (videos or images) inside the frames of a video file. CNN is used to train the suggested model on randomly chosen images from the ImageNet database. The method uses automatic encoding networks to compress images. Compressing secret images into the least visible parts of the cover image is the goal of the training system. It should also determine the best methods for recovering and reconstructing the hidden data with the least amount of loss. The network architecture is based on the structure of auto-encoders, which are often made to learn key features from the input distribution by reconstructing input data after several modifications. The goal of the end-to-end video steganography technique proposed by Xu et al. [9] is to conceal n-bit binary messages within a cover video. It is built on a multi-scale DL network with a decoder, an encoder, and a discriminator, as well as a GAN. Yet, throughout the transmission process, video will inevitably be encoded (the video is first split into image block units of varying sizes, and after that, each unit is subjected to intra-frame or inter-frame prediction). Significant evaluation metrics for Stego's performance include message embedding capacity) Bits Per Pixel (BPP), robustness to video compression (accuracy), message extraction accuracy, and video visual quality (PSNR). A method for concealing a video inside another video is put forth by Weng et al. [10]. There are two contributions: inter-frame residuals and the original secret frame. Inter-frame residuals that have two branches and take advantage of the residuals' sparsity between successive frames. The inter-frame residuals are concealed in a cover video frame by one branch. The original secret frame is concealed in the cover video by the other. Either

the original secret frame or the residuals can be extracted using one of two decoders. The model makes use of DL CNNs, which is the first time this technology has been used in video steganography. PSNR (40.62) is used to assess the model's performance. To finish the video hiding and recovery process, multiple distinct stages are needed. With the use of 3D-CNN as well as micro bottleneck Video Steganography Network (VStegNET)), Mishra et al. [11] present a new framework for video steganography that embeds secret RGB video frames into cover video frames of the same size while capturing temporal and spatial data from video frames. The University of Central Florida 101 (UCF101) action recognition video dataset was used to train the model, and PSNR (35.67), one of several quantitative indicators, was used to assess its performance. Yet, it has a significant model complexity (367.2M model parameters) and uses two distinct 3D U-Nets for performing recovery and hiding. An innovative deep 3D CNN architecture inspired by conventional auto-encoders for video steganography was introduced by Jaiswal et al. [12], who investigate the task of embedding a full-sized video into another video of the same proportions. The suggested model performs better than the state-of-the-art VStegNET. The efficacy of the model is demonstrated by both qualitative and quantitative assessments at the frame and video levels utilizing metrics including SSIM (0.95), APD (2.87), Visual Information Fidelity (VIF) (0.74), and PSNR (36.62). The suggested architecture includes a down-sampling step after every two up-sampling steps, in contrast to traditional decoders that only use continuous up-sampling for reconstruction. This method improves performance by improving feature representation at every stage of reconstruction. Kumar et al. [13] suggest utilizing CNNs for feature extraction as well as processing to conceal one image inside another. Hiding network/encoder (which processes the secret image through extracting features and embedding them into the cover image) and reveal network/decoder (which extracts and reconstructs the hidden secret image to recover it in its original form) are the two main modules of the approach, which builds on the principles of auto-encoder architecture. Images of different sizes from the ImageNet and Pascal Visual Object Classes (Pascal-VOC) datasets are used to test the model. PSNR (37.554) and SSIM (0.97554) measures are utilized to assess its performance. The model had to deal with dynamic or complex data as well as practical issues like computing power and deployment viability. Pyramid Generative Adversarial Network (PyraGAN), an end-to-end video steganography technique based on a pyramid-like generative adversarial network, is proposed by Chai et al. [14] to embed messages made up of randomly generated 0 and 1 bits. Three essential parts make up the model's multiscale down-sampling feature extraction structure: a decoder, a discriminator network, and an encoder. A Versatile Video Coding (VVC) video is used to create a Coding Unit (CU) mask, which improves the network's learning capacity. To further enhance the steganography video's visual quality, an attention mechanism is also included. With a PSNR of 43.109, experimental results show that the suggested steganography network performs exceptionally well in terms of stego video perceptual quality.

With the use of Neural Representation for Videos (NeRV), Biswal et al. [15] investigate a new method of video steganography in which an Implicit Neural Network (INR) conceals a distinct image within each frame of a cover video. To align the weights across the many stages of NeRV with the goals of steganography, a stage-wise gradient scaling technique is suggested. To conceal larger images, an attention-guided technique is also used to concentrate on embedding information in the high-texture regions of the cover frames. The Invertible Neural Networks (INNs) were first introduced by Dinh in 2014. These networks learn a stable bijective mapping between a data distribution [16]. A Large-Capacity and Flexible Video Steganography Network (LF-VSN) is presented by Mou et al. [17]. Multiple videos can be embedded as well as retrieved with the use of a

single INN thanks to the creation of a reversible pipeline that achieves great capacity. A key-controllable approach is suggested for increased flexibility, which enables various receivers to use distinct keys to extract particular hidden videos from the same cover video. High security, adaptability, and a big concealing capacity are features of LF-VSN. Jumping et al. [18] present Hierarchical invertible Network (HiNet(, based on INN, designed to address challenges in image hiding by simultaneously learning both the concealing and revealing processes. This approach enables the embedding of a full-size secret image into a cover image of the same dimensions, ensuring high capacity and imperceptibility. Weida et al. [19] effectively mitigate information loss in invertible NN-based image steganography by incorporating a private key, enhancing both the security and integrity of the hidden information. Several other studies have employed INNs in steganography. For instance, Yanzhen et al. [20] explored hiding N secret images within a cover image, while Hang et al. [21], Liang et al. [22], Shao-Ping et al. [23], and Youmin et al. [24] investigated embedding an image within another image. These INN-based methods effectively retrieve concealed data due to their reversible architectures.

Table 1 summarizes the DL methods for steganography that have been applied in related publications. CNNs are among the most widely used DL techniques by researchers in the field of steganography, such as Kumari et al. [25], Vergara et al. [26], Aiman et al. [27], Xintao et al. [28], and Hadjer et al. [29]. Among these DL methods, Weng et al. [10], Mishra et al. [11], and Aiman and R. [27] are the first authors who employ steganography in video using cover media and CNNs. Recently, GANs and INNs, applied in video steganography and image steganography, have delivered remarkable results. Therefore, researchers are encouraged to adopt these techniques because they have achieved high results.

3. Deep learning video steganography

A decoder and an encoder are typically the two modules that make up DL-based steganography, as seen in Fig. 1. The encoder (sender) starts with an original cover video and uses an embedding algorithm based on DL to conceal the server data (secret message) within the video frames. This produces a container video, which appears normal but contains the hidden information. The decoder (receiver) obtains the container video and applies an extraction algorithm, also powered by DL, to retrieve the hidden data. The process ensures secure communication, as third parties remain unaware of the embedded message. Additionally, steganalysis represents potential detection efforts by adversaries trying to identify whether a video contains hidden data. This method is frequently called an end-to-end model [11].

3.1. Dataset used in video steganography

More than 0.3 million videos with excellent annotations make up the Text Retrieval Conference (TREC) Video Retrieval Evaluation (TRECVID) dataset. The Fast Forward Moving Picture Experts Group (FFMPEG) utility is used for extracting 24 frames from a randomly chosen 2-second clip for every video [30]. A popular video action recognition dataset called UCF101 was created to assess Machine Learning (ML) models in the context of video analysis. A total of 13,320 video clips covering sports, everyday activities, and other topics are arranged into 101 action classes. Each frame includes three Red, Green, and Blue (RGB) color channels and 320 by 240-pixel resolution. Algorithms for real-world human action recognition are benchmarked and trained using this dataset. For processing, every

Table 1. Summarization of DL techniques in video and image steganography.

Ref.	Steganography Technique	Secret media type	Evaluation Metrics	Strength	Limitation
[8]	Auto-encoders CNNs	Image and video	APD = 4.16	Increased ability to hide and increased security	Training model time is high
[9]	GAN and multi-scale DL	n-bit binary	PSNR = 38.493	Achieve large embedding capacity, strong robustness and high visual quality	High training method
[10]	CNNs and inter-frame residuals	Video	PSNR = 39.75	Pure high-capacity image, achieve the model's security against steganalysis, and the robustness to video compression	used a 2D-CNN (lack of motion modelling)
[11]	3D-CNN and micro bottleneck (VStegNET).	Video	PSNR = 35.67	Exploiting 3D-CNN for spatial and temporal feature capture	has a significant model complexity (367.2M model parameters)
[12]	Novel 3D CNN architecture inspired by auto-encoders VStegNET	Video	PSNR = 36.62	Surpasses existing state-of-the-art methods, demonstrating superior performance in both qualitative and quantitative assessments	Training model time is high
[13]	Encoder and decoder CNNs	Image	PSNR = 37.554	Achieve high payload capacity with good imperceptibility	Deal with dynamic or complex data as well as practical issues like computing power
[14]	PyraGAN, based on a pyramid-like GAN	n-bit binary	PSNR = 43.109	A versatile video coding is used to create a coding unit mask, which improves the network's learning capacity	Has significant model parameters
[15]	NeRV	Image	PSNR = 34.168	Using an attention-based strategy increases the capacity to conceal	Has significant model parameters
[17]	LF-VSN with a single INN	Multiple videos	PSNR = 40.97	Embedding great capacity (multiple videos)	Training model time is high
[18]	HiNet by using INN	Image	PSNR = 44.60	Robust against steganalysis attacks	Low embedding capacity
[19]	INN	Image	PSNR = inf	Achieving a balance among capacity, invisibility, and security issues	Training model time is high
[20]	Framework of conditional Invertible Neural Network for Image Steganography (Steg-cINN) based on Conditional INN	n image	Revealing accuracy = 100%	Enhances steganography capacity	High training method

(Continued)

Table 1. Continued.

Ref.	Steganography Technique	Secret media type	Evaluation Metrics	Strength	Limitation
[21]	Practical Robust Invertible network for image Steganography (PRIS)	Image	PSNR = 37.99	To enhance the robustness of NN, they introduce a structured three-step training process	Low embedding capacity
[22]	Conditional INN	Image	PSNR = 43.62	Produce high-quality steganography images that retain rich semantic content and clear spatial details	Embedding capacity is low
[23]	Invertible Steganography Network (ISN) based on INN	Image	PSNR = 38.05	Achieve high payload capacity with good imperceptibility	Training model time is high
[24]	Robust Invertible Image Steganography (RIIS) based on INN.	Image	PSNR = 28.08	Implicitly preserving the crucial details needed for successful retrieval by modulating the high-frequency distribution	Payload capacity limited
[25]	GAN and 3D-CNN.	Image	PSNR = 36.56	Using encryption and scrambling	practical issues like computing power
[26]	Spatial-Temporal Adaptive Filter Network (STFAN).	Image	PSNR = 27.0164	Enhance the efficacy and efficiency of frame processing by taking advantage of the redundancy in the video's temporal dimension	High model training time
[27]	Convolutional video steganography using the temporal residual method	Video	MSE = 0.4167	Using the temporal residual modelling method. The model deals with the residual frame and reference frame individually	Training model time is high
[28]	A novel Stego CNN method	Image	PSNR = 35.852	Automatically learn high-dimensional features, realize high-capacity and good generalization	High training model time
[29]	Mask Region-based Convolutional Neural Network (Mask-RCNN)	Image	PSNR = 36.58	High-capacity embedding, confidentiality is maintained, and ensures image quality	High dimensionality

video is divided into its frames [31]. Video clips reduced to 64 × 64 resolution are included in the Vimeo-90k (video sharing platform) dataset. Five continuous frames make up each cover video, guaranteeing temporal feature continuity as well as similarity [32]. Instead of using videos as a cover, ImageNet relies mainly on images, which are utilized as secret data. Nonetheless, a related dataset known as ImageNet video has over 4,000 video clips with a resolution of 1980 × 1024 pixels and a frame rate of 25 frames per second (fps) [33]. DL and computer vision both make extensive use of the DIVerse 2K)DIV2K(and Microsoft

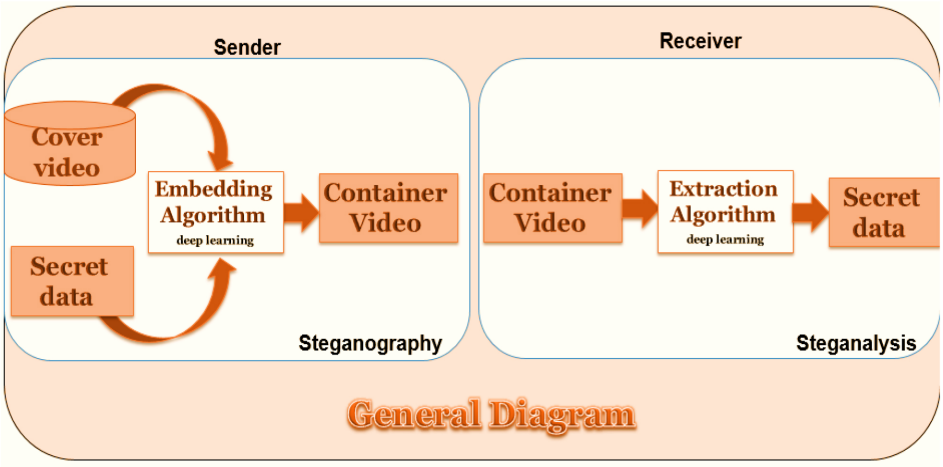

Fig. 1. General diagram of DL video steganography.

Common Objects in Context)MSCOCO(datasets. The 1,000 high-resolution natural scene images that make up DIV2K are divided across 800 training and 100 validation sets. It is typically employed for tasks involving image enhancement and super-resolution. A total of 118,060 training images of everyday objects in a variety of complex backgrounds and at smaller sizes are included in MS COCO. It is extensively utilized for applications including image captioning, segmentation, and object detection [34, 35]. One popular benchmark for motion analysis as well as video object segmentation is the Densely Annotated Video Segmentation (DAVIS) dataset. It includes 150 videos with excellent annotations at 1080p and 480p resolutions [36]. The Ultra Video Group (UVG) dataset includes 16 different types of 4K video at 3840 $\times$ 2160 quality. The dataset is perfect for a high-quality video compression and quality assessment study because these natural video sequences were recorded at either 50 fps or 120 fps [37]. BOSSBase v1.01 is a standardized benchmark dataset, containing 10,000 high-resolution 512 $\times$ 512 images originally captured from seven digital cameras. It includes both pristine “cover” images and manipulated “stego” versions. The dataset provides calibrated, artifact-minimized images converted to 8-bit PNG format to simulate real-world conditions. Since its release in 2010, BOSSBase has served as the foundational resource for developing and evaluating AI-driven steganalysis tools, enabling rigorous testing of data-hiding robustness and detection methods in digital images [38]. The data set utilized for steganography and the reference that was used are listed in Table 2.

3.2. Deep learning techniques in video steganography

There are different techniques to embed a secret message in the video frames. This paper focused on the following techniques:

3.2.1. Convolutional neural networks

CNNs have become quite popular because they can achieve high accuracy, minimize feature maps, employ shared weights and local connections (filters), adapt to different applications, and extract features without the need for manual way [39, 40]. In tasks including classification, pattern recognition, object detection, and image segmentation, their efficacy was proven. Their application in steganography has been investigated recently. As

Table 2. Summary of the database for steganography.

Dataset and Ref.	Number of samples	Availability	Description of the database
TRECVID 2017 [30]	0.3 million video clips	Freeware	Extracting 24 frames from a randomly chosen 2-second clip. Different resolution for each video
UCF101 [31]	13,320 video clips	Freeware	Each frame includes three RGB color channels and a 320 by 240-pixel image, arranged into 101 action classes
Vimeo-90k [32]	90,000 video clips	Not applicable	Five continuous frames make up each cover video (64, 64)
ImageNet [33]	14 million images, 4,000 video clips	Freeware	Clips with a resolution of 1980 × 1024 pixels and a frame rate of 25 fps
DIV2K [34] and MSCOCO [35]	1000 images, 118,060 images	Freeware	Employed for tasks involving image enhancement and super-resolution for DIV2K. MSCOCO(640 × 480).
DAVIS [36]	150 videos	Freeware	Includes 150 videos with excellent annotations at 1080p and 480p resolutions
UVG [37]	4K video	Freeware	Perfect for high-quality video compression and quality assessment study, resolution (3840×2160 (4K UHD), some originals 4096×2160)
BOSSBase V1.01 [38]	10,000 images	Freeware	512 × 512, Portable Grey Map of 8-bit

illustrated in Fig. 2, popular designs for image segmentation, such as Volumetric Net (V-Net), U-Net (“U”-shaped architecture with contracting (encoder) and expanding (decoder) paths) were firstly present in [41], and U-Net + + (Nested skip pathways for better gradient flow (2D/3D)), are currently being employed for video (or image) steganography. With a small number of training images, U-Net, a fully CNN, can achieve great performance. It has two primary paths with 23 convolutional layers each: an expansion path (decoder) and a contraction path (encoder). The number of feature channels doubles at each stage as each layer in the contraction path applies two 3 × 3 filters to extract initial features and boost

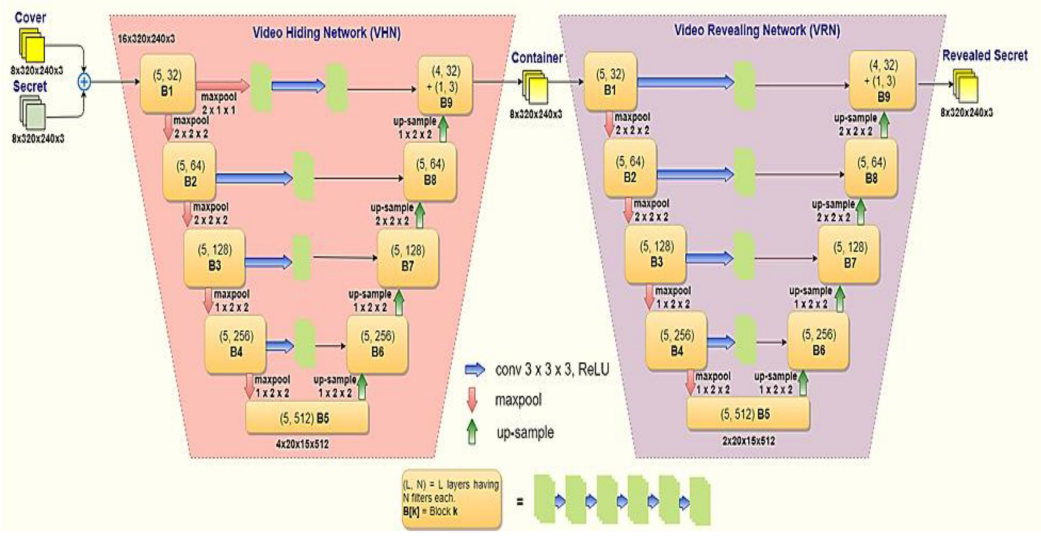

Fig. 2. StegNET’s architecture [11].

Table 3. Summary of the pros and cons of using CNNs in steganography.

Metrics	Prons	Cons
Training time	Efficient on modern Graphics Processing Units (GPUs), enabling embedding and extraction in real-time	Requires significant computational resources, especially for large datasets
Embedding capacity	Allowing high-capacity secret embedding	Overfitting on small datasets may reduce the effectiveness of the secret message
Imperceptibility	Preserving image quality and optimize embedding to be visually undetectable	Poorly trained models can reveal the hidden message
Security	It is difficult to detect by steganalysis due to adaptive feature learning	Vulnerable to DL-based steganalysis if not designed properly
Robustness	CNNs can improve resistance against compression, distortions and noise	It may have difficulty countering advanced steganalysis attacks if adversarial training is not used

feature extraction. The feature maps are then down-sampled using a 2 × 2 max-pooling operation. The feature channels are cut in half at each stage of the expansion path, also known as the decoder, which uses 2 × 2 deconvolution filters. Furthermore, horizontal connections preserve spatial information lost throughout the contraction phase by transferring extracted features from the encoder to the decoder. By keeping crucial features, this improves the quality of the final rebuilt image [42]. To find and alter areas of video frames for embedding, use CNNs. CNNs are effective in hiding information by learning spatial patterns in video frames (effective for spatial domain embedding). For robustness, it could be combined with other networks. Methods for CNN-based video processing including a third dimension in the convolution process, 3D CNNs go beyond conventional 2D CNNs. The network can simultaneously learn temporal and spatial features from the video thanks to this third dimension, which represents time (or depth). The filters (kernels) in a 3D CNN slide over the depth (the time axis, which represents several frames) in addition to the width and height of frames, as in 2D CNNs [11]. There are several pros and cons of using CNNs in steganography, as shown in Table 3 [43, 44].

3.2.2. Generative adversarial networks

GANs [45] have been receiving more and more attention since they were first introduced. In tasks including style transfer, speech synthesis, and image reconstruction (image generation), they have attained state-of-the-art performance. GANs were initially used to apply DL to image steganography. They are quite good at concealing information in images because of their capacity to produce realistic data. The performance and security of GAN-based steganography have increased over time [44]. Fig. 3 depicts the GAN's architecture, which is made up of three primary parts as follows:

1. To create the stego image, Encoder (Generator)G () incorporates the secret message into the cover image. It processes the image using convolutional layers, embeds the hidden message, and preserves the quality of the cover image by using residual blocks or skip connections.
2. Discriminator)D(attempts to differentiate between the innocent cover images and the stego images, which are produced with the use of the change probabilities supplied by CNN that classifies incoming images as real or fake and produces a single value for each image.
3. The decoder ensures dependability in the embedding process by recovering the secret data from the stego image x_m. x_m is produced by the G using the cover image x and secret data m as inputs. Both x_m and x are inputs to the D, which then gives the

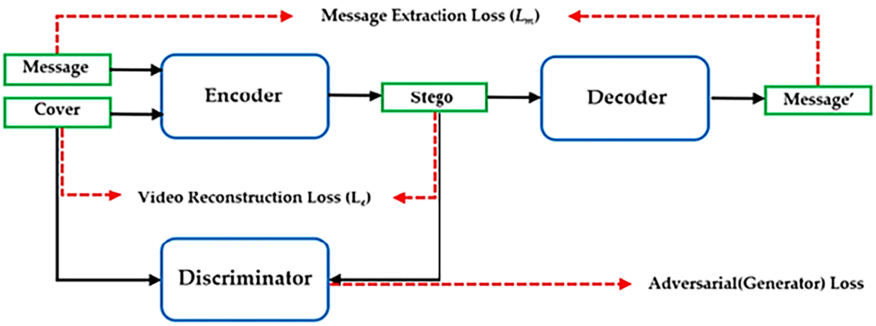

Fig. 3. GANs architecture.

Table 4. Summarization of the pros and cons of using GANs in steganography.

Metrics	Advantage	Disadvantage
Training time	Capable of real-time embedding and extraction with optimized architectures	Requires high computational power (GPUs) and large datasets for effective training
Embedding Capacity	Can embed larger amounts of data while maintaining high visual quality	May suffer from reduced embedding capacity if constraints are too strict
Security	GANs generate highly imperceptible stego-images, making detection by traditional methods difficult	Susceptible to DL-based steganalysis if not properly trained
Imperceptibility	Adversarial training helps generate stego images that closely resemble natural images	Poor training can lead to visible distortions or artifacts in the stego image
Robustness	More resilient to distortions, noise, and compression attacks compared to traditional methods	Still vulnerable to adversarial attacks if not properly trained with diverse data

generator feedback to increase x_m's realism. Lastly, the decoder guarantees that x_m could be reliably used to reconstruct the secret (s) data [9, 14, 46]. There are several advantages and disadvantage of using GANs in steganography as shown in Table 4 [44]. The network framework uses three loss functions: Loss1 (L1) for the encoder, Loss2 (L2) for the decoder, and Loss3 (L3) for the discriminator through multiple rounds of backpropagation and parameter updates these losses are minimized resulting in an optimized model. L1 (MSE loss) ensures the stego video (Vstego) closely resembles the cover video. L2 (cross-entropy loss) measures the decoder's accuracy in reconstructing the hidden message from the input message. L3 (D loss) evaluates the discriminator's ability to distinguish between cover and stego videos. During training the encoder and decoder are jointly optimized using a combined loss [9].

3.3. Evaluation metrics

There are several criteria for steganography to evaluate the results. They can be divided into objective criteria that rely on pixels in their calculation, and subjective criteria that rely on human visual perception. The evaluation metrics commonly used in steganography are briefly outlined in related studies, as shown in Table 5.

- **Objective Metrics** are mathematical and computational models used to evaluate the quality of images, videos, or other media without human intervention. These metrics rely on algorithms to compare processed content with the original. The typical criteria utilized in related publications to assess steganography measure the method's robustness,

security, imperceptibility, and capacity. The frequently utilized metrics unique to steganography are listed below:

1. Imperceptibility measurements assess the stego image's visual quality. The PSNR calculates the ratio of the MSE between the stego image and the cover image to the maximum pixel intensity. Decibels (dB) are used to measure PSNR [47, 48]. The VIF standard is an image quality metric based on information theory that measures the amount of "visual information" shared between the original image and the distorted image (after data hiding). It is calculated through complex steps involving image analysis in the wavelet domain and modelling the statistical properties of human perception [10, 49]. The Structural Similarity Index (SSIM) compares the brightness, contrast, and structural information of the cover and stego image to determine how similar they are perceptually. SSIM values range from 0 to 1, where 1 denotes complete similarity. The average absolute difference between the pixel values of the cover image and stego image is measured by the MSE [50]. The average difference between the pixel values in the stego image and the cover image is determined by the APD.
2. Embedding capacity metrics evaluate the amount of secret data that can be embedded into the image. Typically measured in Bits Per Pixel (BPP) or Bits Per Frame (BPF).
3. Robustness Metrics assess the resilience of the stego video to various attacks or distortions. A robust method allows the secret data to be extracted even after the video is compressed. Evaluates the resilience of the hidden data when noise (e.g., Gaussian noise) is added to the video. Assesses the ability to recover the hidden data after operations like scaling, cropping, rotation, or frame dropping. Bit Error Rate (BER) is the proportion of bits in the secret message that are incorrectly extracted after the video undergoes attacks or distortions.
4. Security Metrics (metrics evaluate the resistance of the stego video to detection or extraction by unauthorized parties). False Positive Rate (FPR) is the probability that a clean (cover) video is incorrectly classified as containing hidden data [51, 52].

❖ **Subjective Metrics** relies on direct human vision (Human Subjective Evaluation), where experts or ordinary individuals assess visual content (images/videos) based on their perception. These methods are considered the most accurate for measuring real visual experience, but they are costly and time-consuming compared to computational metrics such as PSNR or SSIM. These metrics include [53]:

1. Mean Opinion Score (MOS): A group of people (experts or non-experts) is asked to rate the quality of an image/video on a defined scale (e.g., from 1 to 5, where 5 = excellent, 1 = bad).

Table 5. Summarizes the evaluation metrics for steganography that have been applied in related publications.

Metrics	Description
PSNR	Calculates the ratio of the MSE between the stego image and the cover image to the maximum pixel intensity
VIF	Measures 'visual information fidelity' using information theory and wavelet analysis
SSIM	Compares the brightness, contrast, and structural information of the cover and stego image to determine how similar they are perceptually
MSE	Computes the average absolute difference between the pixel values of the cover image and the stego image
BER	The proportion of bits in the secret message that are incorrectly extracted after the image undergoes attacks or distortions
BPP	Embedding capacity metrics evaluate the amount of secret data that can be embedded into the image

2. Differential Mean Opinion Score (DMOS): Evaluates the difference in quality between the original version and the modified version (instead of assessing each separately).

4. Challenges

Some challenges that appeared in video steganography research are:

1. Artificial Intelligence (AI)-based methods, especially DL, have revolutionized image and video steganography. They significantly enhance imperceptibility, making the hidden data nearly undetectable to the human eye. DL techniques improve robustness, ensuring that stego images or videos withstand various types of attacks or distortions. These methods also increase efficiency, reducing the time and computational resources needed for data embedding and extraction.
2. DL methods offer optimal and adaptive solutions for steganography. These techniques can dynamically adjust to different data-hiding challenges. Public datasets (image, video) are essential for evaluating the performance of DL-based approaches. Commonly used datasets include ImageNet, BOSSBase V1.01, UCF101, MS COCO, and Vimeo-90k. These datasets provide diverse video samples for training and testing models. General datasets lack realism and security relevance for steganography, necessitating specialized benchmarks that address challenges like attack resistance and adaptive embedding. Such standards would enable fairer evaluations and advance secure DL-based systems in real-world applications.
3. Balancing visual quality, robustness, and embedding capacity is challenging in DL steganography. High visual quality ensures hidden data remains undetectable to the human eye. Strong resistance to attacks protects embedded data from various threats. Maintaining a high embedding capacity allows for more data to be hidden efficiently. Achieving this balance requires advanced DL architectures and optimization techniques. Striking an optimal balance between visual fidelity and embedding capacity remains a pivotal challenge driving future research in this domain. Achieving this equilibrium requires embedding the maximum feasible payload within video carriers while rigorously preserving imperceptibility—a cornerstone for advancing practical steganographic solutions.
4. Strengthening security by integrating cryptography of hidden data in steganography. Error-correcting codes help maintain data integrity against transmission errors. Combining these techniques strengthens resistance to attacks and unauthorized access. Enhanced security makes video steganography suitable for unsafe application scenarios. This approach ensures reliable data hiding in sensitive communication environments. Continuous improvements are needed to adapt to evolving security threats, and error-correcting codes are essential, which can enhance the security of the secret data and make the video steganography appropriate for various unsafe application scenes.
5. The robustness of video steganography methods is evaluated by testing their resistance to noise, compression, and manipulation attacks, while security is assessed through resistance to steganalysis attacks. However, most methods are not tested against all potential attacks, and many studies fail to report their resistance results, hindering objective comparisons. The absence of unified evaluation metrics further complicates performance assessment, as authors often rely on inconsistent, custom criteria. This underscores the urgent need for standardized benchmarks and specialized databases to ensure comprehensive and fair evaluation of steganography techniques.

5. Conclusions

This review provided an overview of the basic concepts of video steganography in DL. In addition, a comprehensive review of the recent research that has used DL techniques in video steganography is presented. The datasets (the UCF101 is the most used in video steganography) and DL techniques in video steganography were mentioned in detail. CNN and GAN are mostly applied in many experiments in analyzing video steganography. DL-based steganography systems typically comprise two components: an encoder, which embeds secret data into a cover medium (e.g., video) to produce a stego carrier, and a decoder, which extracts the hidden information from the stego carrier. Different evaluation metrics were also used to assess the performance of DL methods for video steganography. These metrics were used to determine the imperceptibility capacity, security and robustness of their approaches. Consequently, PSNR, SSIM, MSE and VIF were widely employed for evaluating cover video and stego video. PSNR has shown significant variation, ranging from 27.0164 in Mou et al.'s method to 45.17 in Vergara et al.'s method, as shown in Table 1. Several critical gaps and limitations persist, including the lack of domain-specific datasets (e.g., medical or military applications) and standardized evaluation metrics to ensure comprehensive and fair evaluation of steganography techniques. Therefore, this study recommends that future research focus on addressing these limitations by developing video databases dedicated to steganography techniques, emphasizing the urgent need to create standardized methods for performance evaluation and comparison between different methods.

Acknowledgment

The authors wish to express their appreciation for the opportunities and resources that contributed to the completion of this research.

Authors contribution

All authors contributed equally to the preparation of this article.

Conflicts of interest

The authors declare no conflicts of interest.

Data availability

No data has been used here.

References

1. V. Kumar, P. Rao, and A. Choudhary, "Image steganography analysis based on deep learning," *Review of Computer Engineering Studies*, vol. 7, no. 1, pp. 1–5, Mar. 2020, doi: 10.18280/rces.070101.
2. M. Sha, "A reliable and secure video steganography method based on 4D-logistic mapping," *Multimedia Tools and Applications*, Feb. 2025, doi: 10.1007/s11042-025-20686-5.
3. N. Subramanian, O. Elharrouss, S. Al-Maadeed, and A. Bouridane, "Image steganography: A review of the recent advances," *IEEE Access*, vol. 9, pp. 23409–23423, Jan. 2021, doi: 10.1109/ACCESS.2021.3053998.

4. A. N. Mazhar and E. F. Naser, "Hiding the type of skin texture in mice based on fuzzy clustering technique," *Baghdad Science Journal*, vol. 17, no. 3, pp. 967–972, Sep. 2020, doi: 10.21123/bsj.2020.17.3(Suppl.).0967.
5. C. Chen, L. Qu, H. Amirpour, X. Wang, C. Timmerer, and Z. Tian, "On the security of selectively encrypted HEVC video bitstreams," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 9, Sep. 2024, Art. no. 288, doi: 10.1145/3672568.
6. S. Salunkhe and S. Bhosale, "RI-CDVS: Robust and imperceptible compressed domain video steganography using H.265 codec," *SN Computer Science*, vol. 4, no. 4, April 2023, Art. no. 357, doi: 10.1007/s42979-023-01681-9.
7. S. Salunkhe, S. Bhosale, and S. V. Narkhede, "An efficient video steganography for pixel location optimization using Fr-WEWO algorithm based deep CNN model," *I. J. Image, Graph. Signal Processing*, vol. 15, no. 3, pp. 14–39, June 2023, doi: 10.5815/ijigsp.2023.03.02.
8. I. A. Abdulmunem, E. S. Harba, and H. S. Harba, "Advanced intelligent data hiding using video stego and convolutional neural networks," *Baghdad Science Journal*, vol. 18, no. 4, 2021, Art. no. 20, doi: 10.21123/BSJ.2021.18.4.1317.
9. S. Xu, Z. Li, Z. Zhang, and J. Liu, "An end-to-end robust video steganography model based on a multi-scale neural network," *Electronics*, vol. 11, no. 24, Dec. 2022, Art. no. 4102, doi: 10.3390/electronics11244102.
10. X. Weng, Y. Li, L. Chi, and Y. Mu, "High-capacity convolutional video steganography with temporal residual modeling," in *Proc. 2019 Int. Conf. on Multimedia Retrieval (ICMR '19)*, Ottawa, ON, Canada, pp. 87–95. doi: 10.1145/3323873.3325011.
11. A. Mishra, S. Kumar, A. Nigam, and S. Islam, "VStegNET: Video steganography network using spatio-temporal features and micro-bottleneck," in *Proc. British Machine Vision Conf. (BMVC),* 2019.
12. A. Jaiswal, S. Kumar, and A. Nigam, "En-VStegNET: Video steganography using spatio-temporal feature enhancement with 3D-CNN and hourglass," in *Proc. 2020 Int. Joint Conf. on Neural Networks (IJCNN)*, Glasgow, UK, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9206921.
13. A. Kumar, R. Rani, and S. Singh, "Encoder-decoder architecture for image steganography using skip connections," *Procedia Computer Science*, vol. 218, pp. 1122–1131, 2023, doi: 10.1016/j.procs.2023.01.091.
14. H. Chai, Z. Li, F. Li, and Z. Zhang, "An end-to-end video steganography network based on a coding unit mask," *Electronics*, vol. 11, no. 7, April 2022, Art. no. 1142, doi: 10.3390/electronics11071142.
15. M. Biswal, T. Shao, K. Rose, P. Yin, and S. McCarthy, "StegaNeRV: video steganography using implicit neural representation," in *Proc. 2024 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshop (CVPRW)*, Seattle, WA, USA, pp. 888–898, doi: 10.1109/CVPRW63382.2024.00094.
16. L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear independent components estimation," 2014, *arXiv no: 1410.8516.*
17. C. Mou, Y. Xu, J. Song, C. Zhao, B. Ghanem, and J. Zhang, "Large-capacity and flexible video steganography via invertible neural network," in *Proc. 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR),* Vancouver, BC, Canada, pp. 22606–22615, doi: 10.1109/CVPR52729.2023.02165.
18. J. Jing, X. Deng, M. Xu, J. Wang and Z. Guan, "HiNet: Deep Image Hiding by Invertible Network," in *Proc. 2021 IEEE/CVF Inter. Conf. on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 4713–4722, doi: 10.1109/ICCV48922.2021.00469.
19. W. Chen and W. Chen, "Enhanced secure lossless image steganography using invertible neural networks," *Journal of King Saud University - Computer and Information Science*, vol. 36, no. 10, Dec. 2024, Art. no. 102259, doi: 10.1016/j.jksuci.2024.102259.
20. Y. Ren, T. Liu, L. Zhai, and L. Wang, "Hiding data in colors: secure and lossless deep image steganography via conditional invertible neural networks," 2022, *arXiv: 07444*.
21. H. Yang, Y. Xu, X. Liu, and X. Ma, "PRIS: Practical robust invertible network for image steganography," *Engineering Applications of Artificial Intelligence*, vol. 133, July 2024, Art. no. 108419, doi: 10.1016/j.engappai.2024.108419.
22. M. Liang and H. Zhao, "Research on image steganography based on a conditional invertible neural network," *Signal, Image and Video Processing*, vol. 19, no. 3, Jan. 2025, Art. no. 262, doi: 10.1007/s11760-025-03818-0.
23. S.-P. L. Lu, R. Wang, Rong, T. Zhong, and P. L. Rosin, "Large-capacity image steganography based on invertible neural networks," in *Proc. 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 10811–10820, doi: 10.1109/CVPR46437.2021.01067.
24. Y. Xu, C. Mou, Y. Hu, J. Xie, and J. Zhang, "Robust invertible image steganography," in *Proc. 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 7865–7874, doi: 10.1109/CVPR52688.2022.00772.
25. G. Kumari and G. Indra, "Video steganography with deep learning," *SSRN*, 2024, doi: 10.2139/ssrn.4814247.
26. G. F. Vergara *et al.*, "Stego-STFAN: A novel neural network for video steganography," *Computers*, vol. 13, no. 7, July 2024, Art. no. 180, doi: 10.3390/computers13070180.

27. F. Aiman and G. R. Manjula, "Video steganography using convolutional neural network and temporal residual method," *International Journal of Computer Applications*, vol. 178, no. 46, pp. 24–29, 2019, doi: 10.5120/ijca2019919375.
28. X. Duan, N. Liu, M. Gou, W. Wang, and C. Qin, "SteganoCNN: Image steganography with generalization ability based on convolutional neural network," *Entropy*, vol. 22, no. 10, Oct. 2020, Art. no. 1140, doi: 10.3390/e22101140.
29. H. Saidi, O. Tibermacine, and A. Elhadad, "High-capacity data hiding for medical images based on the mask-RCNN model," *Scientific Reports*, vol. 14, no. 1, Mar. 2024, Art. no. 7166, doi: 10.1038/s41598-024-55639-9.
30. E. G. Awad, P. Over, J. Fiscus, and A. F. Smeaton, "TRECVID 2017 Multimedia Event Detection (MED) Task Guidelines," National Institute of Standards and Technology (NIST), 2017. [Online]. Available: https://www-nlpir.nist.gov/projects/tv2017/Tasks/med/
31. K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild," Center for Research in Computer Vision, University of Central Florida, Dec. 2012. [Online]. Available: https://www.crcv.ucf.edu/data/UCF101.php.
32. T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Vimeo-90k," vimeo, Feb. 2019. [Online]. Available: http://toflow.csail.mit.edu/.
33. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, J., "ImageNet: A Large-Scale Hierarchical Image Database," Stanford Vision Lab, 2009. [Online]. Available: https://image-net.org/download.php.
34. R. Timofit, E. Agustsson, S. Gu, J. Wu, A. Ignatov, and L. V. Gool, "DIV2K: Diverse 2K Resolution Image Dataset," CVPR Workshops, July 2017. [Online]. Available: https://data.vision.ee.ethz.ch/cvl/DIV2K/.
35. T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," COCO Dataset, 2014. [Online]. Available: https://cocodataset.org/#download.
36. F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross and A. Sorkine-Hornung, "DAVIS: Densely Annotated VIdeo Segmentation Dataset," DAVIS, 2016. [Online]. Available: https://davischallenge.org/davis2016/code.html.
37. A. Mercat, M. Viitanen, and J. Vanne, "UVG Dataset," Ultra Video Group, June 2020. [Online]. Available: https://ultravideo.fi/dataset.html.
38. F. Pibre, J. Pasquet, D. Hillé, and M. Chaumont, "BOSSBase v1.01," BOSS: Benchmarking for Steganalysis, 2011. [Online]. Available: http://agents.fel.cvut.cz/stegodata/BOSSBase_1.01.zip.
39. N. F. Sahib and A. Y. Al-Sultan, "Diagnosis of COVID-19 in CT images based on convolutional neural network (CNN)," in *Proc. 1ST Samarra Int. Conf. for Pure and Applied Science*, *(SICPS2021)*, Samarra, Iraq, doi: 10.1063/5.0121126.
40. S. Ahmed, A. Ali, and E. Naser, "Tesseract OpenCV versus CNN: A comparative study on the recognition of unified modern Iraqi license plates," *Revue d'Intelligence Artificielle*, vol. 37, no. 5, pp. 1331–1339, Oct. 2023, doi: 10.18280/ria.370526.
41. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, Munich, Germany, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
42. V. Himthani, V. S. Dhaka, M. Kaur, G. Rani, M. Oza, and H.-N. Lee, "Comparative performance assessment of deep learning based image steganography techniques," *Scientific Reports*, vol. 12, no. 1, Oct. 2022, Art. no. 16895, doi: 10.1038/s41598-022-17362-1.
43. C. Ofoegbu, E. C. Nwokorie, O. U. Chinyere, and O. S. A. Amadi, "An improved image steganography system using convolutional neural networks," *International Journal of Research Publication and Reviews*, vol. 4, no. 11, pp. 1700–1717, Nov. 2023.
44. H. Kaur and C. L. Chowdhary, "The role of deep learning innovations with CNNs and GANs in steganography," in *Enhancing Steganography through Deep Learning Approaches*, IGI Global, 2024, ch. 4, pp. 75–106. [Online], Available: https://www.igi-global.com/chapter/the-role-of-deep-learning-innovations-with-cnns-and-gans-in-steganography/361549.
45. I. Goodfellow *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.
46. J. Liu *et al.*, "Recent advances of image steganography with generative adversarial networks," *IEEE Access*, vol. 8, pp. 60575–60597, Mar. 2020, doi: 10.1109/ACCESS.2020.2983175.
47. M. K. Oudah, A. N. Abed, R. S.Khudhair and S. M. Kaleefah, "Improvement of image steganography using discrete wavelet transform," *Engineering and Technology Journal*, vol. 38, no. 1, pp. 83–87, Jan. 2020, doi: https://doi.org/10.30684/etj.v39i1B.1574.
48. A. M. Adeshina, S. F. Abdul Razak, S. Yogarayan, and S. Sayeed, "Measuring fidelity of steganography approach in securing clinical data sharing platform using peak signal to noise ratio (PSNR) and structural similarity index measure (SSIM) related works," *International Journal of Computing and Informatics*, vol. 49, no. 11, pp. 45–56, 2025, doi: 10.31449/inf.v49i11.5661.

49. K. N. Hussin, A. K. Nahar, and H. K. Khleaf, "A visual enhancement quality of digital medical image based on bat optimization," *Engineering and Technology Journal*, vol. 39, no. 10, pp. 1550–1570, Oct. 2021, doi: 10.30684/etj.v39i10.2165.
50. S. A. Mahdi and M. A. Khodher, "An improved method for combine (LSB and MSB) based on color image RGB," *Engineering and Technology Journal*, vol. 39, no. 1, pp. 231–242, Mar. 2021, doi: 10.30684/etj.v39i1b.1574.
51. J. Kunhoth, N. Subramanian, S. Al-Maadeed, and A. Bouridane, "Video steganography: Recent advances and challenges," *Multimedia Tools and Applications*, vol. 82, no. 27, pp. 41943–41985, April 2023, doi: 10.1007/s11042-023-14844-w.
52. N. Wu, X. Chen, J. M. Adeke, and J. Zhao, "A cover-independent deep image hiding method based on domain attention mechanism," *Computers, Materials and Continua*, vol. 78, no. 3, pp. 3001–3019, Mar. 2024, doi: 10.32604/cmc.2023.045311.
53. P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," *IEEE Access*, vol. 9, pp. 108069–108082, 2021, doi: 10.1109/ACCESS.2021.3100462.