Proceedings of CONF-MLA 2024 Workshop: Securing the Future: Empowering Cyber Defense with Machine Learning and Deep Learning
DOI: 10.54254/2755-2721/110/2024MELB0094

Enhancing Image Steganography through Generative Adversarial Networks: Applications and Innovations

Xu He

School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China

hexu@bupt.edu.cn

Abstract. Image steganography, a technique for embedding secret information within images, has evolved significantly with the introduction of Generative Adversarial Networks (GANs). Traditional methods often struggled with limitations such as low steganographic capacity and poor image quality. This article explores the integration of GANs into image steganography, focusing on three main applications: carrier modification, carrier selection, and carrier synthesis. GANs enhance the embedding capacity, imperceptibility, and security of steganographic systems by generating encrypted images that are robust against advanced steganalysis techniques. The study examines the advancements and challenges in applying GANs, highlighting the potential for further research and application. It is noted that while GANs offer substantial improvements, the diversity of methods and practical applications remains limited. Future research directions include exploring diverse techniques and enhancing generative models to produce more sophisticated steganographic content that could be integrated into daily use, aiming for broader and more secure applications in secure communications.

Keywords: Image steganography, generate adversarial networks, carrier modification, carrier selection, carrier synthesis.

1. Introduction

Image steganography serves as a critical technology for the secure and covert transmission of confidential information, holding significant value in domains such as national security, copyright protection, and private communications [1]. Traditionally, the practice has faced challenges such as limited capacity for data embedding, poor visual quality of steganographic images, and vulnerability to steganalysis when subjected to complex image processing or attacks [2]. These limitations underscore the need for more advanced solutions that can enhance both the capacity and security of steganographic methods.

As the field of deep learning continues to evolve, Generative Adversarial Networks (GANs) have emerged as a powerful tool for numerous applications, including image steganography. The unique architecture of GANs, comprising a generator and a discriminator that engage in a zero-sum game to improve each other, presents new avenues for embedding secret information into images with enhanced security and imperceptibility [3]. Recent advancements have shown that GANs can significantly improve steganographic methods by generating encrypted images that are robust against modern steganalysis techniques, thereby addressing some of the core issues faced by traditional methods [4].

DOI: 10.54254/2755-2721/110/2024MELB0094

This article delves into the integration of GANs into image steganography, exploring three main aspects: carrier modification, carrier selection, and carrier synthesis using GANs. Each section examines how GANs can be tailored to enhance the effectiveness of steganographic practices, focusing on the capacity for data embedding, the imperceptibility of the modifications, and resistance to detection. By harnessing the power of adversarial training, this study aims to push the boundaries of what is currently achievable in the realm of secure and discreet communication [5]. Through a comprehensive analysis of current techniques and the potential for future innovations, this article provides valuable insights into the evolving landscape of image steganography enhanced by GAN technology.

2. Fundamentals of GANs

2.1. Core principles of GANs

The generator G is essentially a differentiable function capable of theoretically learning any probability distribution and producing realistic-looking images. However, it is important to note that the generated images may not be entirely consistent with authentic images.

Instead, it is an approximate distribution obtained through learning and training data, which makes the generator can be used in the domain of image steganography to generate encrypted images [5].

Its main purpose is to distinguish between generated samples and real data, thereby providing feedback to generator G and continuously improving its performance.

With an initial generator G in place, generated samples and real data are fed into the discriminator D separately. The discriminator D then outputs a value that signifies the probability it assigns to the input being real data. As show in the figure 1.

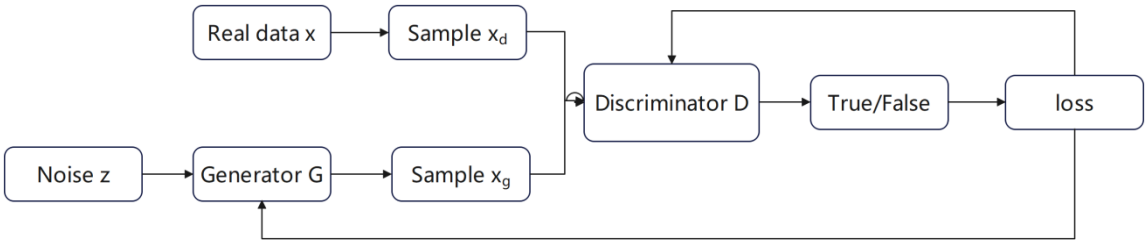

Figure 1. Schematic diagram of GAN model structure (Photo credit: Original).

The objective function of GAN is formulated as follows:

$$\min_G \max_D V(D,G) = E_{x\sim p_{\text{data}}(x)}[lbD(x)] + E_{z\sim p_z(z)}[lb(1-D(G(z)))] \quad (1)$$

In this model, generator G and discriminator D form two players in a zero sum game, who continuously train to improve their judgment and generation abilities in an attempt to win the game. In an ideal scenario, when training is completed, discriminator D should be unable to discern whether the input data originates from authentic sources or has been generated by the generator G.

2.2. Advantages of GANs in image steganography

These approaches embed information by subtly altering pixel values while preserving the overall visual integrity of the image. These transformations shift the image representation from the spatial domain to the frequency domain, enabling the concealment of information within the spectral components of the image.

With the continuous development of feature-based deep neural network steganalysis (such as CNN based steganalysis) and other technologies, the detection performance continues to improve, and traditional image steganalysis algorithms are difficult to resist these advanced steganalysis techniques. This significantly undermines the security of traditional image steganography techniques, prompting the emergence of GAN-based image steganography technology as a potential solution.

DOI: 10.54254/2755-2721/110/2024MELB0094

The image steganography technology based on GAN can theoretically ensure that the distribution of the dataset conforms to the characteristics of a natural image dataset, while greatly improving the embedding payload. It employs a two-person zero-sum game model rooted in game theory, thus avoiding the difficulties of manually designing steganography methods and establishing clear distribution models for real image data, making the design of image steganography methods to some extent detached from manual work and instead relying on machines to independently complete them.

3. Implementing GANs in image steganography

3.1. Carrier modification with GANs

Image steganography approaches utilizing GANs for carrier modification can be broadly classified into three categories, as outlined in [6]: The first type of image steganography method generates a carrier image suitable for image steganography using GAN, and then embeds secret information using traditional methods such as adaptive steganography algorithm. This type of method appeared earlier, and representative models include SGAN, SSGAN, etc [7].

The overall framework of the SSGAN model is shown in the figure. As show in the figure 2.

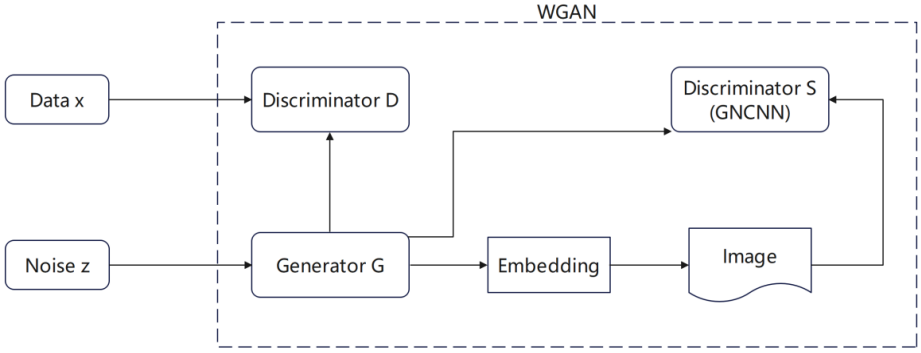

Figure 2. SSGAN model framework (Photo credit: Original).

The second type of image steganography method does not directly use GAN to generate carrier images or encrypted images, but uses GAN to find regions in the image suitable for embedding secret information and output corresponding probability maps. During the iteration process, generator G will continuously modify its judgment of the carrier image area that suitable for embedding secret information, in order to find the hidden areas that are more difficult for discriminator D to recognize. Subsequently, the actual image steganography process is finalized utilizing traditional adaptive steganography techniques.

Representative models of this method include the ASDL-GAN model and the UT-GAN model, which has been improved based on ASDL-GAN. Figure 2 illustrates the comprehensive architecture of the ASDL-GAN model [8, 9]. As show in the figure 3.

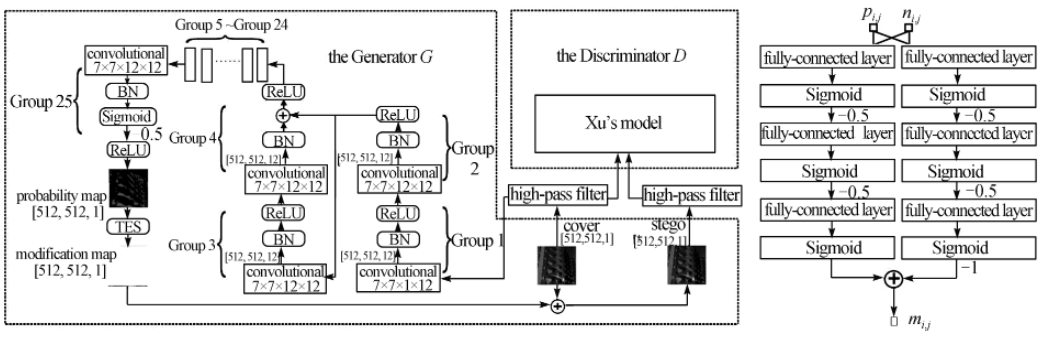

Figure 3. ASDL-GAN model framework [10].

DOI: 10.54254/2755-2721/110/2024MELB0094

The Xu's model depicted in Figure 3 represents the steganalysis model introduced in reference [11]. To facilitate the learning of the modified probability matrix P, Tang et al. introduced a micro-network known as TES, serving as the activation function for P. The configuration of this network is shown in Figure 3. UT-GAN, on the other hand, adopts U-Net as the foundational structure for its generator G, and incorporates the Tanh simulator function, inspired by ASDL-GAN, in place of the TES activation function. It replaces XuNet that only uses one high pass filter with XuNet that uses multiple high pass filters as preprocessing layers to form discriminator D [11].

The third category of image steganography method integrates secret information with the carrier image, incorporating them as part of the input to the generator G. This approach directly leverages GAN to produce encrypted images, eliminating the need for an additional embedding process.

The common idea of such methods can be summarized as a famous steganographic system model - the prisoner model [12]. In this model, Alice and Bob are prisoners held in different prisons, and they can communicate with each other, but the communication content is monitored by guard Eve. They need to transmit confidential information, but it cannot be detected by guards, otherwise the communication will be terminated. In this type of method, generator G plays the role of Alice and directly generates encrypted images that can hide secret information; And the guard Eve plays the role of discriminator D, constantly identifying whether the input image hides secret information. Both of them continuously optimize their algorithms in this adversarial training. At the same time, Bob serves as a decoder to decode encrypted images. Therefore, this type of steganography method does not require manually designed embedding algorithms and can simultaneously train the generator, discriminator, and decoder.

The classic models of this method include HayesGAN and HiDDeN [13, 14]. Building upon this foundation, diverse models boasting distinct advantages have emerged. For instance, Zhang et al. introduced the ISFGAN model, capable of concealing grayscale secret image information within color cover images of identical dimensions, generating secret images that closely resemble carrier images in terms of semantics and color [15]. Additionally, Yu et al. presented the ABDH model, adept at embedding color images into color carriers of the same resolution, effectively withstanding attacks such as noise, cropping, and JPEG compression [16].

3.2. Carrier selection via GANs

Carrier-free steganography, which utilizes GANs for carrier selection, can be broadly categorized into two groups, as outlined in:

The first one is based on mapping methods [17]. The operation process of this type of method can be roughly summarized as follows: This concludes the process of embedding the secret information. This sequence is then concatenated to reconstruct the original secret information.

Subsequently, Zheng et al. further advanced this domain by developing a carrier-free steganography algorithm leveraging Scale-Invariant Feature Transform (SIFT), thereby enhancing its embedding capacity [18].

Considering the weak robustness of block based carrier free image steganography algorithms to geometric attacks. To address both geometric and non-geometric threats more comprehensively, Liu and colleagues presented a disguised image-utilizing carrier-free steganography method, significantly improving robustness. Building upon these foundations [19].

The second type is based on generative methods [20]. Unlike the first type of method, this type of method utilizes a GAN model to directly generate steganographic images. The process is shown in Figure 4, which can also be summarized as follows: First, pre-train a generator network to the point where it is capable of producing an image that resembles a normal and undistort image by taking a noise vector as input. Subsequently, utilizing predefined mapping rules, we transform the secret information into a noise vector format. This generated noise vector is then fed into the pre-trained generator, resulting in the production of an encrypted image, which serves as a carrier for the hidden secret information. In the process of extracting secret information, most methods also require pre training an extractor and inputting the received steganographic image into the extractor to extract the corresponding noise vector. As show in the figure 4.

DOI: 10.54254/2755-2721/110/2024MELB0094

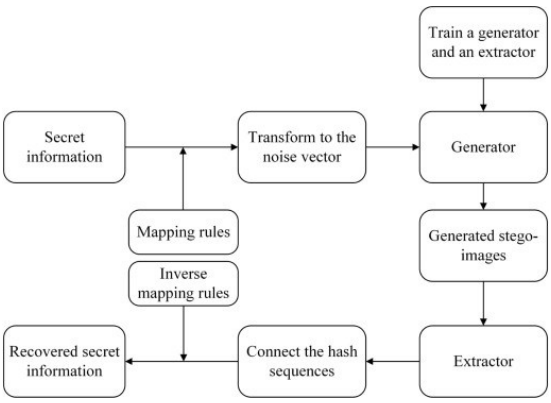

Figure 4. Based on the generated process of carrier free steganography [21].

Chen first proposed a carrier free image steganography method that combines mapping based and generation based methods using StarGAN [21].

3.3. Synthesizing carriers using GANss

The concept of generative steganography, which harnesses the capabilities of Generative Adversarial Networks (GANs) to synthesize carriers, revolutionizes traditional image steganography by dispensing with the need for an original carrier image [22-24]. This approach directly generates encrypted images from secret information, adhering to predetermined rules. This streamlines the process by eliminating the step of embedding secret information into a pre-existing carrier image, thereby mitigating the risk of steganalysis detection. Prior to the advent of deep learning networks, the direct generation of encrypted images encoded with secret data was deemed unfeasible. However, the advent of deep learning, particularly the generative prowess of GANs, has propelled this theoretical concept into the realm of practical application, as illustrated in Figure 5.

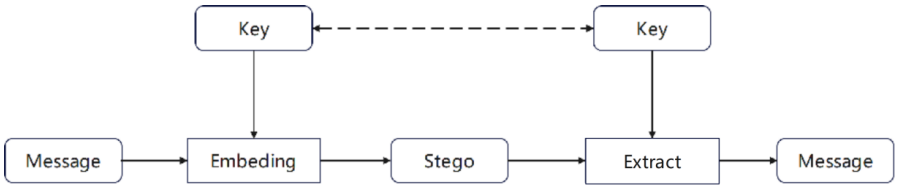

Figure 5. A framework for image steganography based on carrier synthesis [25].

The seminal GSK framework, initially introduced by Ke et al., laid the groundwork for this methodology [25]. Hu et al. presented Embeddless steganography (SWE), which generates encrypted images based on noise vectors mapped from secret information [26].

4. Conclusion

This article has extensively examined the applications and innovations of Generative Adversarial Networks in the field of image steganography. Our study delves into three core areas: carrier modification, carrier selection, and the synthesis of carriers using GANs. The insights provided shed light on the transformative impact of GANs on enhancing the capacity, security, and imperceptibility of steganographic practices. Through adversarial training, GANs can significantly improve the robustness of steganographic images against sophisticated steganalysis techniques. This development marks a significant stride forward from traditional methods, which often struggled with limited steganographic capacity and compromised visual quality.

Despite these advances, the field of GAN-based image steganography still faces numerous challenges. Current methods, while innovative, tend to lack diversity and are predominantly centered around the confrontation between steganography and steganalysis. Moreover, the dependence on algorithm

DOI: 10.54254/2755-2721/110/2024MELB0094

confidentiality raises concerns about the sustainability of security once the algorithms are disclosed. Additionally, practical applications of these technologies in everyday scenarios remain limited, highlighting a gap between theoretical research and real-world usability. Looking ahead, the future research in image steganography should aim to diversify the techniques and explore new paradigms that do not solely focus on the adversarial nature of steganography and steganalysis. Developing methods that can seamlessly integrate into daily applications will enhance the practical value of steganographic techniques. Furthermore, enhancing the generative models to produce more sophisticated and less detectable steganographic content could bridge the current gap between academic research and operational security tools. It is also crucial to continue refining GAN models to address existing challenges more effectively and to push the boundaries of what is currently achievable in secure communication technologies. By pursuing these avenues, we can foster a more secure and versatile future for image steganography.

References

[1] Yuan C, Wang H, He P et al. 2022 GAN-based image steganography for enhancing security via adversarial attack and pixel-wise deep fusion Multimed Tools Appl 81 6681–6701
[2] Goodfellow I J, Pouget-Abadie J, Mirza M et al. 2018 Generative adversarial nets [EB/OL] https://arxiv.org/pdf/1406.2661.pdf
[3] Mertens J F, Zamir S 1971 The value of two-person zero-sum repeated games with lack of information on both sides Int. J. Games Theory 1(1) 39–64
[4] Filler T, Judas J and Fridrich J 2011 Minimizing additive distortion in steganography using syndrome-trellis codes IEEE Trans Inf Forensics Secur 6(3) 920–935
[5] Volkhonskiy D, Nazarov I and Burnaev E 2020 Steganographic generative adversarial networks In: Twelfth Int. Conf. on Machine Vision (ICMV 2019) Int. Soc. for Optics and Photonics 11433 114333M
[6] Shi H, Dong J, Wang W, Qian Y and Zhang X 2017 SSGAN: secure steganography based on generative adversarial networks In: Pacific Rim Conf. on Multimedia Springer 534–544
[7] Tang W, Tan S, Li B and Huang J 2017 Automatic steganography distortion learning using a generative adversarial network IEEE Signal Process Lett 24(10) 1547–1551
[8] Yang J, Ruan D, Huang J, Kang X and Shi Y 2020 An embedding cost learning framework using GAN IEEE Trans Inf Forensics Secur 15 839–851
[9] Simmons G J 1983 The prisoners' problem and the subliminal channel Heidelberg: Springer 51–67
[10] Hayes J and Danezis G 2017 Generating steganography images via adversarial training Adv Neural Inf Process Syst 1954–1963
[11] Zhu J, Kaplan R, Johnson J and Li F 2018 Hidden: hiding data with deep networks Proc Eur Conf Comput Vis 657–672
[12] Zhang R, Dong S and Liu J 2019 Invisible steganography via generative adversarial networks Multimed Tools Appl 78(7) 8559–8575
[13] Yu C 2020 Attention based data hiding with generative adversarial networks Proc. of the AAAI Conf. on Artificial Intelligence 34(01) 1120–1128
[14] Lai J M, Jiang X and Sun T 2024 A review of coverless steganography Neurocomput 566 126945
[15] Zhang Z et al. 2019 Generative steganography by sampling IEEE Access 7 118586–118597
[16] Ke Y, Zhang M, Liu J, Su T and Yang X 2017 Generative steganography with Kerckhoffs' principle based on generative adversarial networks arXiv:1711.04916 https://arxiv.org/abs/1711.04916
[17] Liu M M, Zhang M Q, Liu J, Gao P X and Zhang Y N 2018 Coverless information hiding based on generative adversarial networks J Appl Sci 36(2) 371–382
[18] Odena A, Olah C and Shlens J 2016 Conditional image synthesis with auxiliary classifier GANs arXiv:1610.09585 https://arxiv.org/abs/1610.09585

DOI: 10.54254/2755-2721/110/2024MELB0094

[19] Hu D, Wang L, Jiang W, Zheng S and Li B 2018 A novel image steganography method via deep convolutional generative adversarial networks IEEE Access 6 38303–38314
[20] Chakraborty T, Reddy K S U, Naik S M, Panja M and Manvitha B 2024 Ten years of generative adversarial nets (GANs): a survey of the state-of-the-art Mach Learn: Sci Technol 5 011001
[21] Zhou Z, Sun H, Harit R, Chen X and Sun X 2015 Coverless image steganography without embedding Int. Conf. on Cloud Comput. and Security Springer 123–132
[22] Zheng S, Wang L, Ling B, Hu D 2017 Coverless information hiding based on robust image hashing Int. Conf. on Intell. Comput. Springer 536–547
[23] Zhang X, Peng F and Long M 2018 Robust coverless image steganography based on DCT and LDA topic classification IEEE Trans. Multimed 20(12) 3223–3238
[24] Luo Y, Qin J, Xiang X and Tan Y 2020 Coverless image steganography based on multi-object recognition IEEE Trans Circuits Syst Video Technol 31(7) 2779–2791
[25] Liu Q, Xiang X, Qin J, Tan Y and Zhang Q 2021 A robust coverless steganography scheme using camouflage image IEEE Trans Circuits Syst Video Technol 32(6) 4038–4051
[26] Zou L, Li J, Wan W, Wu Q J and Sun J R 2022 Robust coverless image steganography based on neglected coverless image dataset construction IEEE Trans. Multimed. 1–13