

Review Article

AI-Powered Steganography: Advances in Image, Linguistic, and 3D Mesh Data Hiding – A Survey

De Rosal Ignatius Moses Setiadi ^{1,*}, Sudipta Kr Ghosal ², and Aditya Kumar Sahu ³

¹ Faculty of Computer Science, Dian Nuswantoro University, Semarang, Central Java 50131, Indonesia; e-mail : moses@dsn.dinus.ac.id

² Department of Cyber Forensics and Information Security, Behala Government Polytechnic, 756, Upendra Nath Banerjee Road, Parnasree, Behala, Kolkata-700060, India; e-mail : sudipta.ghosal@gmail.com

³ Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amaravati, Andhra Pradesh 522503, India; e-mail : adityasahu.cse@gmail.com

* Corresponding Author : De Rosal Ignatius Moses Setiadi

Abstract: The rapid evolution of artificial intelligence (AI) has significantly transformed the field of steganography, extending its scope beyond conventional image-based techniques to novel domains such as linguistic and 3D mesh data hiding. This review presents a concise, accessible, and critical examination of recent AI-powered steganography methods, focusing on three distinct modalities: image, linguistic, and 3D mesh. Unlike most surveys focusing solely on one modality, this work highlights some modalities, identifies their unique challenges, and discusses how AI has reshaped embedding mechanisms, evaluation strategies, and security concerns. In image-based steganography, deep models such as GANs and Transformers have improved imperceptibility and extraction accuracy, but face limitations in computational efficiency and extraction consistency. Linguistic steganography, previously hindered by semantic fragility, has been revitalized by large language models (LLMs), enabling context-aware and reversible embedding, though still constrained by metric standardization and synchronization issues. Meanwhile, 3D mesh steganography remains dominated by non-AI methods, offering fertile ground for innovation through geometric deep learning. This review also provides a comparative summary of design principles, performance metrics, and modality-specific trade-offs. The analysis reveals a shift in evaluation paradigms, from numeric fidelity (e.g., PSNR, SSIM) to semantic and perceptual metrics (e.g., LPIPS, BERTScore, Hausdorff Distance). Looking ahead, future directions include cross-modal integration, domain adaptation, lightweight AI models, and the development of unified benchmarks. By presenting recent advances and critical perspectives across underexplored domains, this survey aims to inspire early-stage researchers and practitioners to explore new frontiers of steganography in the AI era.

Keywords: 3D mesh steganography; AI-based steganography; Application of steganography; Critical review of steganography; Generative steganography; NLP steganography; Steganography quality assessment.

Received: February, 25th 2025

Revised: March, 26th 2025

Accepted: March, 31st 2025

Published: April, 4th 2025

Curr. Ver.: April, 4th 2025



Copyright: © 2025 by the authors.
Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>)

1. Introduction

Steganography refers to the art and science of concealing information within an innocuous carrier medium, such as images, audio, video, or text, in a way that conceals the very existence of the hidden data[1]. Unlike cryptography, which obscures the content of a message, steganography aims to obfuscate its existence altogether. This key distinction leads to differing application goals: cryptography prioritizes confidentiality and robustness under exposure, whereas steganography generally emphasizes imperceptibility and undetectability from steganalysis. Watermarking is closely related to steganography, usually focusing on verifying ownership and authenticity[2]. Simply put, watermarking is an embedding that aims to protect the cover with an extraction scheme that can be done blindly or non-blindly. In

contrast to steganography, which functions to protect embedded messages. In terms of aspects, watermarking tends to be robustness as its main metric[1], [3].

In fact, in the development of digital steganography, images have become the most dominant cover media compared to other media, due to the rich redundancy and tolerance of image data perception. Early methods, such as Least Significant Bit (LSB) embedding, utilized pixel-level modifications in the spatial domain, while transform domain approaches (e.g., Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT)) provided better resistance to compression and attacks. Traditional steganography techniques (here referring to early digital steganography before the use of AI/ML methods) are usually classified into spatial domain, transformation domain, statistical, and adaptive schemes[4]. Steganography often operates under the constraint of blind extraction, that is, the ability to recover the hidden message without access to the original cover image.

Traditional image steganography in the spatial domain is much more developed because spatial domain methods, such as LSB, continue to be developed into LSB matching and marching re-visited, and combined with edge detection techniques[5], [6]. Other methods such as Pixel Value Differencing (PVD) are also becoming popular due to their imperceptibility and larger payload[7]. Transform domain-based methods are more widely used when robust steganography is needed against noise or compression. While adaptive steganography is developed based on the cost function to produce secure steganography, which is more resistant to steganalysis attacks such as the pioneering HUGO method[8] followed by HILL[9], WOW[10], MiPOD[11], and MaxMiPOD[12] as one of the latest secure adaptive methods.

There are three core aspects that define steganography performance, namely imperceptibility, which is how indistinguishable the stego object is from the original object, payload capacity, which is the amount of information that can be embedded, and security, which is the inability to be detected by steganalysis, and there is an optional aspect, namely robustness, which means that steganography is resistant to noise, compression or other manipulations[3], [13]. Depending on the application, robustness may be optional or even undesirable, especially when increasing robustness raises suspicion or decreases invisibility. The robustness aspect is used in special cases, namely transmission over noisy or compressed channels[1]. These aspects are inherently trade-offed, e.g., increasing payload may degrade imperceptibility, or improving robustness may sacrifice capacity. Finally, the approach of the steganography method has different emphasis on aspects to make steganography more superior for certain cases, more detailed research [1] classifies steganography as presented in Figure 1.

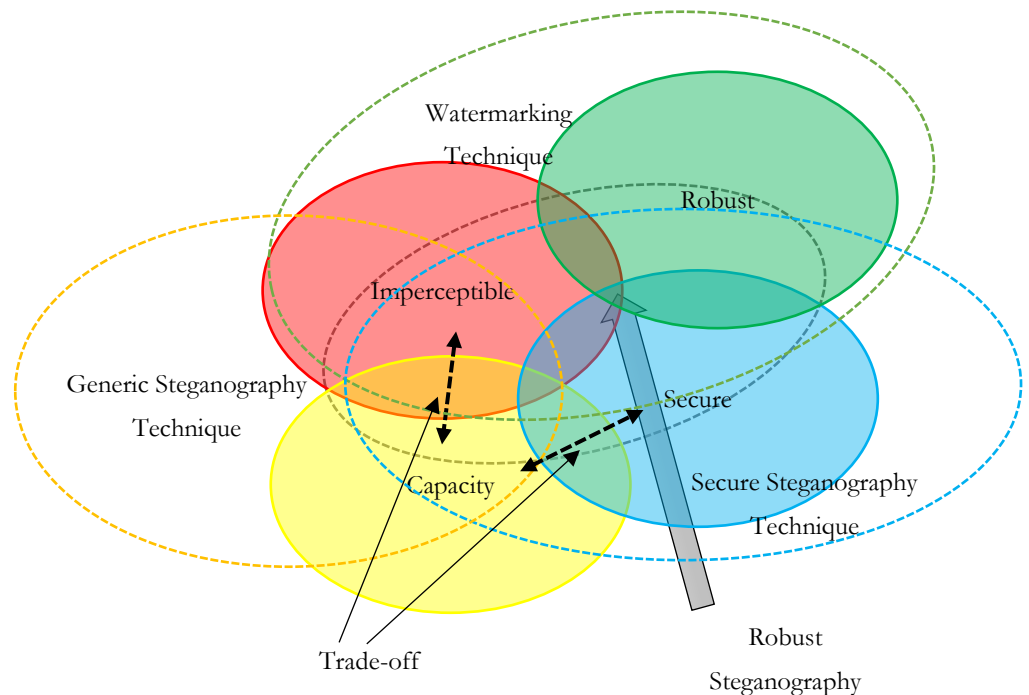


Figure 1. Classification and trade-off aspects of steganography[1].

In parallel, steganalysis has evolved as an adversarial field, developing increasingly powerful techniques to detect hidden messages. This has given rise to a continuous arms race between steganographers and steganalysts[14], [15]. Traditional steganalysis often relies on handcrafted features and statistical analysis; however, the emergence of machine learning (ML) and more recently deep learning (DL) has drastically changed the landscape, enabling both more secure embedding and more powerful detection[16].

The integration of AI into steganography has introduced a paradigm shift. From early ML-based feature selection to end-to-end deep neural networks, generative adversarial networks (GANs), cover selection, and transformer architectures, AI has enabled higher adaptability, semantic embedding, and improved imperceptibility[16]–[18]. This progression has also shifted the focus of steganographic systems from deterministic rules to data-driven learning, resulting in new challenges, such as imperfect extraction, uncertainty in embedding robustness, and lack of explainability.

While image steganography remains the most extensively studied modality, other domains such as text and 3D mesh offer unique challenges and emerging opportunities, especially in the context of AI. This paper focuses on three modalities—image-based, linguistic, and 3D mesh steganography—for the following reasons:

- First, although video and audio are often cited as alternative carriers, many steganographic methods for video merely apply traditional image-based embedding on selected frames or distribute the payload across multiple frames[19]–[21]. As such, they often inherit the principles and limitations of image steganography, rather than introducing fundamentally different mechanisms. Audio steganography, while structurally one-dimensional and conceptually similar to image-based methods, faces significant challenges in terms of imperceptibility and payload capacity due to its lower redundancy and higher temporal sensitivity[22], [23].
- Second, linguistic steganography represents a distinctive domain. Early text-based methods relied on character-level manipulations such as homoglyph substitution or font-based encoding[24], [25]. However, natural language exhibits low tolerance to perturbation—altering even a single word can drastically change meaning or fluency. Techniques such as synonym substitution and syntax transformation were introduced to address this, but they remained brittle and difficult to scale. The rise of natural language processing (NLP) and large language models (LLMs), however, has revitalized this field, enabling text-generation approaches that embed information while preserving semantic and syntactic coherence[26], [27].
- Third, 3D mesh steganography is a promising but underexplored area. Unlike 2D images or sequential text, 3D models possess spatial, topological, and geometric attributes such as vertex coordinates, face connectivity, and surface normals. These properties make embedding more complex but also open new opportunities for imperceptible and high-capacity hiding[28], [29]. Most methods remain handcrafted or statistical in nature, and the application of AI in this domain is still nascent, suggesting room for growth and innovation.

While previous surveys often concentrate on a single modality, this review intentionally provides a tri-modal perspective, i.e. image, text, and 3D mesh to introduce readers to the diverse techniques and challenges across domains. The goal is to offer a concise, accessible, and critical review that not only surveys recent AI-based advances (especially within the last three years), but also inspires early-stage researchers to consider alternative modalities beyond traditional image-based approaches.

Moreover, this paper emphasizes the importance of evaluation metrics, which differ across modalities and objectives. Metrics such as PSNR and SSIM are commonly used in image steganography, yet may not reflect human perceptual quality in AI-generated media. Text-based methods require fluency and semantic preservation, measured by perplexity (PPL), BLEU, or human judgments. 3D mesh methods, on the other hand, rely on spatial distortion measures like Hausdorff distance and coordinate displacement. Thus, this paper includes a comparative discussion of traditional and modern evaluation metrics, aiming to highlight their suitability for various steganographic tasks, particularly in AI-powered frameworks.

By bridging these three modalities and contextualizing their evolution under the influence of deep learning, this paper contributes a structured yet practical reference for understanding and advancing modern steganography across media types. By bridging these three

modalities and contextualizing their evolution under the influence of deep learning, this paper contributes a structured yet practical reference for understanding and advancing modern steganography across media types. The remainder of this paper is organized as follows: Section 2 explores advances in image-based steganography, including CNN, GAN, and Transformer-based approaches; Section 3 discusses linguistic steganography and how LLMs reshape its landscape; Section 4 presents recent developments in 3D mesh steganography and reversible data hiding; Section 5 provides a comparative synthesis, metric analysis, and discussion of future research directions; Section 6 examines the ethical concerns and misuse potential of AI-powered steganography; finally, Section 7 concludes the paper.

2. Image-Based AI Steganography

2.1. Preliminary

Traditional image steganography is generally performed by subtly modifying the pixels or frequency coefficients of an image to embed a secret message. Techniques such as LSB, DCT, and DWT are widely used due to their simplicity and efficiency on computationally limited systems. However, these approaches have limitations in terms of robustness to steganalysis, payload capacity, and adaptability to content variations. With the emergence of artificial intelligence, especially machine learning and deep learning, image steganography has undergone a paradigm shift. In this context, AI is used to select the optimal embedding area, disguise the secret message in the latent feature representation, and optimize the end-to-end embedding and extraction process. Models such as CNN, autoencoder, GAN, and transformer have been applied to improve the visual quality of stego-images and enhance the accuracy of information extraction[30]–[33].

Cover selection[34]–[37] is an AI approach to select the statistically most suitable container image for embedding, while GAN-based steganography [32], [38], [39], which uses adversarial networks to simulate the embedding and detection processes simultaneously, still fall into the steganography category because there is a process of embedding data into the container media. Meanwhile, Convolutional Neural Network (CNN)-based approaches work by learning the spatial representation of the container image. CNNs are used in both the embedding (by the encoder) and decoding (by the decoder) processes by utilizing cascaded convolutions to extract important features from the image[40]. Models such as HiDDeN[41] and SteganoGAN [42] use CNNs to ensure that hidden messages can be embedded without disturbing the visual structure of the image and can still be retrieved with high accuracy.

Different from CNN, Transformer-based approaches such as StegoFormer [31] or CAISFormer [43] utilize a self-attention mechanism to capture global relationships between pixels, allowing the system to be more adaptive in determining embedding locations and controlling modification intensity. The holistic feature transformation makes these approaches superior in undetectability and imperceptibility enhancement, especially in the face of CNN-based steganalysis. However, it should be emphasized that there is a shift in the context of imperceptibility because, in traditional steganography, imperceptibility is evaluated using numerical metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). However, in AI-based methods that rely on latent representation and semantic features, changes in the image are not always detected by these metrics. In contrast, Learned Perceptual Image Patch Similarity (LPIPS) is more relevant because it assesses perceptual similarity based on visual features learned by neural networks, approximating human perception[30]

Other AI-based approaches have also emerged, such as coverless steganography and steganography without embedding (SWE), which store messages in the form of an index or mapping to a collection of images or synthetic images that are automatically generated. In this approach, there is no direct embedding process into the digital media, but the hidden data is represented through content matching or image retrieval and hashing-based retrieval [39], [44]. These techniques generally require large image collections and efficient mapping processes to approach the capacity of traditional steganography. Meanwhile, SWE is a more radical approach, where there is no modification process to the carrier media. Instead of embedding a message into an image, the SWE method uses a generative model to generate an image that semantically or statistically represents the message. One SWE approach is carrier-secret entanglement, where the message is encoded in latent space and can only be interpreted

correctly if the reader has the appropriate generating model. Some SWE approaches also use techniques such as image hashing, bag-of-visual-words, or fingerprinting to map information without pixel modification[45], [46]. So even though the term “steganography” is used in both approaches, technically they do not involve an explicit embedding process, and in many cases the approach is closer to a retrieval system, image synthesis, or entanglement representation than steganography in the classical sense.

Therefore, in this review, we limit the discussion to techniques that contain explicit embedding processes, such as CNN, GAN, and Transformer-based models, which still operate within the framework of conventional steganography but are enhanced by the adaptive and representational capabilities of AI. With the presence of AI, there is a transformation in the performance characteristics of steganography systems. Important aspects in the context of adaptability to content and resistance to steganalysis are significantly improved. However, many AI-based methods do not guarantee 100% extraction accuracy. For example, SteganoGAN and StegoFormer achieve extraction accuracy of around 98–99%, depending on the payload used [31], [42], while the INN-GAN model shows a performance close to 100% without error correction on all tested extraction structures[32].

2.2. State-of-the-art of Image-based AI Steganography

Several recent studies on AI-based image steganography have been conducted. For instance, Wang et al. [32] proposed an image steganography architecture based on Invertible Neural Networks (INN) and GAN, consisting of three main structures: INN-GAN(D), INN-GAN(S), and INN-GAN(US), each employing DenseNet, shared weights, and unshared weights during the extraction phase. The system uses DWT transformation to embed messages in the frequency domain, achieving nearly 100% extraction accuracy without the need for error correction codes. The datasets used include COCO (train/test), VOC, and ImageNet, with image sizes of 256×256 and a fixed embedding rate (ER) of 1 bpp. Experimental results show that INN-GAN(S) achieved PSNR up to 55.16 dB (ImageNet), SSIM up to 0.9991, and Perceptual Loss (LP) of only 0.1976, outperforming methods like SteganoGAN and CHAT-GAN. Steganalysis tests using SRM, ZhuNet, YeNet, and SRNet showed better resistance, especially for INN-GAN(D).

Duan et al. [33] proposed DHU-Net, an image steganography model for high-capacity lossless binary data hiding, based on an enhanced U-Net architecture and flow encoding with BCH error-correcting code. The model introduces binary-to-spatial domain mapping and channel separation optimization to increase capacity without sacrificing visual quality. Evaluation was conducted on ImageNet, CelebA, Div2K, and MS-COCO, achieving PSNR up to 51.72 dB, SSIM 0.999, and 0% error rate after correction using BCH. For robustness evaluation, the model was tested against JPEG compression, Gaussian noise, and blur, showing excellent resilience with a 0% error rate after correction. Metrics used include PSNR, SSIM, and Pe, and steganalysis tests using SRNet, YeNet, XuNet, and StegExpose showed that DHU-Net's ROC curve approached the ideal random line more closely than other models. A critical analysis reveals that DHU-Net successfully combines high capacity, lossless extraction, and attack resilience in one system. However, the architecture's complexity (35M parameters) and high computation load (GFLOPS >17) may limit its deployment on constrained or real-time devices.

Yao et al. [38] proposed DWT-GAN, a GAN-based steganography model that integrates DWT transformation and multi-scale attention convolution modules to improve imperceptibility and message extraction accuracy. This method embeds messages in DWT sub-bands and adaptively controls modification strength on each DWT coefficient using a DNN-based fusion module. Evaluation of COCO, VOC, and ImageNet datasets shows embedding capacity up to 3 bpp, PSNR up to 52.88 dB, SSIM 0.9992, and extraction accuracy (ACC) of 0.994 for 1 bpp payload. For imperceptibility metrics, PSNR, SSIM, and LP were used, while ZhuNet and YeNet served as steganalysis detectors. Compared to HiDDeN, SteganoGAN, and CHAT-GAN, DWT-GAN achieved PSNR improvements of up to 6.65 dB and better steganalysis resistance. A critical analysis shows that frequency domain embedding reduces distortion in the LL sub-band. However, despite the significant improvement in imperceptibility, the model has not yet achieved 100% perfect extraction for high payloads and has a considerable model complexity is 272k parameters, challenging its use in limited-resource applications.

Li et al. [31] proposed StegoFormer, a transformer and CNN-based steganography architecture that integrates a U-shaped network and Swin Transformer block in the encoder to effectively capture local and global information. The architecture is equipped with a decoder for data recovery and a XuNet-based discriminator for adversarial training. To address errors at the center of the stego image, Shifted Window Local Loss (SWLL) and Gaussian Mask Augmentation (GMA) were introduced to strengthen the discriminator. The evaluation was conducted on the BOSSBase[47] dataset with payloads ranging from 0.1–0.5 bpp. StegoFormer achieved top results in PSNR up to 44.35 dB, SSIM is 0.9942, and MS-SSIM is 0.9991, with bitwise error rates as low as 0.00003–0.000056. Security testing using SRNet, YeNet, maxSRMd2, and CSR showed detection error rates as high as 57.67%, indicating better stealth capability.

Zhou et al. [43] proposed CAISFormer, a Transformer-based steganography model with channel-wise attention. The method introduces four main modules: Channel Self-Attention Module (CSAM), Non-Linear Enhancement (NLE), Channel-wise Cross Attention Module (CCAM), and Global-Local Aggregation Module (GLAM). CSAM and CCAM help select key features for embedding and enhance the correlation between cover and secret images. The datasets used include COCO and BOSSBase[47]. Results show an increase in PSNR by more than 0.36 dB (cover-stego) and 0.90 dB (secret-recovery) over previous methods, with a detection ratio drop of 3.43%. Embedding ratio claims reach up to 24 bpp. Imperceptibility was evaluated using PSNR, SSIM, and RMSE, while steganalysis used SRNet and ZhuNet.

Xiao et al. [30] proposed a Transformer-based steganography framework that integrates attention mechanisms and a WGAN adversarial network to improve imperceptibility and robustness. The model includes an encoder with ViT for feature extraction and a decoder with a spatial transformation network to preserve spatial relationships and mitigate positional noise. Datasets used include ImageNet, COCO, and Visual Genome, with random binary strings as input. Evaluation results show PSNR of 35.53, MS-SSIM of 0.958, and LPIPS of 0.048 on ImageNet, with extraction accuracy up to 98.99%. The model maintained an accuracy of >60% after rotation, cropping, and distortion. Random masking during training greatly contributed to model robustness. From a critical view, its main strength lies in high generalization and ZhuNet steganalysis resistance, although it demands considerable computational resources during training.

Li et al. [48] proposed ISteg, a steganography strategy based on an artificial immune system (AIS) designed to enhance stego-image security against conventional and deep learning-based steganalyzers. Instead of replacing the embedding process, ISteg works as a post-processing stage by optimizing the already generated stego-images through existing embedding methods such as HILL, WOW, S-UNIWARD, MIPOD, and UT-GAN. The datasets used include BOSSBase[47], BOWS2[49], and ALASKA2[50], with payloads up to 0.5 bpp. Evaluations using SRM, maxSRMd2, and Zhu-Net show that ISteg increased security, with HILL_ISteg gaining a 3.82% improvement against Zhu-Net at a 0.4 bpp payload. Metrics used include PSNR and SSIM for imperceptibility and detection error rate (PE) for security. Results indicate that ISteg improves security without compromising imperceptibility with average PSNR > 57 dB, and SSIM > 0.9995. Critically, ISteg is highly effective as a universal framework applicable to various embedding methods but introduces additional computational overhead, potentially challenging for real-time applications. The approach stands out due to its post-embedding optimization focus, rarely explored in AI-based steganography literature.

In the following year, Li et al. [51] proposed the Immucover method, which focuses on enhancing the cover image through fuzzy enhancement and the Artificial Immune System (AIS). Immucover consists of three main stages: Texture Region Enhancement (TRE), Edge Detection and Enhancement (EDE), and parameter optimization using AIS. Unlike previous approaches that focus on the embedding process, Immucover optimizes the cover quality to reduce pixel anomalies and improve resistance against steganalysis detection. The dataset used is the same as in the previous study [48], and the method was evaluated using six embedding algorithms, including HILL, WOW, S-UNIWARD, MIPOD, UT-GAN, and SPAR-RL. Security evaluation was conducted using five leading steganalyzers: SRM, maxSRMd2, Xu-Net, and Zhu-Net, with the main metric being PE. The results show that embedding into the immunized cover can increase PE by up to 11.5% compared to conventional methods, without reducing imperceptibility (average PSNR > 45 dB, SSIM > 0.985). Immucover also outperforms other cover enhancement methods such as CAAS and ADS. From a critical

perspective, this approach introduces a new direction in steganography by prioritizing cover quality before embedding, although computational overhead due to AIS optimization remains a potential challenge.

Chen et al. [52] proposed a Steganography Immunoprocessing Framework based on AIS to enhance the security of steganography against attacks from both CNN-based and handcrafted feature-based steganalysis. Unlike ISteg [48], which optimizes a single stego image, this approach is universal and applies to all images generated by the same steganographic process. The immunization process is carried out through dynamic antibody-based optimization, which considers three main characteristics: the gradient from the backpropagation of the CNN steganalyzer, the pixel modification rate from the embedding process, and texture complexity. The datasets used include BOSSBase[47] v1.01, BOWS2[49], and ALASKA2[50], while the base embedding methods include S-UNIWARD, WOW, HILL, MiPOD, and SPAR-RL. The evaluation was conducted using steganalyzers SRM, maxSRMd2, and ZhuNet, with PE (detection error rate) as the main metric, and PSNR and SSIM for imperceptibility. Experimental results showed an increase in PE of up to 11.5% at a payload of 0.5 bpp, with average visual quality exceeding 55 dB PSNR and 0.99 SSIM. Critically, this framework stands out for strengthening stego image resistance against both CNN-based and handcrafted steganalyzers simultaneously, although challenges remain in terms of computational time and optimal parameterization.

2.3. Finding and Critical Analysis

Based on a review of recent AI-based steganography methods, it can be concluded that state-of-the-art developments can be categorized into several prominent focus areas:

- First, several approaches such as ISteg, Immucover, and the Steganography Immunoprocessing Framework explicitly target security against steganalysis, both handcrafted feature-based (SRM, maxSRMd2) and CNN-based (ZhuNet, YeNet, SRNet). Immunization strategies based on the Artificial Immune System (AIS) have proven effective in increasing the detection error rate (PE) by up to 11.5% without degrading visual quality. This approach is appealing because it functions as a universal post- or pre-processing step, independent of the embedding method used, although trade-offs arise in the form of computational overhead and the complexity of dynamic parameterization.
- Second, imperceptibility improvements are achieved through the integration of deep learning architectures, especially Transformers and attention mechanisms. For example, CAISFormer, StegoFormer, and TransStego demonstrate high PSNR and SSIM results, with additional perceptual metrics such as LPIPS used to assess visual quality from the human perception perspective. This approach shifts the evaluation paradigm from classic numerical evaluation (e.g., PSNR) toward perceptual fidelity, aligning with trends in generative modeling and adversarial training. However, the use of LPIPS in some studies is still inconsistent with its official implementation (e.g., merely based on MSE between feature maps), so the validity of perceptual claims needs to be analyzed more thoroughly.
- Third, some papers present claims of extremely high payload capacity, even up to 24 bpp for RGB images or 8 bpp for grayscale images, as claimed by CAISFormer and some Transformer-based variants. In theory, this approaches the information channel capacity limits of digital media and appears less realistic for practical scenarios without sacrificing imperceptibility or security. In practice, other methods such as DWT-GAN and INN-GAN targeting 1–3 bpp already show challenges in maintaining visual fidelity and extraction accuracy. This indicates that the classic trade-off between payload, imperceptibility, and security remains an unresolved challenge.

An important finding to highlight is in the evaluation of imperceptibility. Conventional metrics such as PSNR and SSIM are still widely used to assess the visual quality of stego images. PSNR measures the mean squared error between pixels, while SSIM evaluates structural similarity based on luminance, contrast, and local texture. However, both have fundamental limitations as they do not fully reflect human visual perception. Images with high PSNR values may still exhibit noticeable visual artifacts[53], and SSIM remains sensitive to small spatial distortions and lacks sensitivity to semantic differences. In response to these limitations, modern metrics such as Learned Perceptual Image Patch Similarity (LPIPS) and Fréchet Inception Distance (FID) are strongly recommended for future research to evaluate imperceptibility based on human perception and semantic distribution. LPIPS measures the

similarity between two images based on feature maps extracted from pre-trained CNNs such as VGG or AlexNet. Unlike PSNR, which only compares pixels, LPIPS assesses differences at the feature level and weights them using trained parameters aligned with human visual perception. Therefore, LPIPS can capture semantic differences and local changes that may not be visible to pixel-based metrics. In several studies [32], [38], LPIPS has been applied merely as a perceptual loss calculated through MSE between feature maps, without using the trained weights as specified in the official LPIPS version. This may result in inaccurate evaluations if not explicitly clarified. The LPIPS formula is shown in Equation (1), which illustrates how feature differences across layers are weighted and averaged to produce the final score.

$$\text{LPIPS}(x, y) = \sum_l \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \|w_l \odot (F_l^x[h, w] - F_l^y[h, w])\|_2^2 \quad (1)$$

Where x is the cover; y is the stego image, F_l^x and F_l^y are the feature map of the layer l in x and y ; w_l are the learned weights for layer l ; $H_l W_l$ are the height and width of the feature map in the layer l ; and $\odot$ is Hadamard product. The smaller the LPIPS value, the more similar the two images are perceptually.

On the other hand, FID measures the distance of feature distributions between stego images and cover images. Features are taken from the Inception-v3 network at a particular layer. FID calculates the Fréchet (Wasserstein-2) distance between two multivariate Gaussian distributions estimated from the features. FID is calculated by Equation (2).

$$\text{FID} = \|\mu_x - \mu_y\|_2^2 + \text{Tr}(\Sigma_x + \Sigma_y - 2(\Sigma_x \Sigma_y)^{1/2}) \quad (2)$$

Where μ_x, Σ_x are mean and covariance of cover features; μ_y, Σ_y are mean and covariance of stego features; Tr is trace operator; $(\Sigma_x \Sigma_y)^{1/2}$ is the matrix root of the product of two covariance matrices. FID takes into account the differences in the statistical distribution of features, not just a single pair of images, and is considered a strong indicator of image quality and realism.

Another important finding is the shift in the role of AI models, from being merely an embedding tool to an end-to-end steganography system designer that combines embedding, detection avoidance, and robustness to disturbances (rotation, blur, noise, etc.). Models such as TransStego introduce generalization capabilities and robustness to geometric transformations without fine-tuning, indicating a new direction in the development of adaptive and contextual steganography. Overall, although significant progress has been made in terms of extraction accuracy and resistance to steganalysis, challenges remain in the form of architectural complexity, high computational dependency, and non-uniform standardization of measurement tools across studies. Thus, future studies need to prioritize the development of methods that are not only technically superior, but also lightweight, scalable, and able to withstand real-world attacks, while adopting evaluation metrics that more accurately reflect human perception.

3. Linguistic Steganography

3.1. Preliminary

Linguistic steganography is a branch of steganography that conceals secret information within natural language, fundamentally different from image steganography which manipulates pixels or frequency transformations in the visual domain. The primary focus of linguistic steganography is to maintain the fluency and semantic coherence of the text, while reliably embedding and extracting messages without disrupting the naturalness of the language. Technically, linguistic approaches cannot rely on visual metrics such as PSNR, SSIM, LPIPS, or FID to assess imperceptibility; instead, they depend on linguistic metrics and even human evaluation to judge the quality and logic of the generated text[54], [55]. Linguistic approaches are more challenging due to the low redundancy of text and its higher sensitivity to alterations. As such, text steganography methods are developed with a focus on linguistic imperceptibility, including syntactic and semantic alignment rather than bit-level perturbation.

Historically, linguistic steganography began with primitive techniques such as invisible ink and typographic layout manipulations. In the digital era, it has evolved into three main categories: (1) Text Modification-Based Steganography (TMS), which modifies existing text (e.g., synonym substitution, punctuation, sentence structure); (2) Coverless Linguistic Steganography (CLS), which does not alter text but selects it from a large corpus based on the hidden message; and (3) Generative Linguistic Steganography (GLS), which generates text from scratch based on the secret message using NLP and DL techniques [54], [56].

Among the three, GLS has become the main focus of recent research due to its larger embedding capacity, the absence of a required cover, and its ability to generate new text that adaptively conceals the message[57]. GLS typically consists of two main processes [54], [56]:

- Text Generation: Utilizing large language models (LLMs) such as GPT-3 or BERT to generate natural language text.
- Encoding/Decoding: Embedding message bits into the text using algorithms like arithmetic coding, Huffman coding, or adaptive grouping that selects words from candidate lists based on probability distributions generated by the model.

GLS thus enables the embedding of secret bits into linguistically natural structures. Models such as RNN-Stega, VAE-Stega, and StegaText have explored various embedding schemes to enhance imperceptibility and resistance to steganalysis. Some important aspects highlighted in linguistic steganography include:

- Perceptual imperceptibility: The semantic readability and fluency of the text.
- Statistical imperceptibility: The probabilistic distribution similarity between the stego and original text.
- Cognitive imperceptibility: The contextual meaning preservation in specific scenarios.

Performance evaluation no longer relies solely on extraction accuracy but also employs NLP metrics such as perplexity (PPL), mutual information (MI), BLEU, ROUGE, and METEOR. With the advancement of NLP technology and generative models, linguistic steganography has become not only more sophisticated but also more difficult to detect, thus requiring increasingly complex and contextual evaluations. This approach opens new pathways for concealing information that is hidden not only from machines but also from human perception.

3.2. State-of-the-art Linguistic Steganography

One of the pioneering GLS projects that utilizes LLMs and has been implemented in real-time is the Stegasuras application introduced by Ziegler et al.[58], which can be accessed at <https://steganography.live/>. This system employs arithmetic coding combined with GPT-2 to embed secret messages directly into generated natural language text, without modifying existing text. Evaluation results show excellent performance in terms of efficiency (up to 3 bits per word/bpw), extremely low KL divergence, and high-quality language output based on human evaluation. Stegasuras is among the first generative GLS systems to be deployed interactively and in real-time (see Figure 2).

Xiang et al. [59] proposed Causal Perception Guided Linguistic Steganography (CPG-LS), which combines a BERT-based lexical substitution model with a causal perception network. The system identifies words in the text that can be substituted with minimal risk of detection using a causal score based on word saliency and replaceability. The dataset used consists of 10,000 texts from the CC-100 corpus[60]. Evaluation metrics include PPL, Mauve, and bpw for payload capacity, and three DNN-based steganalysis models: BiLSTM-D, R-BI-C, and TS-RNN. Results show that CPG-LS achieves fluency comparable to synonym substitution methods but with higher payload and improves anti-steganalysis capability by 10.1% compared to ref [61]–[63] as baselines.

Chang [64] proposed a reversible linguistic steganography system utilizing Bayesian Masked Language Modeling (BMLM). The system embeds messages by substituting words based on BERT model predictions, guided by prediction uncertainty to adaptively route embedding. Evaluation was conducted on ewight selected paragraphs from "Alice in Wonderland" classic English literature txt using bpw for payload, cosine similarity for imperceptibility, and flag bits for reversibility. Experiments show that this approach achieves perfect reversibility without added overhead while maintaining semantic and sentiment similarity between cover and stego texts. A critical analysis highlights that although the system maintains both imperceptibility and reversibility, the computational burden caused by Monte Carlo dropout

remains high and needs optimization for real-time implementation. This work is a pioneering contribution to reversible generative steganography based on prediction uncertainty.

Figure 2. Stegasuras interface (<https://steganography.live/>), an implementation of Ziegler et al.’s [58] neural linguistic steganography using arithmetic coding and GPT-2. Users can input a context, embed a secret message, and generate fluent cover text in real-time. The interface includes options such as a message encoding algorithm, steganography strategy, and language model temperature (τ), which controls output diversity. Although labeled “Encrypt/Decrypt”, the system performs steganography—concealing message presence rather than encrypting content.

Yang et al. [65] proposed a Semantic-Preserving Linguistic Steganography method based on Pivot Translation and Semantic-Aware Bins Coding (SaBins). This technique uses round-trip translation (English $\rightarrow$ German $\rightarrow$ English) to produce semantically preserved paraphrases and embeds secret information using semantic-based encoding. The evaluation was performed on public datasets featuring diverse text styles, using metrics such as PPL, semantic similarity (BLEU, METEOR), and detection accuracy by steganalysis. Results show significant improvement in payload capacity without sacrificing semantic fidelity, and strong resistance to detection. This method strikes a balance between the high capacity of GLS and the semantic preservation of MLS while reducing the dependency on shared models between sender and receiver.

Ding et al. [66] developed NMT-Stega, a linguistic steganography model based on Neural Machine Translation (NMT) with features such as Global Context Fusion and Language Model Reference (LMR). Using Patient-Arithmetic Encoding, the system improves embedding capacity over previous methods relying on conventional waiting mechanisms. Evaluation using BLEU, KL divergence, PPL, and three steganalysis models shows that the model

maintains both semantic and statistical imperceptibility. Compared to baselines, it achieves a higher ER while preserving text quality and resistance to detection. However, model complexity and the requirement for bilingual text pairs can pose challenges in monolingual or real-time scenarios.

Xiang et al. [67] introduced Hiding Information in Syntax Space (HISS-Stega), a framework that utilizes syntax-controlled paraphrasing and syntactic bins coding strategies. The model automatically generates syntactic variants of source sentences to embed data while maintaining semantic coherence. Benchmark tests show that HISS-Stega outperforms baselines in MAUVE and BERTScore, and achieves the lowest detection accuracy and F1 scores. With an ER reaching 0.3332 bpw, the system offers competitive payload capacity while remaining resistant to detection. However, its reliance on syntactic structures can limit applicability in informal or highly unstructured text domains.

Namitha et al. [68] introduced StegAbb, a GPT-3-based steganographic tool for covert communication between Software Defined Radios (SDRs). The method uses GPT-3 to generate text from encrypted binary acronyms via abbreviation expansion. Evaluation metrics include PPL, human opinion score, and decoding accuracy. Results show a perplexity as low as 2.75, a human rating of 4.052/5, and 100% decoding accuracy, with an embedding capacity of up to 4 bpw. The stego detection rate by external models is only 0.5245, close to the ideal value of 0.5. Although innovative and deployable in real systems, the method relies heavily on the GPT-3 temperature parameter and has not been tested on complex narrative text.

Ding et al. [69] contributed a recent advancement in GLS by combining conditional generation-based methods using LSTM and multi-head attention with substitution-based methods using BERT Masked Language Model. It also strengthens statistical feature consistency using Graph Attention Network (GAT) to capture spatial token relationships. The dataset is drawn from the OPUS corpus[70], trained in a PyTorch environment using a TITAN RTX GPU. Evaluation includes ER, PPL, KL divergence, and semantic similarity via SpaCy, as well as anti-steganalysis accuracy using three detectors: LS-CNN, TS-RNN, and BERT Classifier. Results show that this model outperforms previous methods in imperceptibility, achieving KL divergence ≈ 1.112 and PPL ≈ 2.864 at ER ≈ 1.98 bit/token, with lower detection accuracy by steganalysis models.

Qi et al. [71] introduced a provably secure approach that addresses segmentation ambiguity. This issue arises when detokenization and retokenization between sender and receiver produce mismatched token representations, causing message extraction failure. To resolve this, the authors proposed SyncPool, which creates an ambiguity pool based on token prefix relationships and uses a cryptographically secure pseudorandom number generator (CSPRNG) to synchronize token selection between sender and receiver. The method does not alter the original probability distribution of the language model, thus preserving theoretical security with KL divergence remaining zero. The evaluation was conducted using LLaMA2 and Baichuan2 models on the IMDb dataset in both English and Mandarin, using metrics such as PPL, KL divergence, utilization, and total error. Results show that SyncPool successfully eliminates extraction errors to 0%, without disturbing token distribution or triggering detection by distributional steganalysis. This method marks a significant step toward truly provably secure linguistic steganography based on large language models, although it still demands strict synchronization of random generators and introduces additional entropy management complexity.

3.3. Finding and Critical Analysis

Based on the review of recent GLS methods above, it can be concluded that linguistic steganography presents significantly greater challenges and complexity compared to image-based steganography. One of the key findings is the inherent trade-off between payload capacity, fluency, and distributional/statistical security. Models such as Stegasuras and StegAbb can achieve payloads of up to 3–4 bpw; however, at such high rates, fluency decreases and semantic quality begins to deteriorate. Conversely, models like CPG-LS and HISS-Stega prioritize fluency and detection resistance but must compromise on the number of bits that can be embedded.

Technically, GLS also introduces new complexities in the selection and evaluation of performance metrics, which cannot solely rely on extraction accuracy or traditional generative metrics. Some of the key linguistic metrics used in the literature include:

- Perplexity (PPL): Measures how "surprising" a text is to a language model; lower values indicate more natural text. It is suitable for evaluating fluency.
- BLEU, ROUGE, METEOR: N-gram-based similarity metrics for assessing semantic similarity between sentences. Commonly used for semantic fidelity evaluation.
- KL divergence: Measures the difference in token probability distributions between stego text and original text. This metric is crucial for assessing statistical imperceptibility.
- Cosine Similarity / BERTScore: Evaluates semantic similarity using sentence or word embedding vectors.
- MAUVE: A GAN- and mutual information-based metric that evaluates how closely the text distribution resembles that of a natural language model output.

Each metric serves a different context. PPL and MAUVE are best suited for evaluating fluency and probabilistic distribution, while BLEU and METEOR focus on semantic correctness. However, no single metric can comprehensively assess all aspects of linguistic imperceptibility, making a combination such as PPL, KL divergence, and BERTScore the most recommended for holistic evaluation.

Regarding extraction capability, not all GLS methods guarantee 100% extraction accuracy. Some approaches, such as Bayesian Masked Language Modeling, demonstrate perfect reversibility by accounting for prediction uncertainty. Others, like StegAbb and NMT-Stega, remain vulnerable to extraction failure due to tokenization ambiguity or unsynchronized sampling distributions. Provably secure approaches such as SyncPoo [49] theoretically ensure 100% message extraction without compromising distributional security, but require full synchronization between sender and receiver—including the use of the same pseudorandom generator and identical language models.

The critical conclusion from this study is that GLS provides flexible and adaptive message embedding within text, but demands compromises among naturalness, payload, and security. No single system yet achieves optimal performance across all three aspects. Additionally, challenges such as token ambiguity, lack of standardized evaluation metrics, and the infrastructure burden of large-scale models remain barriers to real-world adoption. Future research should focus on developing more representative and lightweight imperceptibility evaluation methods and exploring compact models for mobile or embedded applications.

4. 3D Mesh Steganography

4.1. Preliminary

A 3D mesh is a representation of a three-dimensional object in the form of a collection of vertices, edges, and faces, where the geometric and topological information of the object is explicitly structured. Unlike image steganography, which embeds information into a two-dimensional pixel matrix, and linguistic steganography, which operates on symbolic representations of natural language, 3D steganography involves manipulating the spatial structure of objects in 3D space. The unique characteristics of 3D data, such as vertex coordinates, face connectivity, mesh topology, and texture, make the embedding and extraction process far more complex, while also offering a more flexible and hidden space for data insertion [29], [72], [73].

3D mesh-based steganography first emerged in response to the need for secure transmission of 3D models, a demand that has rapidly grown with the advancement of CAD technology, virtual modeling, and 3D printing. The main motivation behind this approach is to embed additional information, such as copyright data, authentication markers, or secret messages, into 3D models without compromising their visual and geometric integrity [74], [75]. According to the literature, there are three primary embedding domains [29]:

- Geometrical domain: modifies vertex coordinates directly (e.g., LSB, PCA-based methods);
- Topological domain: embeds data through changes in vertex or face ordering (e.g., permutation-based techniques);
- Representation domain: leverages spectral transforms, wavelets, or complex structures like Laplacian meshes.

Each domain entails trade-offs between capacity, imperceptibility, and robustness. LSB-based methods, such as the Gaussian-Curvature-Constraint-Based Method proposed by [76] can achieve capacities of up to 93 bpv but are highly vulnerable to detection and noise. In contrast,

statistical or transform domain approaches like spectral mesh or wavelet offer greater resistance to smoothing and affine transformations, though at the cost of lower capacity and greater distortion [29] presents a simple illustration of 3D mesh steganography using Spatial LSB Substitution, which belongs to the Geometrical Domain category that modifies coordinate space. Figure 3 presents an illustration example of a simple 3D mesh steganography based on Spatial LSB Substitution which falls into the Geometrical Domain category which modifies the domain coordinates.

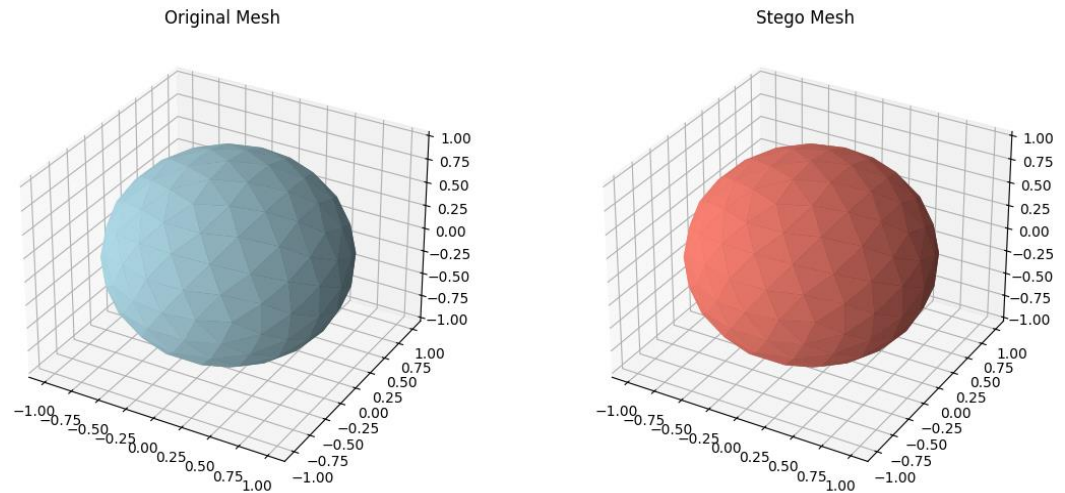


Figure 3. 3D mesh steganography based on Spatial LSB Substitution (Geometrical Domain).

Evaluation of the sample shown in Figure 3 indicates that 3D mesh steganography based on LSB embedding in vertex coordinates can embed binary messages with minimal geometric distortion. With a payload of 512 bits on an icosphere mesh comprising 162 vertices (approximately 3.16 bpv), this method yielded a very low MSE of 1.12×10^{-15} and a PSNR of 149.49 dB, suggesting almost negligible numerical difference. Hausdorff distance and average displacement per vertex were also in the range of 10^{-7} . Although the numerical evaluation suggests minimal geometric changes, visual differences were still observable due to vertex shifts affecting shading and lighting during rendering. This highlights that in 3D mesh, imperceptibility is not only dependent on geometric shape but also on visual effects arising from minor spatial adjustments.

As the need for more adaptive and attack-resistant systems grows, machine learning and deep learning-based approaches have begun to be adopted. Several AI and deep learning methods have been applied, including the use of autoencoders or GNNs (Graph Neural Networks) to support reversible embedding or optimal embedding location prediction. More explicitly, Graph Convolutional Networks (GCN)[77] have been used to enhance robustness against vertex reordering and noise. In general, the main challenges in 3D mesh steganography include:

- Preserving visual quality and topological structure, as excessive modifications can degrade the geometry object.
- Robustness against common transformations such as rotation, translation, mesh simplification, and smoothing.
- Reversibility for applications such as digital forensics or hidden metadata in CAD designs.

Evaluation in 3D steganography commonly uses the following metrics:

- Imperceptibility: using PSNR 3D, Hausdorff Distance, Root Mean Square Error (RMSE), and angle distortion.
- Robustness: assessed through transformations such as simplification, reordering, and smoothing.
- Capacity: measured in bits per vertex (bpv).
- Security: evaluated using steganalysis detection rates via classifiers such as SVM or ensemble methods against mesh features like YANG40 [78], LFS76[79].

Given the complexity and flexibility of 3D data structures, this field offers great potential but requires the development of highly contextual and domain-aware steganographic techniques. Future research challenges include optimizing adaptive embedding, defending against universal steganalysis, and integrating with multimodal systems.

4.2. State-of-the-art of 3D Mesh Steganography

The development of 3D mesh steganography is increasingly oriented towards reversible and encryption-based techniques to meet the demand for privacy protection and data integrity in 3D media. One of the main focuses is to increase embedding capacity without sacrificing visual quality or the accuracy of model recovery. For instance, Lyu et al. [80] proposed a Reversible Data Hiding (RDH) method for encrypted 3D mesh models based on multi-MSB prediction and entropy coding. The model divides the vertices into an embedded set and a prediction set based on index parity, then utilizes neighboring MSB prediction to define an adaptive embedding space. For space efficiency, label information is compressed using arithmetic coding. The dataset includes Beetle, Mushroom, Elephant, and Mannequin models from the Princeton Shape Benchmark (PSB)[81]. Evaluation metrics include ER, SNR, Hausdorff Distance, and runtime. Results show a maximum ER of 27.96 bpv on the Elephant model, with distortion-free model recovery and competitive execution time. Critical analysis highlights that although the method is efficient and reversible, its reliance on local predictive structure may limit performance on meshes with complex topology or high noise. Its primary advantage lies in the combination of efficiency, high capacity, and accurate model restoration, making it a strong candidate for cloud-secure 3D data sharing..

Zhang et al. [82] proposed MMPRDH-ED, a reversible data hiding method for encrypted 3D meshes that combines triangular mesh subdivision (TMS), multi-MSB prediction (MMP), and a multi-layer embedding strategy (MLES). The method applies the RRBE (reserving room before encryption) approach to increase the number of embeddable vertices. The dataset includes 3D models from PSB[81] and four specific models: Beetle, Cow, Lamp, and Vase. Experimental results show that MMPRDH-ED achieves an ER of up to 44.99 bpv, with an average of 39.27 bpv. Evaluation using Hausdorff Distance and SNR shows very low distortion and nearly perfect model recovery i.e., $\text{SNR} > 100$ dB for $n \geq 5$. The method also maintains separability between the extraction and recovery processes and uses stream encryption for efficiency. Critical analysis suggests that although it excels in capacity and reversibility, computational complexity increases due to mesh subdivision and complex vertex management. Furthermore, it relies heavily on full trust in cloud service providers if encryption is outsourced.

Zhang et al. [83] proposed an adaptive 3D mesh steganography algorithm based on Feature-Preserving Distortion (FPD), which incorporates embedding cost computation from steganalytic features to minimize detection. The algorithm defines an embedding domain via decimal-to-integer conversion and applies Q-layered Syndrome Trellis Code (STC) for embedding. The dataset includes PSB[81], Princeton ModelNet (PMN)[84], and The Stanford Models (TSM)[85]. Evaluation metrics include embedding capacity up to 6 bpv, imperceptibility using PSNR and visual difference, and security against six steganalyzers including LFS52, LFS64, and NVT+. Results show the method achieves the highest average out-of-bag (OOB) error of up to 30% even at large payloads. Its main strength lies in adaptivity and reduced embedding impact on geometric features, but it remains vulnerable to attacks like vertex reordering and smoothing due to the topological sensitivity of the embedding domain.

Wang et al. [86] proposed an RDH scheme for encrypted 3D mesh models using adaptive vertex grouping, multi-MSB prediction, and Huffman coding to increase embedding capacity. Two encryption schemes are introduced: stream cipher and secret sharing over Galois Field $\text{GF}(2^8)$ for application flexibility. The dataset comprises 100 models from PSB[81], including "Dolphin", tested with various scale values m . Experimental results show that ER reaches up to 50 bpv, with an average exceeding 35 bpv, outperforming previous methods. Evaluation includes ER, SNR, coordinate change rate (CCR), and model recovery accuracy. Strengths include high capacity and separability for data extraction and model recovery. Weaknesses include the need for additional metadata and potentially high computational load, especially with secret sharing in resource-constrained environments.

Gao et al. [87] proposed an RDH scheme for encrypted 3D mesh models using adaptive vertex grouping, multi-MSB prediction, and Huffman coding to increase embedding capacity.

Two encryption schemes are introduced: stream cipher and secret sharing over Galois Field $GF(2^8)$ for application flexibility. The dataset comprises 100 models from PSB[81], including "Dolphin", tested with various scale values m . Experimental results show that ER reaches up to 50 bpv, with an average exceeding 35 bpv, outperforming previous methods. Evaluation includes ER, SNR, coordinate change rate (CCR), and model recovery accuracy. Strengths include high capacity and separability for data extraction and model recovery. Weaknesses include the need for additional metadata and potentially high computational load, especially with secret sharing in resource-constrained environments.

Zhang et al. [74] proposed Multi-layer Synchronized RDH (MS-RDH), a spatial-domain RDH algorithm for 3D mesh models addressing mesh asynchrony—discrepancies in vertex traversal order during embedding and extraction. This method combines minimum distortion analysis, logistic mapping strategies, and interleaved bimodal mapping rules in a multilayer embedding scheme. The dataset includes six mesh models from PSB[81] such as Bunny, Fandisk, and Dragon. Experiments show SNR values above 92.15 dB even in worst-case scenarios, and ER reaching 8.83 bpv on models like Buddha, with low geometric distortion i.e., Hausdorff distance $< 10^{-5}$. MS-RDH's main strength lies in its combination of high capacity and strong visual imperceptibility while maintaining full reversibility. However, the method is non-blind and less robust against external attacks, thus requiring integration with additional security mechanisms or robustness enhancements.

Hou et al. [72] proposed a Separable Reversible Data Hiding method in Encrypted 3D Mesh Models based on octree spatial subdivision and multi-MSB prediction to enhance embedding capacity in encrypted domains. This approach partitions 3D models into adaptive sub-blocks using octree, then exploits spatial correlation within each sub-block for multi-MSB prediction. Data is embedded into encrypted parts via significant bit substitution. PSB[81] is used as the dataset, evaluated on models such as Bunny. Results show an average embedding capacity of up to 55.02 bpv and ER of 95%, with model recovery quality measured via SNR and Hausdorff Distance. The model is recoverable losslessly, and auxiliary data can be extracted without error, proving its reversible and separable nature. Strengths include high capacity, space efficiency, and topology independence, making it applicable to point clouds. However, performance depends on the choice of subdivision parameters and thresholds, which, if suboptimal, may degrade embedding quality.

Gao et al. [88] introduced a crypto-space steganography approach for encrypted 3D mesh models using innovative strategies based on greedy selection and shortest expansion. The aim is to increase vertex utilization ratio without additional metadata through efficient topology grouping. The method uses multi-MSB prediction and Huffman coding to compress vertex coordinates before encryption via stream cipher. The dataset is sourced from PSB[81], including models like Aircraft and Plane. Experimental results show an average ER of 49.39 bpv. Evaluation includes ER, vertex utilization ($\sim 90\%$), prediction accuracy, entropy, and number of coordinates changing rate (NCCR), with entropy nearing the maximum value of 4.32 and NCCR reaching 100%. The main strength is its high capacity and no need for extra metadata. The drawback lies in its vulnerability to common attacks such as rotation or mesh noise, as geometric robustness is not addressed.

The most recent work by Peng et al. [89] introduced MeshPAD, a novel approach based on geometric deep learning for payload-adaptive 3D mesh steganography. Unlike heuristic methods such as Vertex Normal-based Distortion (VND) and Gaussian Curvature-based Distortion (GCD)[90], MeshPAD uses a graph autoencoder to assess edge sensitivity and SplineCNN to predict alteration probability based on the specified payload. Evaluation on the Humanbody and Manifold40 datasets shows up to 3% improvement in undetectability at half payloads compared to heuristic methods. Even under high payload ($\phi = 0.9$), MeshPAD maintains detection performance over 10% better than older methods. Visual analysis shows the method avoids predictable embedding patterns, making it more resistant to steganalysis. Overall, MeshPAD represents a major breakthrough surpassing previous state-of-the-art methods, particularly in payload adaptivity, detection resilience, and stability across varied topological structures.

4.3. Finding and Critical Analysis

Based on a review of recent methods in 3D mesh steganography and reversible data hiding (RDH), it can be concluded that current research directions are primarily focused on

three dominant objectives: (1) increasing embedding capacity, (2) enhancing imperceptibility or minimizing visual distortion, and (3) strengthening reversibility and data security in encrypted domains. Most methods prioritize spatial efficiency and geometric prediction strategies, with an emphasis on the accurate recovery of the original model after extraction.

One notable finding is the continued dominance of non-AI approaches such as multi-MSB prediction, mesh subdivision, Huffman coding, and spatial partitioning. For example, methods by Lyu et al., Zhang et al. (MMPRDH-ED), Wang et al. (CPP), and Hou et al. (Oc-tree-MSB) all rely on rule-based or deterministic prediction techniques without involving machine learning models, despite being published within the last three years in high-impact journals. Only one recent study, MeshPAD by Peng et al.[89] utilized geometric deep learning through a combination of graph autoencoders and SplineCNN to perform adaptive and context-aware embedding based on payload and mesh topology. This indicates that 3D mesh steganography remains an open domain for further exploration of AI-based methods, particularly in applying GNNs, spatial Transformers, and generative models.

In terms of evaluation, there is a notable distinction between the use of PSNR (Peak Signal-to-Noise Ratio) and SNR (Signal-to-Noise Ratio). PSNR is commonly used in both image and 3D mesh domains to measure the ratio between the original signal and the distortion introduced by embedding, in a logarithmic scale. PSNR calculates the maximum relative error based on the maximum value scale (typically 255 in images), while SNR measures the average signal-to-noise ratio. In the mesh domain, SNR is often used to evaluate vertex coordinate recovery, whereas PSNR remains popular for its communicability as a standard numeric quality metric. However, it is important to note that PSNR and SNR are not the only relevant metrics.

Evaluation of imperceptibility and robustness in 3D mesh steganography should be expanded using metrics such as:

- Hausdorff Distance and Root Mean Square Error (RMSE): to measure the geometric distance between the original and stego mesh.
- Coordinate Change Rate (CCR) and Number of Coordinate Changing Rate (NCCR): to measure the number of modified vertices.
- Entropy and Vertex Utilization Rate: to assess embedding efficiency.
- Steganalysis Feature Sets such as YANG40, LFS76, and DNN-based detectors: to evaluate resistance to detection.

Some newer metrics such as NVT+, LFS52, and OOB errors in stego detection are increasingly used in recent studies, such as in the MeshPAD method. This reflects a shift toward distributional security-based evaluation rather than purely numeric distortion.

In terms of datasets, the PSB[81] dominates nearly all studies. This is due to its diversity in object shapes, resolution, and topology, along with its wide public availability. Models such as Bunny, Armadillo, Dragon, and Elephant from PSB have become standard benchmarks as they represent practical challenges in complex geometry and vertex density variation. Other datasets like ModelNet, Manifold40, and Humanbody have started to be adopted in AI-driven approaches as they are better suited for training and validating deep learning-based models. Recommendations for future research include:

- Further exploration of AI and geometric deep learning applications in 3D steganography, particularly using GNNs and spatial Transformers.
- The need for standardized and comprehensive evaluation tools, incorporating spatial, statistical, and distributional metrics.
- Development of benchmark 3D mesh datasets with ground-truth steganography data to support more objective cross-method validation.
- Evaluations should include robustness to topological manipulations such as smoothing, simplification, reordering, and rendering effects (e.g., shading).

Overall, 3D meshes steganography holds significant potential as a new frontier in steganographic research, requiring contextual, robust, and adaptive approaches tailored to the unique characteristics of spatial 3D data.

5. Comparative Summary and Discussion

This survey has successfully fulfilled its objective to provide a structured, critical, and comparative review of AI-powered steganography across three distinct modalities: image, linguistic, and 3D mesh. Rather than focusing solely on the well-established image domain, this

work introduces underexplored but promising areas—text-based (linguistic) and 3D mesh steganography, to expand the reader’s understanding of modern steganographic systems in the AI era. Each modality shows differing design philosophies, performance priorities, and evaluation criteria, as summarized in Table 1.

Table 1. Comparative Summary of AI Steganography Across Modalities.

Aspect / Modality	Image-Based	Linguistic	3D Mesh
Primary Challenge	Undetectability vs. Payload	Semantic Fluency vs. Payload	Reversibility vs. Topological Fidelity
Main Approach	CNN	LLM (GPT, BERT)	Rule-based
	GAN	Masked LM	GNN
	Transformer		Spatial Partitioning
Imperceptibility	PSNR	PPL	PSNR
	SSIM	BLEU	SNR
	LPIPS →recommended for AI based	BERTScore	Hausdorff
	FID→recommended for AI based	KL div	Coordinate Change Rate (CCR)
Payload Unit	bits per pixel (bpp)	bits per word/token (bpw)	bits per vertex (bpv)
Extraction Metric	BER	Bit recovery rate	BER
	Extraction Accuracy	Flag bits	Reversibility
		BLEU semantic match	CCR
Steganalysis Metric	Detection Error Rate (PE)	KL-Div	OOB Error
	ROC-AUC (ZhuNet, SRNet)	Distribution Shift	Entropy Difference
		LLM Classifier	Detection Rate (LFS, VND)
Robustness Concern	Compression cropping noise	Token ambiguity synonym mapping	Reordering simplification shading
AI Impact	Embedding	Generation	Still emerging, mostly non-AI methods
	Robustness	Reversibility	
	Generation	Fluency	

A comparative analysis across the three modalities, i.e., image, linguistic, and 3D mesh, reveals distinct developmental trajectories, technical emphases, and evaluation strategies. As summarized in Table 1, each modality prioritizes different aspects of steganographic performance. Image steganography has evolved into a mature domain where AI techniques, particularly convolutional networks, and GANs, have enabled high imperceptibility and extraction accuracy. Nevertheless, some deep models incur significant computational costs and imperfect extraction under high payload conditions. Linguistic steganography, long hindered by semantic rigidity, has been rejuvenated through the rise of LLMs and neural generation, which now support both reversible and context-aware embedding. However, challenges such as metric standardization (e.g., when to use PPL vs. KL divergence vs. semantic similarity) and synchronization ambiguity persist. Meanwhile, 3D mesh steganography is still largely dominated by non-AI or heuristic approaches. The recent integration of geometric deep learning, such as in MeshPAD, signals a promising direction, with this modality uniquely capable of embedding in spatial structures and offering high capacity. Overall, the modality-specific differences in embedding principles, security concerns, and metrics underscore the need for more unified and nuanced evaluation frameworks.

Notably, this review also reinforces the importance of appropriate evaluation metrics. While PSNR and SSIM are widely used in image-based approaches, perceptual metrics like LPIPS and FID are increasingly relevant for deep learning models. Text steganography demands semantic-aware metrics like BERTScore and KL divergence, while 3D mesh evaluations rely on geometric and topological measures such as Hausdorff distance and coordinate change rates. Researchers are encouraged to move beyond numeric fidelity alone and consider semantic and perceptual realism. Looking forward, several directions emerge:

- Cross-modal and multimodal steganography: While not the focus of this review, future work could investigate the integration of multiple modalities (e.g., image+text, 3D+sound) for more robust and versatile hiding mechanisms.

- Domain adaptation and personalization: Embedding mechanisms that adapt to individual cover characteristics and user-defined constraints may enhance security and utility.
- Lightweight and mobile-ready AI models: Efficient architectures for real-time or edge devices will become increasingly important.
- Unified metrics and benchmarks: The field would benefit from standardized, task-specific datasets and evaluation frameworks tailored to AI-based steganography.

In conclusion, this paper aims not only to survey recent progress but also to inspire further exploration, especially in domains that are still underexplored. By drawing attention to linguistic and 3D mesh steganography alongside conventional image-based methods, we hope to encourage broader and more innovative applications of steganography in the AI era.

6. Ethical Concerns and Misuse Potential of AI-Powered Steganography

The increasing sophistication of AI-based steganography brings not only technical opportunities but also ethical dilemmas that deserve attention. As embedding and extraction techniques grow more precise and imperceptible, the line between legitimate usage and potential abuse becomes increasingly blurred. This section explores the ethical dimensions and misuse scenarios associated with the adoption of steganographic AI models, particularly in the context of dual-use technology.

6.1. Dual-Use Dilemma and Emerging Risks

The concept of dual-use AI, where technology designed for beneficial purposes can be repurposed for malicious activities, has been extensively discussed in recent policy and cybersecurity literature. According to the OECD's AI Principles and Policy Observatory, AI technologies with high capability and low traceability, such as steganography, must be evaluated under a framework of proportionality and risk management[91].

Several documented incidents highlight the misuse of steganography. Europol's Internet Organised Crime Threat Assessment (IOCTA 2021) [92] reported cases of extremist groups using steganographic images in public forums to share encrypted messages. Similarly, [93], [94] show that stegomalware malware hidden within media files using deep learning methods, has been actively deployed to bypass traditional antivirus systems. These examples underscore the dual-use nature of steganographic AI, where the same tools enabling privacy can also facilitate the obfuscation of harmful content.

6.2. Ethical Implications for Research and Publication

In the academic and industrial research community, openness and reproducibility of scientific work are central values. However, in the context of dual-use technologies such as AI-powered steganography, the publication of source code, pre-trained models, or detailed architectural blueprints must be accompanied by ethical reflection. Organizations such as the OECD AI Policy Observatory and the Partnership on AI have proposed frameworks for responsible publication and risk-sensitive AI deployment. These frameworks emphasize the importance of pre-release risk assessments, particularly for tools that can be used for concealment, evasion, or manipulation. For example, the Partnership on AI's guidelines highlight that developers should consider "safety-critical" properties and propose mechanisms like access restrictions, delayed-release, or third-party audits before the dual-use AI models publication [91].

A particularly relevant concern arises with synchronization-based linguistic steganography using LLMs. If such systems are made publicly available without safeguards, they may be misappropriated for untraceable covert communication by malicious actors, as outlined in the OECD's classification of high-risk AI use cases[91], and extensively discussed in Brundage et al.'s report on the malicious use of AI [95]. To address these concerns, research outputs in this field are recommended to include disclaimers, usage restrictions, and where possible, implement watermarking or auditable stego-generation logs. Furthermore, for particularly sensitive implementations, e.g., those offering high payload capacity and statistical imperceptibility. Developers are encouraged to consider controlled model licensing or gated access models, as exemplified in recent large-model deployments, e.g., GPT-4, Claude, and Gemini.

6.3. Towards Ethical AI-Steganography Design

Moving forward, the development of AI-based steganography systems should not solely prioritize performance metrics such as payload capacity, imperceptibility, or robustness to detection. Instead, it should also be grounded in design principles that support transparency, accountability, and lawful traceability. Techniques such as embedding reversibility, forensic watermarking, or even AI-detectable perturbations could be integrated to allow for legitimate auditing without undermining the system's protective goals.

As steganographic systems evolve to leverage large-scale generative models, their outputs become increasingly difficult to distinguish from natural content, raising concerns for misuse and forensic invisibility. Researchers have recently proposed the incorporation of zero-bit watermarking or statistical fingerprinting in language model outputs as a means to identify AI-generated content, approaches that may be adapted to trace steganographic text in security-sensitive environments[96], [97].

Addressing these challenges will require interdisciplinary collaboration among computer scientists, ethicists, legal scholars, and cybersecurity professionals. Governance frameworks and ethical standards, similar to those being developed for AI-generated deepfakes, synthetic media, and LLM output moderation, must be adapted for steganographic contexts as well. This is especially urgent as barriers to deploying sophisticated AI steganography continue to fall due to the open availability of large pre-trained models and the ease of fine-tuning. Without proactive ethical design, the dual-use nature of steganographic AI may outpace our ability to regulate or mitigate its unintended consequences.

7. Conclusion

This review has critically surveyed and comparatively analyzed recent advances in AI-powered steganography across image, linguistic, and 3D mesh modalities. Each domain presents distinct challenges, with AI offering substantial improvements in adaptability, imperceptibility, and embedding capacity. While image steganography continues to advance through adversarial and Transformer-based models, linguistic steganography leverages the generative power of LLMs for contextually aware embedding. 3D mesh steganography, while still dominated by rule-based methods, shows emerging potential through geometric deep learning.

This paper has successfully made a major contribution by presenting a tri-modal synthesis rarely seen in existing reviews, covering strategies for embedding images, text, and 3D data; a comparative framework on design priorities, robustness issues, and evaluation metrics across modalities; and identifying recent metric trends that are relevant to AI-generated media. Finally, it discusses the ethical implications of increasingly sophisticated steganography systems, especially the risk of dual use in linguistic applications.

Moving forward, we emphasize the need for cross-modal integration, domain personalization, and standard benchmarks to support reproducibility. Most importantly, as AI democratizes access to steganography tools, future work must embed ethical design principles, such as forensic traceability and model accountability - into system architectures. Only by aligning performance, practicality, and responsible use can AI-powered steganography realize its full potential in secure and trusted communications.

Author Contributions: Conceptualization: DRIMS; Methodology: DRIMS; Validation: DRIMS, SKG, and AKS; Formal analysis: DRIMS; Investigation: DRIMS, SKG and AKS; Resources: DRIMS; Writing—original draft preparation: DRIMS; Writing—review and editing: DRIMS, SKG, and AKS; Visualization: DRIMS; Supervision: DRIMS; Project administration: DRIMS; Funding acquisition: DRIMS, SKG, and AKS. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

- [1] D. R. I. M. Setiadi, S. Rustad, P. N. Andono, and G. F. Shidik, "Digital image steganography survey and investigation (goal, assessment, method, development, and dataset)," *Signal Processing*, vol. 206, p. 108908, May 2023, doi: 10.1016/j.sigpro.2022.108908.
- [2] N. Rijati, A. K. Sahu, D. R. I. M. Setiadi, and A. Sambas, "DWT-DCT Image Watermarking with Quantum-inspired Optimization," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 1, pp. 1034–1044, Feb. 2025, doi: 10.22266/ijies2025.0229.74.
- [3] K. Hu, M. Wang, X. Ma, J. Chen, X. Wang, and X. Wang, "Learning-based image steganography and watermarking: A survey," *Expert Syst. Appl.*, vol. 249, no. PC, p. 123715, Sep. 2024, doi: 10.1016/j.eswa.2024.123715.
- [4] S. Rahman *et al.*, "A Comprehensive Study of Digital Image Steganographic Techniques," *IEEE Access*, vol. 11, no. January, pp. 6770–6791, 2023, doi: 10.1109/ACCESS.2023.3237393.
- [5] D. R. I. M. Setiadi, "Improved payload capacity in LSB image steganography uses dilated hybrid edge detection," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 2, pp. 104–114, Feb. 2022, doi: 10.1016/j.jksuci.2019.12.007.
- [6] B. Patwari, U. Nandi, and S. K. Ghosal, "Image steganography based on difference of Gaussians edge detection," *Multimed. Tools Appl.*, vol. 82, no. 28, pp. 43759–43779, Nov. 2023, doi: 10.1007/s11042-023-15360-7.
- [7] A. K. Sahu, G. Swain, M. Sahu, and J. Hemalatha, "Multi-directional block based PVD and modulus function image steganography to avoid FOBP and IEP," *J. Inf. Secur. Appl.*, vol. 58, p. 102808, May 2021, doi: 10.1016/j.jisa.2021.102808.
- [8] T. Pevný, T. Filler, and P. Bas, "Using high-dimensional image models to perform highly undetectable steganography," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6387 LNCS, pp. 161–177, 2010, doi: 10.1007/978-3-642-16435-4_13.
- [9] B. Li, M. Wang, J. Huang, and X. Li, "A new cost function for spatial image steganography," in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct. 2014, pp. 4206–4210. doi: 10.1109/ICIP.2014.7025854.
- [10] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," in *2012 IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec. 2012, pp. 234–239. doi: 10.1109/WIFS.2012.6412655.
- [11] V. Sedighi, R. Cogranne, and J. Fridrich, "Content-Adaptive Steganography by Minimizing Statistical Detectability," *IEEE Trans. Inf. Forensics Secur.*, vol. 11, no. 2, pp. 221–234, Feb. 2016, doi: 10.1109/TIFS.2015.2486744.
- [12] D. R. I. M. Setiadi, S. Rustad, P. N. Andono, and G. F. Shidik, "Maximizing complex features to minimize the detectability of content-adaptive steganography," *Multimed. Tools Appl.*, Aug. 2024, doi: 10.1007/s11042-024-20056-7.
- [13] S. Kaur, S. Singh, M. Kaur, and H.-N. Lee, "A Systematic Review of Computational Image Steganography Approaches," *Arch. Comput. Methods Eng.*, vol. 29, no. 7, pp. 4775–4797, Nov. 2022, doi: 10.1007/s11831-022-09749-0.
- [14] T. Muralidharan, A. Cohen, A. Cohen, and N. Nissim, "The infinite race between steganography and steganalysis in images," *Signal Processing*, vol. 201, p. 108711, 2022, doi: 10.1016/j.sigpro.2022.108711.
- [15] P. Bedi, A. Singhal, and V. Bhasin, "Deep learning based active image steganalysis: a review," *Int. J. Syst. Assur. Eng. Manag.*, vol. 15, no. 3, pp. 786–799, Mar. 2024, doi: 10.1007/s13198-023-02203-9.
- [16] B. Song, P. Wei, S. Wu, Y. Lin, and W. Zhou, "A survey on Deep-Learning-based image steganography," *Expert Syst. Appl.*, vol. 254, no. May, p. 124390, Nov. 2024, doi: 10.1016/j.eswa.2024.124390.
- [17] 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, 2021, doi: 10.1109/ACCESS.2021.3053998.
- [18] M. H. Noor Azam, F. Ridzuan, M. N. S. Mohd Sayuti, A. H. Azni, N. H. Zakaria, and V. Potdar, "A systematic review on cover selection methods for steganography: Trend analysis, novel classification and analysis of the elements," *Comput. Sci. Rev.*, vol. 56, no. January, p. 100726, May 2025, doi: 10.1016/j.cosrev.2025.100726.
- [19] J. Kunthoth, N. Subramanian, S. Al-Maadeed, and A. Bouridane, "Video steganography: recent advances and challenges," *Multimed. Tools Appl.*, vol. 82, no. 27, pp. 41943–41985, Nov. 2023, doi: 10.1007/s11042-023-14844-w.
- [20] B. Chen, Y. Hong, and Y. Nie, "Deep video steganography using temporal-attention-based frame selection and spatial sparse adversarial attack," *J. Vis. Commun. Image Represent.*, vol. 104, p. 104311, Oct. 2024, doi: 10.1016/j.jvcir.2024.104311.
- [21] S. R. B., M. G. R., and M. C. Belavagi, "A Mellin transform based video steganography with improved resistance to deep learning steganalysis for next generation networks," *MethodsX*, vol. 13, p. 102887, Dec. 2024, doi: 10.1016/j.mex.2024.102887.
- [22] Y. Sanjalawe, S. Al-E'mari, S. Fraihat, M. Abualhaj, and E. Alzubi, "A deep learning-driven multi-layered steganographic approach for enhanced data security," *Sci. Rep.*, vol. 15, no. 1, p. 4761, Feb. 2025, doi: 10.1038/s41598-025-89189-5.
- [23] D. T. Firdaus, N. J. D. La Croix, T. Ahmad, D. Mukanyiligira, and L. Sibomana, "Steganographic model to conceal the secret data in audio files utilizing a fourfold paradigm: Interpolation, multi-layering, optimized sample space, and smoothing," *J. Saf. Sci. Resil.*, vol. 6, no. 2, pp. 138–149, Jun. 2025, doi: 10.1016/j.jnlssr.2024.09.004.
- [24] C. Y. Chang and S. Clark, "Linguistic steganography using automatically generated paraphrases," *NAACL HLT 2010 - Hum. Lang. Technol. 2010 Annu. Conf. North Am. Chapter Assoc. Comput. Linguist. Proc. Main Conf.*, no. June, pp. 591–599, 2010.
- [25] C.-Y. Chang and S. Clark, "Practical Linguistic Steganography using Contextual Synonym Substitution and a Novel Vertex Coding Method," *Comput. Linguist.*, vol. 40, no. 2, pp. 403–448, Jun. 2014, doi: 10.1162/COLI_a_00176.
- [26] J. Wu, Z. Wu, Y. Xue, J. Wen, and W. Peng, "Generative Text Steganography with Large Language Model," in *Proceedings of the 32nd ACM International Conference on Multimedia*, Oct. 2024, pp. 10345–10353. doi: 10.1145/3664647.3680562.
- [27] M. Steinebach, "Natural Language Steganography by ChatGPT," in *Proceedings of the 19th International Conference on Availability, Reliability and Security*, Jul. 2024, pp. 1–9. doi: 10.1145/3664476.3670930.
- [28] X. Xiang, Y. Tan, J. Qin, and Y. Tan, "Advancements and challenges in coverless image steganography: A survey," *Signal Processing*, vol. 228, no. May 2024, p. 109761, Mar. 2025, doi: 10.1016/j.sigpro.2024.109761.
- [29] H. Zhou, W. Zhang, K. Chen, W. Li, and N. Yu, "Three-Dimensional Mesh Steganography and Steganalysis: A Review," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 12, pp. 5006–5025, Dec. 2022, doi: 10.1109/TVCG.2021.3075136.
- [30] C. Xiao, S. Peng, L. Zhang, J. Wang, D. Ding, and J. Zhang, "A transformer-based adversarial network framework for steganography," *Expert Syst. Appl.*, vol. 269, no. July 2024, p. 126391, Apr. 2025, doi: 10.1016/j.eswa.2025.126391.

- [31] Z. Li *et al.*, “Adversarial feature hybrid framework for steganography with shifted window local loss,” *Neural Networks*, vol. 165, pp. 358–369, Aug. 2023, doi: 10.1016/j.neunet.2023.05.053.
- [32] K. Wang, Y. Zhu, Q. Chang, J. Wang, and Y. Yao, “High-accuracy image steganography with invertible neural network and generative adversarial network,” *Signal Processing*, vol. 234, no. January, p. 109988, Sep. 2025, doi: 10.1016/j.sigpro.2025.109988.
- [33] X. Duan, C. Li, B. Wei, G. Wu, C. Qin, and H. Nam, “DHU-Net: High-capacity binary data hiding network based on improved U-Net,” *Neurocomputing*, vol. 576, no. November 2023, p. 127314, Apr. 2024, doi: 10.1016/j.neucom.2024.127314.
- [34] P. N. Andono and D. R. I. M. Setiadi, “Quantization selection based on characteristic of cover image for PVD Steganography to optimize imperceptibility and capacity,” *Multimed. Tools Appl.*, vol. 82, no. 3, pp. 3561–3580, Jan. 2023, doi: 10.1007/s11042-022-13393-y.
- [35] Z. Wang, X. Zhang, and Z. Yin, “Joint Cover-Selection and Payload-Allocation by Steganographic Distortion Optimization,” *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1530–1534, 2018, doi: 10.1109/LSP.2018.2865888.
- [36] X. Li, D. Guo, and C. Qin, “Diversified Cover Selection for Image Steganography,” *Symmetry (Basel)*, vol. 15, no. 11, p. 2024, Nov. 2023, doi: 10.3390/sym15112024.
- [37] Z. Wang and X. Zhang, “Secure Cover Selection for Steganography,” *IEEE Access*, vol. 7, pp. 57857–57867, 2019, doi: 10.1109/ACCESS.2019.2914226.
- [38] Y. Yao, J. Wang, Q. Chang, Y. Ren, and W. Meng, “High invisibility image steganography with wavelet transform and generative adversarial network,” *Expert Syst. Appl.*, vol. 249, no. PA, p. 123540, Sep. 2024, doi: 10.1016/j.eswa.2024.123540.
- [39] Ambika, Virupakshappa, and S. Veerashetty, “Secure communication over wireless sensor network using image steganography with generative adversarial networks,” *Meas. Sensors*, vol. 24, no. July, p. 100452, Dec. 2022, doi: 10.1016/j.measen.2022.100452.
- [40] X. Duan *et al.*, “High-Capacity Image Steganography Based on Improved FC-DenseNet,” *IEEE Access*, vol. 8, pp. 170174–170182, 2020, doi: 10.1109/ACCESS.2020.3024193.
- [41] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, “HiDDeN: Hiding Data With Deep Networks,” in *Computer Vision – ECCV 2018*, 2018, pp. 682–697. doi: 10.1007/978-3-030-01267-0_40.
- [42] K. A. Zhang, A. Cuesta-infante, L. Xu, K. Veeramachaneni, and J. Carlos, “SteganoGAN: High Capacity Image Steganography with GANs,” *arXiv*, Jan. 2019, [Online]. Available: <http://arxiv.org/abs/1901.03892>
- [43] Y. Zhou, T. Luo, Z. He, G. Jiang, H. Xu, and C.-C. Chang, “CAISFormer: Channel-wise attention transformer for image steganography,” *Neurocomputing*, vol. 603, no. July, p. 128295, Oct. 2024, doi: 10.1016/j.neucom.2024.128295.
- [44] L. Meng, X. Jiang, and T. Sun, “A review of coverless steganography,” *Neurocomputing*, vol. 566, p. 126945, Jan. 2024, doi: 10.1016/j.neucom.2023.126945.
- [45] Y. Zhang, Y. Chen, H. Dou, C. Tan, Y. Luo, and H. Sang, “Image steganography without embedding by carrier secret information for secure communication in networks,” *PLoS One*, vol. 19, no. 9, p. e0308265, Sep. 2024, doi: 10.1371/journal.pone.0308265.
- [46] X. Liu, Z. Ma, J. Ma, J. Zhang, G. Schaefer, and H. Fang, “Image Disentanglement Autoencoder for Steganography without Embedding,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, vol. 2022-June, pp. 2293–2302. doi: 10.1109/CVPR52688.2022.00234.
- [47] Digital Data Embedding Laboratory, “DDE Download Section,” *DDE Laboratory Binghamton University*. <https://dde.binghamton.edu/download/>
- [48] W. Li, H. Wang, Y. Chen, S. M. Abdullahi, and J. Luo, “Constructing Immunized Stego-Image for Secure Steganography via Artificial Immune System,” *IEEE Trans. Multimed.*, vol. 25, pp. 8320–8333, 2023, doi: 10.1109/TMM.2023.3234812.
- [49] A. Gupta, “BOWS2,” *Mendeley Data*, 2023. <https://data.mendeley.com/datasets/kb3ngxfmjw/1>
- [50] Troyes University of Technology, “ALASKA2 Image Steganalysis,” *Kaggle.com*. <https://www.kaggle.com/c/alaska2-image-steganalysis>
- [51] W. Li, H. Wang, and Y. Chen, “From Cover to Immucover: Adversarial Steganography via Immunized Cover Construction,” *IEEE Trans. Fuzzy Syst.*, vol. 32, no. 3, pp. 1233–1247, Mar. 2024, doi: 10.1109/TFUZZ.2023.3321740.
- [52] Y. Chen, H. Wang, W. Li, and W. Li, “A Steganography Immunoprocessing Framework Against CNN-Based and Handcrafted Steganalysis,” *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 6055–6069, 2024, doi: 10.1109/TIFS.2024.3409075.
- [53] D. R. I. M. Setiadi, “PSNR vs SSIM: imperceptibility quality assessment for image steganography,” *Multimed. Tools Appl.*, vol. 80, no. 6, pp. 8423–8444, Mar. 2021, doi: 10.1007/s11042-020-10035-z.
- [54] L. Xiang, R. Wang, Z. Yang, and Y. Liu, “Generative Linguistic Steganography: A Comprehensive Review,” *KSII Trans. Internet Inf. Syst.*, vol. 16, no. 3, pp. 986–1005, Mar. 2022, doi: 10.3837/tiis.2022.03.013.
- [55] M. Knöchel and S. Karius, “Text Steganography Methods and their Influence in Malware: A Comprehensive Overview and Evaluation,” in *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, Jun. 2024, pp. 113–124. doi: 10.1145/3658664.3659637.
- [56] M. A. Majeed, R. Sulaiman, Z. Shukur, and M. K. Hasan, “A Review on Text Steganography Techniques,” *Mathematics*, vol. 9, no. 21, p. 2829, Nov. 2021, doi: 10.3390/math9212829.
- [57] R. Yan, T. Song, and Y. Yang, “TokenFree: A Tokenization-Free Generative Linguistic Steganographic Approach with Enhanced Imperceptibility,” in *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2024, pp. 449–455. doi: 10.1109/SMC54092.2024.10831652.
- [58] Z. Ziegler, Y. Deng, and A. Rush, “Neural Linguistic Steganography,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Sep. 2019, pp. 1210–1215. doi: 10.18653/v1/D19-1115.
- [59] L. Xiang, J. Xia, Y. Liu, and Y. Gui, “CPG-LS: Causal Perception Guided Linguistic Steganography,” *IEEE Signal Process. Lett.*, vol. 30, pp. 1762–1766, 2023, doi: 10.1109/LSP.2023.3332298.
- [60] P. Koehn, “CC-100: Monolingual Datasets from Web Crawl Data,” *Statistical and Neural Machine Translation*. <https://data.statmt.org/cc-100/>

- [61] H. Ueoka, Y. Murawaki, and S. Kurohashi, "Frustratingly Easy Edit-based Linguistic Steganography with a Masked Language Model," *arXiv*, Apr. 20, 2021. [Online]. Available: <http://arxiv.org/abs/2104.09833>
- [62] X. Zheng and H. Wu, "Autoregressive Linguistic Steganography Based on BERT and Consistency Coding," *Secur. Commun. Networks*, vol. 2022, pp. 1–11, May 2022, doi: 10.1155/2022/9092785.
- [63] K. Winstein, "Framework for Lexical Steganography." 1998. [Online]. Available: <https://web.mit.edu/keithw/tlex/lsteg.pdf>
- [64] C.-C. Chang, "Reversible Linguistic Steganography With Bayesian Masked Language Modeling," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 2, pp. 714–723, Apr. 2023, doi: 10.1109/TCSS.2022.3162233.
- [65] T. Yang, H. Wu, B. Yi, G. Feng, and X. Zhang, "Semantic-Preserving Linguistic Steganography by Pivot Translation and Semantic-Aware Bins Coding," *IEEE Trans. Dependable Secur. Comput.*, vol. 21, no. 1, pp. 139–152, Jan. 2024, doi: 10.1109/TDSC.2023.3247493.
- [66] C. Ding, Z. Fu, Z. Yang, Q. Yu, D. Li, and Y. Huang, "Context-Aware Linguistic Steganography Model Based on Neural Machine Translation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 868–878, 2024, doi: 10.1109/TASLP.2023.3340601.
- [67] L. Xiang, C. Ou, and D. Zeng, "Linguistic Steganography: Hiding Information in Syntax Space," *IEEE Signal Process. Lett.*, vol. 31, pp. 261–265, 2024, doi: 10.1109/LSP.2023.3347153.
- [68] M. V. Namitha, G. R. Manjula, and M. C. Belavagi, "StegAbb: A Cover-Generating Text Steganographic Tool Using GPT-3 Language Modeling for Covert Communication Across SDRs," *IEEE Access*, vol. 12, pp. 82057–82067, 2024, doi: 10.1109/ACCESS.2024.3411288.
- [69] C. Ding, Z. Fu, Q. Yu, F. Wang, and X. Chen, "Joint Linguistic Steganography With BERT Masked Language Model and Graph Attention Network," *IEEE Trans. Cogn. Dev. Syst.*, vol. 16, no. 2, pp. 772–781, Apr. 2024, doi: 10.1109/TCDS.2023.3296413.
- [70] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," *Proc. 8th Int. Conf. Lang. Resour. Eval. Lr. 2012*, pp. 2214–2218, 2012.
- [71] Y. Qi, K. Chen, K. Zeng, W. Zhang, and N. Yu, "Provably Secure Disambiguating Neural Linguistic Steganography," *IEEE Trans. Dependable Secur. Comput.*, vol. PP, pp. 1–14, 2024, doi: 10.1109/TDSC.2024.3519322.
- [72] G. Hou, B. Ou, M. Long, and F. Peng, "Separable Reversible Data Hiding for Encrypted 3D Mesh Models Based on Octree Subdivision and Multi-MSB Prediction," *IEEE Trans. Multimed.*, vol. 26, pp. 2395–2407, 2024, doi: 10.1109/TMM.2023.3295578.
- [73] H. Zhou, K. Chen, Z. Ma, F. Wang, and W. Zhang, "Basic Concepts," in *Triangle Mesh Watermarking and Steganography*, Singapore: Springer Nature Singapore, 2023, pp. 11–42. doi: 10.1007/978-981-19-7720-6_2.
- [74] G. Zhang, Z. Sui, C. Sun, Q. Liu, and X. Cheng, "A multi-layer mesh synchronized reversible data hiding algorithm on the 3D model," *Multimed. Syst.*, vol. 30, no. 1, p. 52, Feb. 2024, doi: 10.1007/s00530-023-01253-0.
- [75] H. Zhou, K. Chen, Z. Ma, F. Wang, and W. Zhang, "3D Mesh Steganography," in *Triangle Mesh Watermarking and Steganography*, Singapore: Springer Nature Singapore, 2023, pp. 109–143. doi: 10.1007/978-981-19-7720-6_4.
- [76] Ying Yang, N. Peyerimhoff, and I. Ivriissimtzis, "Linear Correlations between Spatial and Normal Noise in Triangle Meshes," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 1, pp. 45–55, Jan. 2013, doi: 10.1109/TVCG.2012.106.
- [77] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 1025–1035, Jun. 2017, [Online]. Available: <http://arxiv.org/abs/1706.02216>
- [78] Z. Li and A. G. Bors, "3D mesh steganalysis using local shape features," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 2144–2148. doi: 10.1109/ICASSP.2016.7472056.
- [79] Z. Li and A. G. Bors, "Steganalysis of 3D objects using statistics of local feature sets," *Inf. Sci. (Njy)*, vol. 415–416, pp. 85–99, Nov. 2017, doi: 10.1016/j.ins.2017.06.011.
- [80] W.-L. Lyu, L. Cheng, and Z. Yin, "High-capacity reversible data hiding in encrypted 3D mesh models based on multi-MSB prediction," *Signal Processing*, vol. 201, p. 108686, Dec. 2022, doi: 10.1016/j.sigpro.2022.108686.
- [81] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The princeton shape benchmark," in *Proceedings Shape Modeling Applications, 2004.*, 2004, vol. 08540, pp. 167–388. doi: 10.1109/SMI.2004.1314504.
- [82] G. Zhang, X. Cheng, F. Yang, A. Wang, X. Zhang, and L. Liu, "High-capacity multi-MSB predictive reversible data hiding in encrypted domain for triangular mesh models," *J. Vis. Commun. Image Represent.*, vol. 103, no. April, p. 104246, Aug. 2024, doi: 10.1016/j.jvcir.2024.104246.
- [83] Y. Zhang, J. Zhu, M. Xue, X. Zhang, and X. Cao, "Adaptive 3D Mesh Steganography Based on Feature-Preserving Distortion," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 8, pp. 5299–5312, Aug. 2024, doi: 10.1109/TVCG.2023.3289234.
- [84] Zhirong Wu *et al.*, "3D ShapeNets: A deep representation for volumetric shapes," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 1912–1920. doi: 10.1109/CVPR.2015.7298801.
- [85] M. Levoy, "The Stanford 3D Scanning Repository," *Stanford University Computer Graphics Laboratory*. <http://graphics.stanford.edu/data/3Dscanrep/>
- [86] X. Wang, J.-C. Liu, C.-C. Chang, and C.-C. Chang, "Reversible Data Hiding in Encrypted 3D Mesh Models Based on Multi-Group Partition and Closest Pair Prediction," *Futur. Internet*, vol. 16, no. 6, p. 210, Jun. 2024, doi: 10.3390/fi16060210.
- [87] K. Gao, J.-H. Horng, and C.-C. Chang, "Reversible Data Hiding for Encrypted 3D Mesh Models With Secret Sharing Over Galois Field," *IEEE Trans. Multimed.*, vol. 26, pp. 5499–5510, 2024, doi: 10.1109/TMM.2023.3334972.
- [88] K. Gao, J. Horng, C.-C. Chang, and C. Chang, "Crypto-space steganography for 3D mesh models with greedy selection and shortest expansion," *Displays*, vol. 87, no. August 2024, p. 102961, Apr. 2025, doi: 10.1016/j.displa.2024.102961.
- [89] W. Peng, K. Tang, W. Tang, Y. Su, M. Fang, and P. Li, "MeshPAD: Payload-aware mesh distortion for 3D steganography based on geometric deep learning," *Expert Syst. Appl.*, vol. 272, no. February, p. 126684, May 2025, doi: 10.1016/j.eswa.2025.126684.
- [90] H. Zhou, K. Chen, W. Zhang, Y. Yao, and N. Yu, "Distortion Design for Secure Adaptive 3-D Mesh Steganography," *IEEE Trans. Multimed.*, vol. 21, no. 6, pp. 1384–1398, Jun. 2019, doi: 10.1109/TMM.2018.2882088.
- [91] W. Thawornwichian, "OECD Framework for the Classification of AI systems," Feb. 2022. doi: 10.1787/cb6d9eca-en.
- [92] "Europol: Internet Organised Crime Threat Assessment (IOCTA) 2021," *Comput. Fraud Secur.*, vol. 2021, no. 12, p. 4, Dec. 2021, doi: 10.1016/S1361-3723(21)00125-1.

-
- [93] R. Chaganti, V. Ravi, M. Alazab, and T. D. Pham, "Stegomalware: A Systematic Survey of MalwareHiding and Detection in Images, Machine LearningModels and Research Challenges," *arXiv*. Oct. 06, 2021. [Online]. Available: <http://arxiv.org/abs/2110.02504>
 - [94] L. Caviglione and W. Mazurczyk, "Never Mind the Malware, Here's the Stegomalware," *IEEE Secur. Priv.*, vol. 20, no. 5, pp. 101–106, Sep. 2022, doi: 10.1109/MSEC.2022.3178205.
 - [95] M. Brundage *et al.*, "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," *arXiv*. Feb. 20, 2018. [Online]. Available: <http://arxiv.org/abs/1802.07228>
 - [96] X. Zhong, A. Dasgupta, and A. Tanvir, "Watermarking Language Models through Language Models," *arXiv*. Nov. 07, 2024. [Online]. Available: <http://arxiv.org/abs/2411.05091>
 - [97] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature," *arXiv*. Jan. 26, 2023. [Online]. Available: <http://arxiv.org/abs/2301.11305>