APSIPA Transactions on Signal and Information Processing, 2024, 13, e30

Overview Paper

A Comprehensive Survey of Digital Image Steganography and Steganalysis

Weiqi Luo[1], Kangkang Wei[2], Qiushi Li[3], Miaoxin Ye[1], Shunquan Tan[3*], Weixuan Tang[4] and Jiwu Huang[5]

[1] *School of Computer Science and Engineering, Sun Yat-sen University, China*
[2] *School of Software, Nanchang University, China*
[3] *Shenzhen Key Laboratory of Media Security, Shenzhen University, China*
[4] *Institute of Artificial Intelligence, Guangzhou University, China*
[5] *Guangdong Laboratory of Machine Perception and Intelligent Computing, Faculty of Engineering, Shenzhen MSU-BIT University, China*

ABSTRACT

In the realm of digital communications, steganography and steganalysis have become a solution for securely exchanging covert information. This survey initiates with an exploration of the widely used passive-warden scenario model, analyzing its significance, key performance indicators, relevant databases, and clarifying some commonly misunderstood fundamental concepts associated with this model. Subsequently, the paper comprehensively examines the evolution and current state of digital image steganography and steganalysis, highlighting the transition from traditional handcrafted based methods to sophisticated deep learning based techniques developed over the past two decades. It offers thorough descriptions and evaluations of typical methods in both steganography and steganalysis, with a particular emphasis on deep learning-based techniques that have emerged in recent years. Furthermore, the survey identifies significant challenges currently faced in translating theoretical research into practical applications. By integrating

[*] Corresponding author: Shunquan Tan, tansq@szu.edu.cn. This work was supported by National Natural Science Foundation of China under Grant 62472458, 62272314 and 62302458.

Received 27 June 2024; revised 15 September 2024; accepted 14 October 2024
ISSN 2048-7703; DOI 10.1561/116.20240038

these insights, the survey not only charts the historical development and technological advancements in steganography and steganalysis but also establishes a proactive agenda for future research aimed at enhancing security in covert communications.

Keywords: Steganalysis, steganography

1 Introduction

The primary goal of steganography is to facilitate secure, undetectable communication, thereby preventing any suspicion about the transmission of a hidden message. Unlike traditional cryptography, which protects information by encrypting the content into a seemingly indecipherable format, steganography hides the existence of the communication itself, embedding messages within ordinary, non-suspicious data. Historically, the practice of steganography dates back to ancient times when ingenious methods were used to hide messages. As the world transitioned into the digital age, particularly with the proliferation of internet and the rise of digital media, these developments have dramatically altered the landscape of steganographic practices, shifting them from the physical to the digital domain. Digital steganography utilizes sophisticated algorithms to conceal secret information within multiple digital formats such as text, audio, images, and video files. Subsequently, these files with embedded secret information are transmitted via public channels to achieve covert communication. In recent years, the increasing instances of covert communication through steganographic technologies have highlighted its growing use in both harmless and harmful contexts. This has raised concerns about the privacy and security aspects of these technologies. Concurrently, steganalysis technologies have emerged, and the field of steganalysis has developed as an essential countermeasure to identify and assess steganographic methods. Steganalysis primarily employs advanced techniques, including statistical analysis, machine learning, and pattern recognition, to detect differences in carrier files before and after secret information has been embedded, thereby identifying steganographic operations. This interaction is akin to a cat-and-mouse game. As shown in the timeline diagram of typical steganographic and steganalysis methods over the past two decades in Figure 1, the fields of steganography and steganalysis have significantly advanced by continuously evolving their techniques through ongoing competition.

In the late 20th century, image steganography based on LSB (Least Significant Bit) replacement and matching became fundamental in digital steganography. LSB replacement overwrites the LSB of pixel to correspond with message bit. This method introduced the structural asymmetry - specifically,

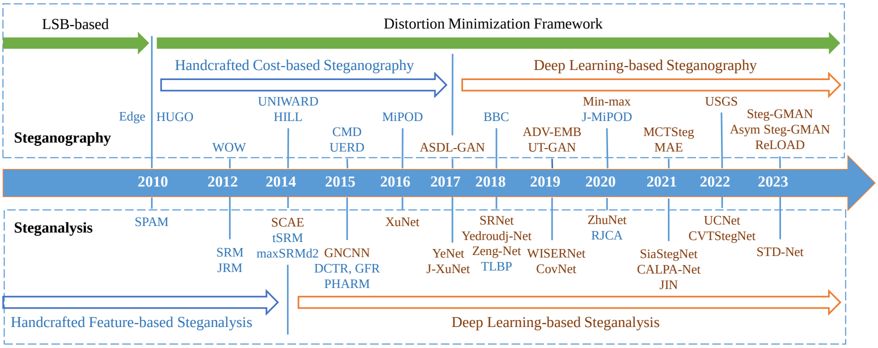


Figure 1: Timeline diagram of typical steganographic and steganalysis methods over the past 20 years. Note that the methods depicted in blue font are based on handcrafted-based techniques, while those in orange font represent deep learning-based approaches.

never decreasing even pixels and never increasing odd pixels when hiding data, which can be easily detected by simple statistical methods [36, 30, 71]. Some techniques like [40] developed to attack this form of steganography, even when the modification rate is as low as 0.0005 per pixel. In contrast, LSB matching involved adjusting the pixel value by either increasing or decreasing it by one to match the message bit, effectively overcoming the structural asymmetry issue. It is crucial to recognize that these naive methods often assume that the LSB of an image is random enough. Thus, the selection of embedding positions in a cover image relies on a pseudorandom number generator, which can introduce noticeable embedding artifacts, particularly in flat or uniform areas of the image, thereby compromising security. Several steganalysis algorithms [133, 76, 42, 70, 62, 107] proposed to detect LSB based steganography, achieving high detection accuracy. In response to these vulnerabilities, researchers have developed content-adaptive steganography methods to enhance the security. For instance, Luo *et al.* firstly proposed an edge-adaptive steganographic algorithm [106] (called "Edge" for short), which prioritizes embedding secret information into content edges while preserving smooth regions from modifications. Compared with previous LSB based methods, this approach significantly improved the imperceptibility and security. Consequently, more and more researchers have begun to focus on the impact of image content on steganography security.

In previous steganographic methods, we typically need to manually establish the embedding rules and the modification positions to ensure that the extractor could accurately retrieve the secret message. In 2007, Fridrich *et al.* introduced a general and efficient framework [39] for developing steganographic schemes that minimize the statistical impact of embedding. This framework simplifies the design of steganography into two main steps: first, designing an embedding cost for each unit, where the cost indicates the difficulty of detecting modifications after embedding: the higher the cost, the easier the detection; second, applying certain coding methods, such as Syndrome-Trellis Codes (STC) [34] and Steganographic Polar Codes (SPC) [93] techniques, for data hiding while minimizing the total embedding costs of pre-assigned embedding units. At the receiver end, the recipient does not need to be aware of the specific steganographic modifications or embedding costs at all. Instead, they can directly extract the secret information from the stego image using the corresponding decoding rules. In 2010, HUGO [120] emerged as the first method to adopt the minimal distortion framework to construct content-adaptive steganography. HUGO primarily achieves resistance to steganalysis based on second-order features by designing a distortion function that considers the impact of pixel modifications on feature perturbation. Compared to earlier LSB-based methods, HUGO significantly enhances security performance. Subsequently, a series of content-adaptive steganographic methods appeared, such as WOW [53], UNIWARD [56], HILL [78], CMD [79], UERD [47], MiPOD [132], and J-MiPOD [18], which progressively improved the security performance of image steganography. Like HUGO, these methods are based on this framework and primarily focus on designing the embedding costs in the first step. To achieve high security, the main idea of these methods was to assign small cost values to image areas that are difficult to model, while assigning large cost values to easily modeled areas such as regular textures and flat regions. However, the setup of these embedding costs and certain rules primarily depend on human experience, which significantly limits the improvement of steganographic security due to human factors. Since 2016, the advancement of deep learning methods [75], especially convolutional neural networks (CNNs) has led to significant interest in applying deep learning to image steganography, culminating in the development of methods like ASDL-GAN [154] , the first deep learning-based steganographic approach that employs generative adversarial networks (GANs) for automated embedding cost learning. Although the security performance of ASDL-GAN still does not match that of traditional steganography like S-UNIWARD, it has introduced a novel approach to steganography research. Unlike methods that manually set embedding costs, deep learning-based steganography is data-driven and is better able to learn embedding costs. Since then, a variety of deep learning-based steganographic methods have emerged, which can be categorized into two main categories: adversarial sample-based methods for adjusting the embedding costs of existing steganographic algorithms, such as ADV-EMB [152], Min-max [3], MAE [98] and USGS [99], and GAN-based methods for automatically learning the embedding costs of images, such as ASDL-GAN [154], UT-GAN [187], Steg-GMAN [60] and its asymetry version (asym Steg-GMAN) [61]. Currently, modern deep learning-based steganography methods have surpassed traditional techniques that relied on manually set embedding costs, thus becoming the dominant trend in steganography research.

Similar to the development trajectory of steganography, early steganalysis methods that contended with the minimal distortion steganography framework still relied on handcrafted statistical features. These features were designed to detect changes in the statistics of various high-frequency components and the correlations between image pixels (or DCT coefficients) caused by steganographic manipulations. In 2010, the Subtractive Pixel Adjacency Matrix (SPAM) method [119] was developed, which was particularly effective for detecting LSB matching steganography. However, as content-adaptive steganography methods based on the distortion minimization framework evolved, the performance of SPAM needed further improvement. In 2012, Fridrich *et al.* [37] introduced the Spatial Rich Model (SRM), which derived features from a series of different image high-frequency components and incorporated ensemble classifiers [66] for classification. SRM has become one of the most classic algorithms in steganalysis, influencing many subsequent methods based on both handcrafted-based and deep learning-based approaches. Within the rich model family, representative works in the spatial domain include tSRM [153], maxSRMd2 [25], and in the JPEG domain, JRM [72], DCTR [54], GFR [136], and PHARM [55]. Though methods incorporating local binary pattern (LBP) features like TLBP [77] and the Reverse JPEG Compatibility Attack (RJCA) [11] were later proposed, the rich model family had numerous relevant research until the advent of deep learning. In 2014, Tan *et al.*, recognizing structural similarities between deep neural networks and SRM, firstly introduced a deep steganalysis framework called Stacked Convolutional Auto-Encoders (SCAE) [144], achieving detection results comparable to SPAM. Since then, deep learning-based steganalysis technology has developed rapidly. By integrating knowledge from the field of steganalysis, various modules, including specialized preprocessing layers, activation layers, and pooling layers, have been introduced and designed. Representative works include spatial domain models such as GNCNN [122], XuNet [181], YeNet [191] , Yedroudj-Net [194], SRNet [6], WISERNet [206], CovNet [28], ZhuNet [212], SiaStegNet [195], and CVTStegNet [103], JPEG domain models like J-XuNet [182], ZengNet [205], as well as the universal models UCNet [167]. Notably, in 2016, XuNet outperformed rich models in steganalysis for the first time, and in 2019, SRNet became the first fully end-to-end, data-driven deep learning steganalysis model, independent of any handcrafted feature-based knowledge. It is worth noting that in the Kaggle ALASKA II competition [17], novel deep learning models from the field of computer vision were widely applied to steganalysis. This success is attributed to pretraining on large-scale datasets like ImageNet, which enabled effective transfer to JPEG steganalysis tasks, surpassing the performance of SRNet. Inspired by this, Butora *et al.* [12] generated the JIN dataset based on ImageNet for pretraining JPEG steganalysis models, significantly enhancing their performance in JPEG steganalysis tasks. Since then, some new deep learning frameworks, such as EfficientNet [143], Transformer [172, 101], etc., has been quickly used for steganalysis. In 2022, Luo *et al.* [103] first utilized Transformers for steganalysis and proposed CVTStegNet based on a CNN-Transformer architecture. From the early SPAM with 686-D features to the SRM with 34,671-D features, and now to deep learning networks with complex structures and large parameter sizes, the performance of steganalysis has progressively improved alongside advancements in steganography techniques. Currently, deep learning-based methods have surpassed handcrafted-based methods, becoming the mainstream in steganalysis. However, in the pursuit of higher accuracy, models have become increasingly complex and redundant. To address this issue, Tan *et al.* proposed two specialized steganalysis model compression frameworks, CALPA-Net [146] and STD-Net [145], aimed at reducing model redundancy while maintaining high accuracy.

The above content briefly reviews the history and key developments in steganography and steganalysis. For a deeper exploration, this survey will analyze and discuss these fields in detail, following the pipeline outlined in Figure 2. The subsequent sections are arranged as follows: Section 2 outlines the preliminaries of image steganography. Sections 3 and 4 then present typical steganographic and steganalysis techniques, respectively. Section 5 explores the challenges currently faced in the fields of steganography and steganalysis. Concluding remarks are provided in Section 6.

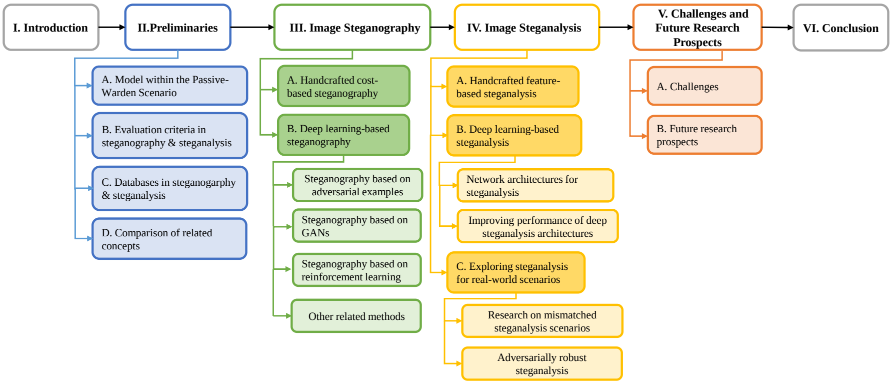

I. Introduction
II.Preliminaries
III. Image Steganography
IV. Image Steganalysis
V. Challenges and Future Research Prospects
VI. Conclusion
A. Model within the Passive-Warden Scenario
B. Evaluation criteria in steganography & steganalysis
C. Databases in steganogarphy & steganalysis
D. Comparison of related concepts
A. Handcrafted cost-based steganography
B. Deep learning-based steganography
Steganography based on adversarial examples
Steganography based on GANs
Steganography based on reinforcement learning
Other related methods
A. Handcrafted feature-based steganalysis
B. Deep learning-based steganalysis
Network architectures for steganalysis
Improving performance of deep steganalysis architectures
C. Exploring steganalysis for real-world scenarios
Research on mismatched steganalysis scenarios
Adversarially robust steganalysis
A. Challenges
B. Future research prospects

Figure 2: Survey pipeline. Zoom in to clearly view the titles of each section.

2 Preliminaries

In this section, we first introduce the model of image steganography within the passive-warden scenario. Next, we outline the key evaluation criteria for steganography and steganalysis within this scenario. Following this, we describe the image databases commonly used in steganography and steganalysis. Lastly, we present some concepts related to steganography, and provide a brief comparative analysis of these concepts.

2.1 Model Within the Passive-Warden Scenario

As illustrated in Figure 3, a common model used to explain the dynamics and interactions of steganography and steganalysis involves three characters: Alice, Wendy, and Bob. Alice is the sender who wants to communicate a secret message M to Bob. She uses steganography to embed this message within an innocuous-looking carrier X, such as an image, video, or audio file, using a secret key K. The objective is to keep the hidden message undetectable to anyone except the intended recipient. Wendy, serving as the adversary or steganalyst, uses various steganalyzers to detect the presence of hidden information in cover, or even disrupts their communications. Bob, the intended recipient, uses a corresponding decoding technique and key to extract the hidden message M' from the carrier file Y', which may be disrupted by the transmission channel or Wendy. The ongoing challenge between Alice's evolving steganography techniques and Wendy's efforts to detect them underscores a continuous dance of concealment and discovery. In their dynamic interplay, both steganography and steganalysis have experienced a spiral enhancement, constantly pushing the boundaries of each other's capabilities.

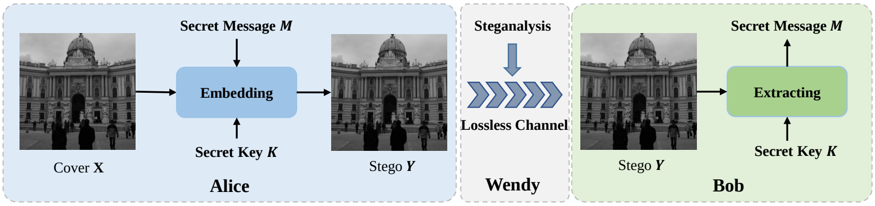


Figure 3: The relationship between image steganography via cover modification and steganalysis

In this paper, we focus on image steganography via cover modification[1] within the passive-warden scenario, where the transmission channel is lossless,

[1] Typically, there are three different fundamental architectures that determine the internal mechanism of the embedding and extraction algorithms in steganography: steganography by cover selection, by cover synthesis, and by cover modification [38]. Among these,

and the adversary, Wendy, solely monitors communications without making any modifications on Y, hence $Y = Y'$. In this scenario, the fundamental principle of steganography involves embedding a secret message M into a given cover image X by altering specific embedding units, such as pixel values in the spatial domain and DCT coefficients in the JPEG domain, according to predefined modification rules $Emb(\cdot)$ and a key K. This process results in a stego image Y containing the secret information as follows:

$$Y = Emb(X, M, K)$$

When Bob receives the resulting stego image $Y'(= Y)$ via lossless channel, he can recover the corresponding secret message using the extraction function $Ext(\cdot)$:

$$M' = Ext(Y', K) = Ext(Y, K) = M$$

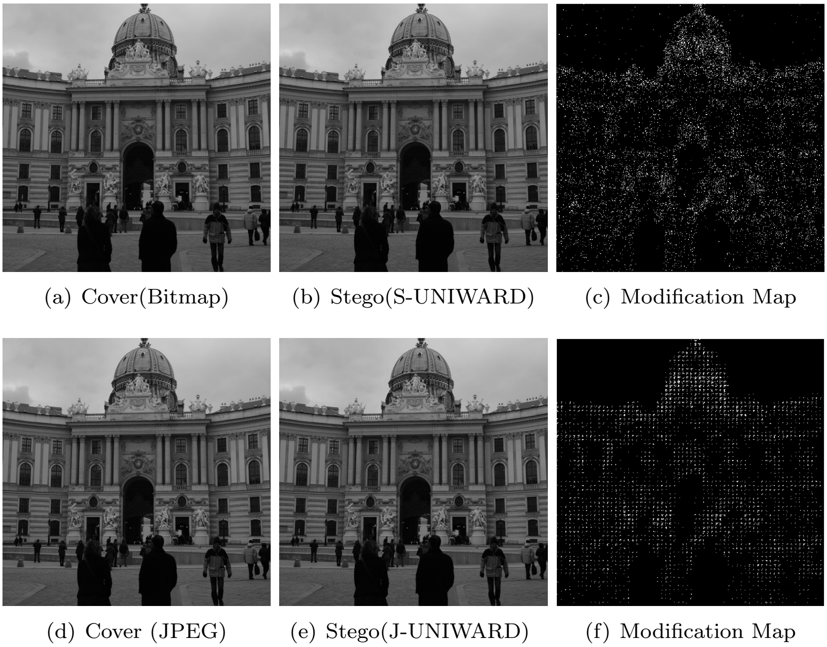
It is important to note that existing steganographic methods within the passive-warden scenario do not consider robustness metrics. If the channel experiences loss (i.e., $Y' \neq Y$), these methods cannot guarantee the correct extraction of the hidden message M at the receiver's end. In such cases, consideration of the active-warden scenario, which is beyond the scope of this paper, becomes necessary.

2.2 Evaluation Criteria in Steganography and Steganalysis

To ensure covert communication, steganography must guarantee that the cover X and the stego image Y are visually indistinguishable. To this end, the majority of steganographic algorithms make only minor modifications (typically ±1) to the embedding units. In addition, the extent of these modifications is relatively small, primarily determined by the amount of secret information to be embedded. Typically, the embedding payload is measured by metrics such as bpp (bits per pixel), bpc (bits per channel), and bpnzac (bits per non-zero AC coefficient). Moreover, modern steganographic methods often embed the secret information in regions of the image with relatively complex content. These methods therefore impose specific constraints on the amplitude, quantity, and location of modifications to maintain the fidelity of the stego image. Consequently, unlike other data hiding techniques, such as watermarking, most current steganography studies do not provide metrics like PSNR or SSIM since these values are generally high, as illustrated in Figure 4.

It is essential to recognize that, although it is difficult for the human eye to distinguish cover image and its corresponding stego image, the steganography modifications inevitably alter some inherent statistical characteristics within the cover image X. This change, particularly in the correlation properties

steganography primarily embeds secret information by modifying the embedding units of the cover; this approach is the most extensively studied paradigm in steganography today.



(a) Cover(Bitmap) (b) Stego(S-UNIWARD) (c) Modification Map

(d) Cover (JPEG) (e) Stego(J-UNIWARD) (f) Modification Map

Figure 4: Illustration of the cover image (i.e., (a) is a bitmap and (d) is a JPEG image of QF 90), stego image (i.e., (b) and (e) are the stego obtained by S-UNWARD and J-UNIWARD at 0.4 bpp/bpnzAC, respectively), and the corresponding modification map using two typical steganographic methods: S-UNIWARD in the spatial domain and J-UNIWARD in the JPEG domain. In the examples shown, the modification rates are 7.65% and 9.19%, respectively, with PSNR values of 59.26dB and 52.77dB before and after steganography.

among adjacent embedding units, introduces detectable artifacts. These alterations form the basis for detection by steganalysis methods, which typically analyze statistical differences in images before and after steganography using techniques such as statistical analysis, deep learning, and others to identify them. Consequently, steganalysis is essentially treated as a specialized binary classification problem, distinct from other image classification tasks due to the visually imperceptible nature of its objects (cover and corresponding stego). Steganalysis primarily focuses on detecting statistical differences caused by minor modifications to a small number of embedding units, presenting a binary classification challenge centered on weak signal detection.

Like other binary classification tasks, the detection accuracy P_A is widely regarded as the most crucial criterion for evaluating the performance of steganalysis (as well as steganography). It is defined as:

$$P_A = 1 - P_E \tag{1}$$

where P_E denotes the average detection error, further defined as:

$$P_E = (\frac{FP}{TN + FP} + \frac{FN}{TP + FN})/2 \tag{2}$$

Here, FP refers to the number of cover images misclassified as stego images, TN represents the number of cover images correctly classified as cover images, FN indicates the number of stego images misclassified as cover images, and TP refers to the number of stego images correctly classified as stego images.

For steganalysis, a higher detection accuracy P_A signifies a stronger discriminative ability of the steganalysis method. Conversely, for steganography, a lower detection accuracy P_A (equivalent to a higher detection error P_E) indicates a higher security performance of the steganographic method.[2]

The embedding payload is the most important factor that affects the detection error P_E. Typically, as the embedding payload increases, the detection error P_E correspondingly decreases. According to experimental results in existing literatures, when the embedding payload exceeds 0.40 bpp/bpnzac/bpc, the detection errors P_E evaluated on the mainstream databases BOSSBase [1] and BOWS2 [2] for most existing steganaographic methods become relatively low (less than 10%) as detected by the current best steganalysis methods, indicating that these steganographic methods are no longer secure. Therefore, under the passive-warden scenario, high embedding capacities are typically not pursued, generally staying at 0.4 bpp/bpnzac/bpc or less. At this embedding rate, imperceptibility can be maintained, so the focus is more on the steganographic security.

2.3 Databases in Steganography and Steganalysis

Standard image databases are essential for testing and comparing the performance of emerging algorithms, ensuring that advancements in the field are measurable and reproducible. Table 1 provides a brief overview of the databases commonly used in steganography and steganalysis.

Prior to 2018, the datasets predominantly utilized for steganography and steganalysis were BOWS2 [2] and BOSSBase [1]. Both datasets consist of 10,000 grayscale images with dimensions of 512×512, covering a variety of image types such as life scenes, scenic spots, and buildings. BOWS2 dataset was developed for a watermarking competition. BOSSBase, created by B.

[2] In addition to detection accuracy (error), several other evaluation criteria are used, including wAUC (weighted Area Under the Curve of the Receiver Operating Characteristic) [12], MD5 (Missed Detection at a 5% False Alarm Rate) [20], and FP50 (False Positive Rate at 50% Missed Detection) [20].

[3] Data available at: http://bows2.ec-lille.fr/

[4] Data available at: http://agents.fel.cvut.cz/boss

[5] Data available at: https://alaska.utt.fr/

[6] Data available at: https://github.com/YangzlTHU/IStego100K

Table 1: Datasets commonly used in steganography and steganalysis research, along with their specific details

Dataset	Number	Sizes	Type	Format	Year
BOWS2 [3]	10,000	512×512	Grayscale	PGM	2007
BOSSBase [4]	10,000	512×512	Grayscale	PGM	2011
ALASKA II [5]	80,005	17 different sizes	Grayscale & Color	PGM & PPM & JPG	2019
IStego100K [6]	208,104	1024×1024	Color	JPG	2019

Patrick *et al.*, was specifically designed for the HUGO steganalysis competition. We need to note that these fixed-size images are derived from full-resolution RAW data through subsequent processing steps such as image demosaicking, conversion to 8-bit grayscale, downsampling, and center-cropping. These post-processing steps, especially with various downsampling methods, can leave significant artifacts, impacting steganographic security [73]. We should keep in mind that these public databases have different statistical properties of neighboring pixels compared to natural images.

With the advancement of steganalysis, especially those involving deep learning techniques, a small number of image samples are often insufficient. Deep learning models require large datasets to effectively learn complex steganographic patterns and prevent overfitting. In 2019, two larger image databases were constructed to meet these needs: ALASKA II [17] and IStego100K [190]. ALASKA II, created by Remi *et al.* for the Steganalysis Challenge, aims to provide a large and diverse dataset of photographic images to bridge the gap between laboratory research and real-world applications. This dataset primarily includes images of landscapes, buildings, and everyday scenes, and features 17 different image sizes. It contains 80,005 images in each of the sizes 256×256 and 512×512, and also includes a set of 80,005 images in various sizes ($M = N$, both M and N are in $\{512,640,720,1024\}$, where M and N denote the width and height of the images, respectively) to facilitate the analysis of steganography in images of arbitrary sizes. Meanwhile, Yang *et al.* developed and released IStego100K, a dataset containing over 200,000 images with dimensions of 1024×1024. This collection includes stego images obtained using various steganographic methods and parameters, such as embedding payload and quality factor. The dataset is designed to promote the development of universal steganalysis techniques. Subsequently, there are a number of related steganalysis methods [64, 48, 121] using this dataset for experiments.

Furthermore, some research teams have improved the performance of steganographic and steganalysis methods by constructing their own image databases or utilizing established databases from the field of computer vision. For instance, Tang *et al.* [154] developed proprietary database SZUBase (not available online), comprising 40,000 full-resolution raw images captured by various cameras. Similarly, Butora *et al.* [12] selected 896,357 images of 256×256 size from the ImageNet dataset[7] to investigate the effect of pre-training model.

2.4 Comparison of Related Concepts

In recent years, some steganographic techniques have emerged that differ from the research scenario of this paper, such as generative steganography [224, 100, 225] and robust steganography [218, 87, 207]. To clearly differentiate from these techniques, we refer to our research scenario as "Dominant Steganography" in this section. This term highlights that, based on a rough estimate from the Web of Science, over 90% of current steganography research focuses on the scenario examined in this paper (i.e., steganography by cover modification within the passive-warden scenario), which significantly overshadows other types of steganography. Additionally, concepts often confused with steganography include watermarking [22, 38]. The collective relationship among these research directions is depicted in Figure 5. It is important to recognize that these areas share significant overlap and utilize similar technical methods. However, they differ fundamentally in their philosophical bases, which influence their requirements and, consequently, the design of their solution techniques. In the following, we will briefly distinguish between these concepts.

- **Information Hiding:** Information hiding (or data hiding) is a broad concept that involves concealing any form of information within a host medium, making it undetectable or inaccessible to unauthorized parties. Its goals are to maintain the confidentiality and integrity of information across diverse applications. Therefore, both watermarking and various forms of steganography can be considered as techniques of information hiding tailored for specific applications.

- **Watermarking:** Watermarking involves embedding markers in multimedia content for purposes such as copyright protection, content authentication, and traceability, among others. Depending on the specific application requirements, watermarking techniques can be categorized into robust, fragile, and semi-fragile watermarks, and so on.

- **Dominant Steganography:** Dominant steganography involves embedding a secret message within a cover by altering the embedding units in a way that is imperceptible, thereby facilitating covert communication. Its

[7] Available at: https://image-net.org/download.php

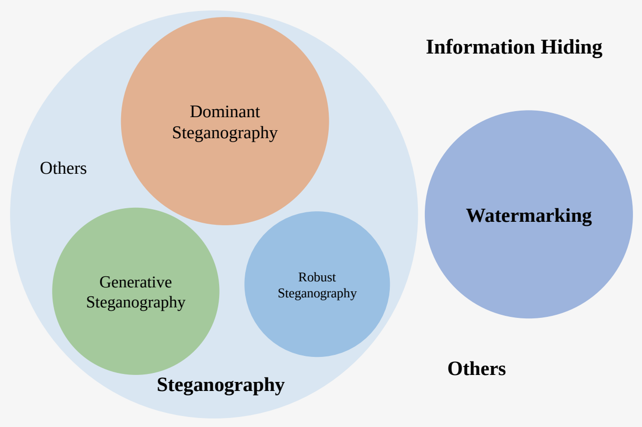


Figure 5: Venn diagram illustrating the relationships among information hiding, watermarking, and steganographic techniques, including dominant steganography, generative steganography, and robust steganography.

design requirements emphasize security, specifically by evading detection from existing steganalysis tools. This is the main research scenario of this paper.

- **Generative Steganography:** Generative steganography, a type of steganography by cover synthesis, primarily employs deep learning methods such as Generative Adversarial Networks (GANs) and diffusion models to create carriers that inherently contain the hidden information. To maintain security, it is crucial that images generated with and without secret information appear to belong to the same probability distribution. Research shows that there are significant statistical differences between Artificial Intelligence Generated Content (AIGC) and natural images captured by devices like cameras or cellphones. Detection algorithms, such as those proposed by Xi *et al.* [176] Wang *et al.* [160], and Luo *et al.* [108], can effectively distinguish between these types of images. Therefore, generative steganography typically evaluates the security by comparing generated images with and without secret messages rather than comparing with natural images.

- **Robust Steganography:** Robust steganography is somewhat akin to robust watermarking in that the technique aims to fortify steganographic methods against common multimedia processing operations such as JPEG compression, making the hidden information resilient to destruction. Additionally, unlike robust watermarking, robust steganography also needs to consider security against steganalysis. Current reports suggest that, due to the inclusion of robustness requirements, its security performance is significantly lower than of dominant steganographic methods at the same embedding capacity, especially evaluated on the modern deep learning-based steganalyzers.

3 Image Steganography

As previously introduced, modern image steganography predominantly follows the minimum distortion framework [39]. In this framework, the steganographic design methodology comprises two primary steps: designing the embedding cost function and implementing message embedding using specific encoding techniques, as illustrated in Figure 6. Over the past decade, significant research in image steganography has focused on the first step, with efforts concentrated on developing increasingly complex and advanced cost functions to enhance steganographic security. The existing methods for measuring distortion (or cost) can be divided into additive and non-additive categories. Additive distortions assume independence among the embedding units in an image, meaning the overall distortion of the cover image equals the sum of the individual embedding unit costs. In contrast, non-additive distortions, such as [33, 79, 26], account for interactions among embedding units. In practice, cover images are typically segmented into disjoint sublattices, with each calculating distortions and embedding secret information independently. This segmentation allows for cumulative distortion from all sublattices, as changes in one sublattice can affect subsequent calculations in neighboring areas. Currently, additive distortions are the most prevalent. Once the costs are defined, the second step commonly utilizes existing encoding techniques, particularly Syndrome-Trellis Codes (STC), for information embedding. Based on the methods used to design the embedding costs, steganography methods are generally classified into two categories: handcrafted cost-based and deep learning-based approaches. The subsequent sections will explore those representative methods from each of these categories.

3.1 Handcrafted Cost-based Steganography

Handcrafted cost-based steganography typically follows the steps illustrated in Figure 7 for embedding cost design. Firstly, the statistical features S of the cover image X are obtained through methods such as variance estimation and residual extraction. These features reflect the texture and complexity characteristics of each embedding unit within the cover image. Based on these characteristics, a distortion for each embedding unit is calculated according to specific rules, resulting in the embedding cost ρ. This process underscores

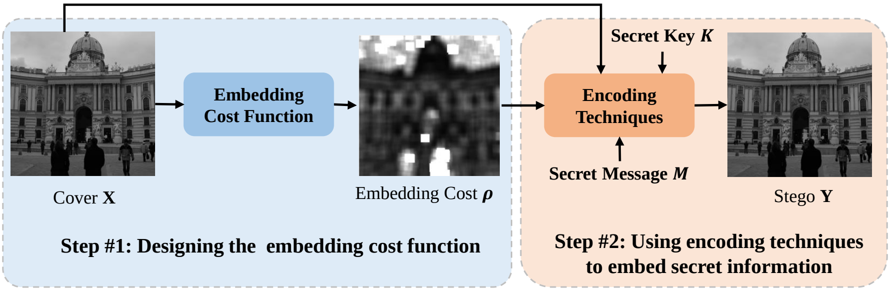


Figure 6: Illustration of the minimum distortion framework for image steganography in spatial domain.

the heavy reliance on manually set rules in this type of steganography. For example, the choice of model for estimating local variance in an image, the selection of high-frequency filters and their parameters for residual extraction, and the method of defining the distortion for an embedding unit based on the statistical feature S are all manual factors that significantly influence the steganographic security. Thus, different embedding cost functions can be obtained based on various experiential settings.

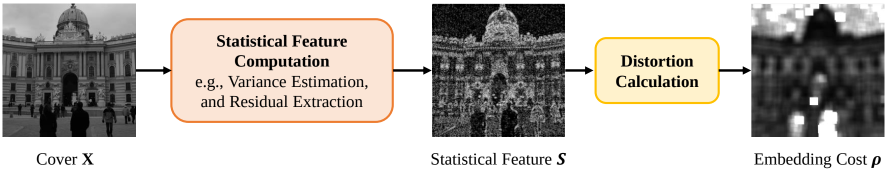

Cover **X**
Statistical Feature Computation
e.g., Variance Estimation, and Residual Extraction
Statistical Feature ***S***
Distortion Calculation
Embedding Cost ***ρ***

Figure 7: Illustration of the process of embedding cost design using a handcrafted based steganography.

In spatial domain, HUGO [120] is the first steganography method within the minimal distortion framework, and it significantly outperforms previous LSB based steganography. WOW [53] used the directional high-pass filtering to design the cost function. Building upon the WOW, HILL [78] constructed a high-pass filter and two low-pass filters to calculate the costs, making their distortion distribution more concentrated. UNIWARD [56] defined the embedding distortion as a sum of relative changes of coefficients calculated by a directional filter bank, which can be applied in an arbitary domain, with its spatial domain version being S-UNIWARD. MiPOD [132] introduced an additive distortion measurement that ensures limited detectability under a multivariate generalized Gaussian model. Chen *et al.* [15] tried to magnify the cover image first to highlight fine details before calculating the distortion with exsiting steganographic methods. CPP [223] combined various steganographic methods, redefining controversial pixels to improve security performance. Filler *et al.* [33] proposed an non-additive framework based on the Gibbs sampler to search for the optimal embedding schemes. To be specific, it divides the image into many disjoint sub-images and then embeds multiple times iteratively with additive distortion measurement on the sub-images, ensuring that the distortion satisfies a certain specific distribution. CMD [79] and Synch [26] both proposed the strategy of synchronizing modification directions when adjusting the initial distortion calculated from existing methods, which further enhance the security. DeJoin [215] improved the modification strategy of synchronous neighboring embedding by analyzing and defining the joint distortion between image pixel blocks. ASYMM [59] proposed a novel model-based steganographic scheme to reduce dependency on heuristic parameters. By applying the idea of Gibbs sampler, the adjacent embedding information is incorporated to optimally embed message into each sub-lattices. Su *et al.* [138] proposed a Gaussian Markov random field based model to capture the correlation of locally adjacent pixels, thus generating symmetric embedding probabilities. CMD-C [151] introduced the synchronous concept of CMD into color images, guiding pixels at the same positions in different color channels to be modified in the same direction, thereby capturing the correlation between color channels. ACMP [96] considered inter-channel correlations in color images, using a strategy that adjusts channel modification probabilities for payload distribution. GINA [159] aligned the R and B channel modifications with the G channel and adaptively distributes embedding capacity across all three channels.

For JPEG image steganography, the corresponding version of UNIWARD, known as J-UNIWARD [56], achieved high security at that time. UED [46] tried to ensure that embedding modifications are uniformly distributed across discrete quantized DCT coefficients, minimizing the changes in first and second-order statistics of image DCT coefficients after embedding. UERD [47] comprehensively considered factors such as block complexity and DCT coefficient mode to determine distortion, which further improves time complexity compared to UED. Wang *et al.* [162] proposed a cost function combining block fluctuation and quantization step. Wei *et al.* [168] proposed a cost function combining block residuals and quantization step, and it can effectively resist residual detection. J-MiPOD [19, 18] extended the MiPOD in spatial domain for JPEG image. DCDT [139] employed a generalized Distortion Cost Domain Transformation (DCDT) function to calculate the distortion cost. When equipped with HILL, DCDT outperforms other JPEG steganographic schemes such as UERD in resisting detection by GFR and SCA-GFR. Chen *et al.* [16] enhanced the cost function of JPEG steganography by leveraging microscale textures. Linear unsharp masking serves as the microscope, and an inter-block spreading rule is introduced to further enhance security. BBC [92] introduced non-additive JPEG steganography by suppressing block effects caused by inter-block modifications. BBM [158] enhanced existing additive JPEG steganography methods by controlling modification directions at different positions within DCT blocks to minimize the changes in DCT coefficients to spatial pixel values, and combining with the BBC method further enhances JPEG steganography. Wang *et al.* [161] adjusted the existing embedding cost using DCT block similarity and channel similarity, resulting in a significant improvement in security. Recently, Li *et al.* [90] first used a deblocking method to construct an optimal polar map selection strategy, then design a modulation method based on statistical models to better utilize the optimal polar map in the quantized Gaussian embedding model. Butora *et al.* [8] introduced a JPEG steganography method that utilizes side information to constrain the boundary of the likelihood ratio test for a decompressed JPEG image. This is achieved by minimizing the Kullback-Leibler divergence between the cover and stego distributions.

Unlike the process as illustrated in Figure 7 , Chen *et al.* [13, 14] proposed a method to minimize the residual distance between stego and cover images through a stego post-processing strategy, thereby enhancing steganography security. Additionally, Li *et al.* [89] proposed the ISteg algorithm based on artificial immune systems and immune evolution models, employing intelligent optimization search to post-process the cover-stego images generated by existing steganography algorithms to produce immune steganographic images that minimize the distance to cover image features, Subsequently, Li *et al.* [88] proposed an immune image steganography method that utilizes fuzzy enhancement and artificial immune systems to achieve adaptive enhancement of texture regions and edge regions. Ye *et al.* [192] presented a residual-guided learning method for image steganography.

In Table 2, we provide a summary of representative handcrafted cost-based image steganography methods.

Table 2: Summary of representative handcrafted cost-based image steganographic methods

Domain	**Method**	**Year**	**Highlight**
Spatial	HUGO	2010	First method to adopt minimal distortion framework
	WOW	2012	Use the directional high-pass filtering to design cost function
	HILL	2014	A high-pass filter and two low-pass filters to calculate the costs
JPEG	J-UNIWARD	2014	Modification of positions with small perturbations in the distributional properties of JPEG coefficients
	UERD	2015	Block complexity and SCT coefficient mode to determine distortion
	J-MiPOD	2020	Extend the MiPOD for JPEG image

3.2 Deep Learning-based Steganography

Recently, deep learning technology has significantly impacted various fields of artificial intelligence, including image steganography. Positioned at the intersection of image processing and pattern recognition, image steganography has seamlessly integrated into this technological wave. Unlike traditional handcrafted techniques, methods based on deep learning rely less on manual expertise and more on data-driven processes, enabling autonomous derivation of embedding cost functions. Consequently, steganography utilizing deep learning has become the dominant research direction in this field, with some modern methods surpassing traditional ones to achieve state-of-the-art security. Currently, the derivation of cost functions primarily utilizes three approaches: adversarial examples, GANs, and deep reinforcement learning, which will be explored in subsequent sections. Finally, we will introduce some recent deep learning-based information hiding techniques developed in recent years that may be employed in image steganography to enhance security.

3.2.1 Steganography Based on Adversarial Examples

Currently, mainstream steganalysis methods rely on architectures based on convolutional neural networks (CNNs). Effectively evading detection by these CNN-based steganalyzers is key to enhancing the security of modern image steganography algorithms. Szegedy *et al.* [141] discovered that making imperceptible changes to input samples can lead neural networks to output incorrect classification results with high confidence. These intentionally modified samples, aimed at attacking neural networks, are known as adversarial examples. Numerous studies have indicated that adversarial examples have generalization properties, meaning that an example generated for a specific neural network (targeted network) can also affect other neural networks or machine learning models that rely on manual features to some extent. Adversarial examples expose the vulnerability of existing machine learning models and also provide new avenues for enhancing the security of steganography methods by reducing the discriminative ability of steganalysis tools. In the following, we first present the basic framework of steganography based on adversarial examples, and then highlight some typical works in this area.

Basic Framework: The basic framework for steganography using adversarial samples is illustrated in Figure 8. This framework consists of two main steps: pre-training a steganalyzer and adversarial adjustment. In step #1, we begin by collecting a cover set C_p and then use an existing steganographic method S_{init} to obtain the original embedding cost set ρ_{C_p}. Subsequently, we generate the corresponding stego set S_p using STC. Based on the cover set C_p and stego set S_p, we train a deep learning-based steganalyzer (also referred to as a targeted steganalyzer). In step #2, for each cover image X (not included in C_p), we first extract the original embedding cost ρ using the existing steganography method S_{init}. We then use the resulting targeted steganalyzer in step #1 to guide the adjustment of the embedding cost ρ, resulting in a modified cost ρ'. Finally, the stego image Y can be generated using the adjusted embedding cost. It is evident that steganography based on adversarial examples essentially involves enhancing an existing steganographic method. This enhancement is achieved by adjusting the existing embedding costs using a pre-trained deep neural network steganalyzer. The main difference among most related algorithms lies in the strategies used to adjust the embedding cost in the second step. Additionally, the choice and number of targeted steganalyzers, as well as the specific initial steganography method to be enhanced in the first step, significantly impact the security of the final steganographic method.

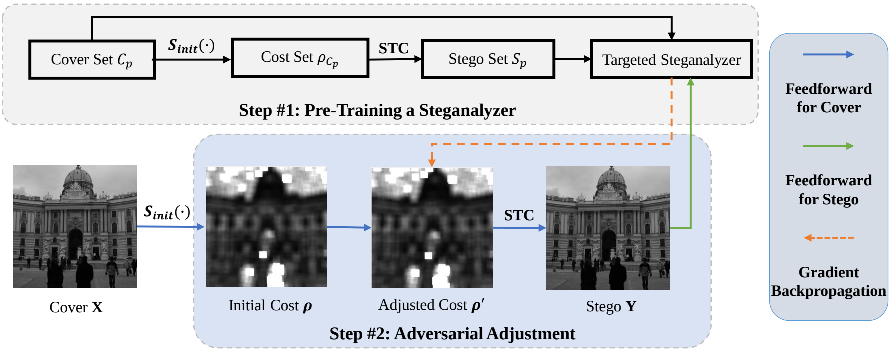


Figure 8: Diagram of the basic framework for image steganography using adversarial examples

It should be noted that due to the vulnerabilities inherent in deep neural networks, adjusting existing embedding costs can effectively counter the targeted steganalyzer used in the first step. Therefore, when evaluating steganographic methods based on adversarial examples, it is necessary to either retrain the targeted steganalyzer or employ other steganalysis models to analyze the security performance.

Typical Works: Ma *et al.* [110] introduced the adversarial example technique into image steganography. This method selectively modifies the flipping direction of binary STC embedding units based on the gradients of the target neural network with respect to the cover image, generating adversarial examples. Experimental results demonstrate that this method effectively deceives the targeted steganalysis network, although it becomes ineffective once the targeted network is retrained. Building on this, Ma *et al.* [109] further developed the technique by modifying the softmax output probabilities to specific vector values and adjusting the embedding cost of existing steganography methods based on the gradient of the cover image according to these vector values. This expansion of the adversarial example technique applies to ternary STC embedding. Zhang *et al.* [219] iteratively utilized adversarial example techniques and controlled the intensity of adversarial noise to construct enhanced cover images. These enhanced cover images using existing cost functions for embedding, and possess higher security than the original cover images. CAAS [105] is also a method for enhancing cover images against adversarial attacks. It adaptively generates perturbations based on the texture information, gradients, and image segmentation methods of the cover image, resulting in an enhanced cover image and improving the security of steganography. Existing adversarial example-based steganography methods enhance existing manually crafted cost functions to achieve higher security than the original methods. ADV-EMB [152] randomly divided embedding units and secret information into two disjoint parts in the same proportion. After embedding partial secret information into the first part of embedding units with original embedding costs, it adjusts the embedding costs of the rest embedding units and facilitates the embedding of the remaining secret information. The division ratio is an important parameter in ADV-EMB, and the parameter is tested from small to large until the generated cover image can deceive the targeted steganalyzer. AEN [111] first removed the checkerboard pattern of the gradient of the cover image and normalizes the processed gradient of the cover image, thereby constructing an adaptive adversarial modification embedding cost intensity based on this normalized cover gradient. However, this method has limited improvement on spatial steganography security and cannot be applied to JPEG image steganography. MAE [98] proposed to use a mixed gradient of the cover image and its corresponding multiple stego images to determine the modification direction of the embedding cost, and carefully select the embedding cost to be modified based on the amplitude of the gradients and the original embedding cost. This method can effectively enhance the security performance of existing spatial steganography methods by modifying less than 6% of the embedding cost. SGS [135] proposed a new adversarial embedding framework for stego image generation and selection. This method first generates multiple candidate stego images by randomly adjusting the original embedding cost and then selects the final stego image by minimizing the high-pass residual distance between the cover image and the stego image. The method can enhance the security of spatial steganography methods. USGS [99] proposed a stego image generation and selection method based on the adversarial embedding framework. This method can generate more diverse candidate stego images and simultaneously use high-level features (the ability to deceive targeted models) and low-level features (minimizing residual distance) of steganalysis to select the final stego image. USGS can enhance both spatial steganography and JPEG steganography, and its enhancement effect is much better than that of SGS. ESGS [82] improved the cover image selection method based on the SGS method. It selects the final stego image by minimizing the steganalysis feature (i.e., SRM) distance between the stego image and the cover image. This method can enhance the security of spatial steganography. ITE-SYN [127] initially embedded using CMD, then randomly selects a sub-image at each iteration to adversarially adjust the cost and re-embed until the resulting stego image can successfully deceive the targeted discriminator. Additionally, the strength of perturbation gradually increases with each iteration, ensuring minimization of adversarial perturbation strength. Backpack [4] iteratively attacks the targeted steganalyzer, updating embedding cost using gradient descent and employing a min-max strategy to find the optimal stego image during iterations. JAS [31] first computed joint costs and adjusts costs based on joint gradients feedback from the targeted steganalyzer, enhancing the security of existing steganography methods based on joint costs. Xie *et al.* [178] mapped the original embedding costs to 0 and 1, thereby dividing the cover image into embedding and non-embedding regions. They further adjust the costs of embedding regions based on gradients, resulting in new embedding costs. This method greatly enhances steganography security while maintaining low computational complexity. GEAP [126] introduced the adversarial embedding mechanism into color image steganography by changing the adversarial loss of each sub-image color pixel vector, effectively resisting targeted steganalyzers based on deep networks. Min-max [5, 3] proposed a new adversarial embedding iterative strategy based on ADV-EMB. During the iterative process, this method uses a min-max strategy to select the appropriate stego image, i.e., selecting the stego image that is the most difficult to be detected under the strongest steganalyzer. JS-IAE [115] proposed a new adversarial embedding iterative strategy. In each round, this method carefully selects a portion of embedding cost to be modified, and the modification is based on the embedding cost of the previous round rather than the original embedding cost in each round as in ADV-EMB.

3.2.2 Steganography Based on GANs

Generative Adversarial Networks (GANs) [43] constitute an innovative deep learning framework involving two neural networks: a generator and a discriminator. These networks engage in a game-like scenario where the generator aims to create fake data that appears real, and the discriminator strives to distinguish between real and generated data. This adversarial interaction enhances the quality of the generated outputs, making GANs highly effective for tasks like generating realistic images, enhancing photos, and creating art. The adversarial game between the generator and discriminator in GANs parallels the interaction between steganography and steganalysis. Modeling steganography and steganalysis using neural networks could leverage the GAN mechanism to enhance the security performance of steganographic methods (i.e., improving their ability to evade detection by steganalysis discriminators). Unlike steganography based on adversarial examples, which heavily depends on existing steganography methods, GANs offer a fresh perspective on image steganography, capable of learning embedding probabilities from scratch. Although research in GAN-based image steganography is still in its infancy and limited, its potential has already garnered significant attention. In the following sections, we will first outline the basic framework of steganography utilizing GANs, and then explore some key works in this field.

Basic Framework: The basic framework for steganography using GANs is illustrated in Figure 9. This framework includes two main components: a steganographic generator and a steganalysis discriminator. The generator aims to generate a simulated stego image Y from an input cover image X. It is important to note that to minimize the steganography modifications and to utilize STC for actual information embedding, the generator does not generate the stego Y directly from the cover X. Instead, it employs a probabilistic generation network, often structured like U-Net, to transform the cover X into an embedding probability map P (rather than embedding cost ρ). This map is then processed by an STC simulator, which simulates information embedding to produce an embedding modification map M. This map M is then superimposed on the input cover image X to form the stego image Y. On the other side of the GAN, the discriminator is tasked with distinguishing between the generator-produced stego image Y and the input cover image X. Given the availability of numerous effective CNN-based steganalysis networks in steganalysis, these networks are often adapted for and integrated directly into the discriminator. Consequently, most research based on GANs focuses on the design of the generator, especially the probabilistic generation network aspect.

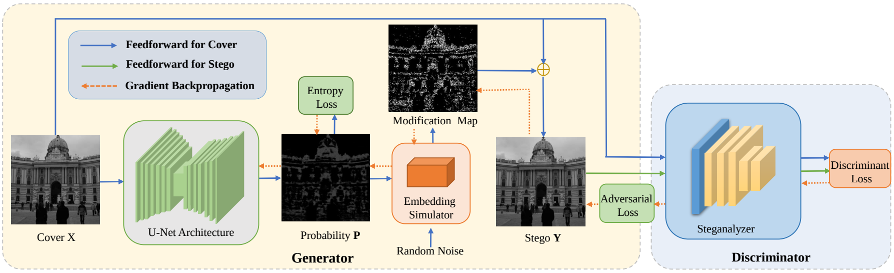


Figure 9: Generative adversarial network-based image steganography framework in the spatial domain.

Once the training of the GAN is complete, the trained generator is used to generate the probability map P for each input cover. This map P is then transformed into the embedding cost ρ based on the following formula. Finally, STC is used to embed the secret message, resulting in the final stego image.

$$\begin{cases} \rho_{i,j}^{+} = ln(1/(p_{i,j}^{+} + \epsilon) - 2), \\ \rho_{i,j}^{-} = ln(1/(p_{i,j}^{-} + \epsilon) - 2), \\ \rho_{i,j}^{0} = 0. \end{cases} \tag{3}$$

where $p_{i,j}^{+}$ and $p_{i,j}^{-}$ ($p_{i,j}^{+} = p_{i,j}^{-} = p_{i,j}/2$ for symmetric scenarios) denote the embedding probabilities for modifications of ±1 at the (i,j)-th position of the probability map P, respectively. Meanwhile, $\rho_{i,j}^{+}$, $\rho_{i,j}^{-}$, and $\rho_{i,j}^{0}$ represent the corresponding embedding costs for modifications by $+1$, -1, and 0. Additionally, a small value, ϵ, is added to prevent division by zero, typically set to 10^{-5}.

Typical Works: ASDL-GAN [154] is the first method to apply GANs to image steganography in spatial domain. Its generator comprises a 5×5 high-pass filter and 25 layers of 7×7 convolutional filters. The discriminator features a 5×5 high-pass filter followed by the steganalyzer of Xu-Net [181]. Experimental results demonstrate that through iterative learning, ASDL-GAN increasingly directs steganographic modifications towards areas of the image with more complex textures. However, the security performance of the resulting stego images still does not match that of the traditional S-UNIWARD. Expanding on ASDL-GAN, Yang *et al.* developed UT-GAN [187], which incorporates a U-Net-based generator to produce embedding probability maps and includes a TanH-simulator for approximating the STC embedding process. Additionally, the use of multiple high-pass filters in the discriminator has been implemented. Experimental results indicate that UT-GAN surpasses the traditional S-UNIWARD in terms of security. Wu *et al.* [171] enhanced the UT-GAN generator by integrating multiple intermediate feature maps into the U-Net architecture and increasing the maximum feature channel count to 256. This modification enables the creation of more complex embedding probability maps, resulting in improved security performance over both ASDL-GAN and the original UT-GAN. Li *et al.* [81] further refined the UT-GAN generator by introducing cross-feedback channels between the upsampling and downsampling segments of the network. This enhancement allows the detailed information gradients to be better integrated throughout the model, enhancing the quality and security of the steganography. Unlike previous methods that utilize a single steganalyzer in the discriminator, Steg-GMAN [60] employed multiple steganalyzers and adopts an adaptive strategy for updating model parameters. During the training phase, each iteration focuses on updating the weakest steganalyzer in the discriminator and utilizing the gradient from the strongest steganalyzer to refine the generator. This approach strives to maintain a balance between the generator and discriminator, enriching the gradient information supplied to the generator and significantly boosting steganographic security. Building on Steg-GMAN, Huang *et al.* [61] introduced a new architecture featuring an asymmetric dual-branch generator network, along with an innovative adversarial loss function. This loss function prompts the generator to incorporate features such as image residuals, embedding probability graphs, and gradient symbols. The aim is to effectively target textured regions within cover images and employ gradient symbols to confront the discriminator, thus promoting asymmetric embedding cost learning and achieving the highest level of security in the spatial domain. JS-GAN [188] represented the first GAN-based steganography for JPEG images, facilitating end-to-end training through the integration of an IDCT module and a simulated embedding module. However, its security performance is somewhat inferior to traditional JPEG steganography methods. Building on JS-GAN, JS-GAN (ESI) [185] employed a CNN network to estimate the original spatial domain image, thereby acquiring estimated edge information. Utilizing this information to tailor the learned embedding costs asymmetrically, it achieves superior security compared to J-UNIWARD.

3.2.3 Steganography Based on Reinforcement Learning

Reinforcement learning (RL) is a distinct machine learning paradigm, differing from supervised and unsupervised learning. Central to RL is an agent that interacts with an environment to make decisions by observing its state, taking action, and receiving a reward. This process repeats across multiple rounds, allowing the agent to accumulate a sequence of rewards. The objective of RL is to devise an optimal policy that maximizes the total reward obtained from the environment. RL has demonstrated its effectiveness in exploring policies through a trial-and-error approach and has been applied in various fields, including gaming, robotics, and autonomous driving. Despite its rapid advancement in these areas, the application of RL in steganography remains relatively undeveloped. In this section, we begin by outlining the basic RL framework applied to steganography, followed by a discussion of some typical works in this field.

Basic Framework: The basic framework for steganography with RL is illustrated in Figure 10. In this framework, the agent represents the steganographer, while the environment represents the steganalyst. The agent's goal is to learn the optimal policy for embedding secret messages to maximize the reward. The environment assigns rewards based on the security performance of the generated stego image, linking the reward to factors such as the ability to counter steganalyzers and preserve image texture. Under this RL framework, the agent dynamically updates the steganographic policy through a trial-and-error process until a secure steganographic embedding policy is achieved.

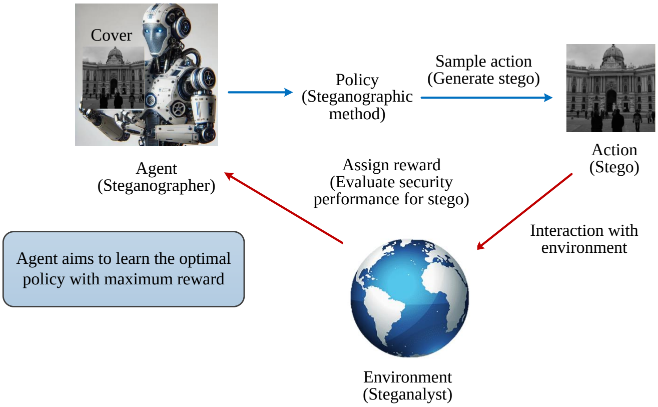


Figure 10: Basic framework for image steganography using reinforcement learning.

Under this RL framework, two essential properties can be employed to learn embedding costs. First, the non-differentiable mechanism can be utilized to construct a sampling-based simulator. This simulator can replace the neural network-based simulator in GANs, generating discrete modifications during the feedforward process and propagating significant gradients during the backpropagation process. In this context, the agent adopts a one-step Markov Decision Process (MDP), with the policy being updated by policy gradients [140]. Second, the framework facilitates the processing of sequential signals. Notably, the one-step MDP can be extended to a multi-step MDP, which is capable of formulating sequential signals. Here, the agent sequentially modifies the embedding costs across multiple steps to achieve optimal security performance. In this case, the Asynchronous Advantage Actor-Critic (A3C) framework [114] is utilized to learn the policy.

Typical Works: Tang *et al.* [148] proposed a cost learning method called SPAR-RL based on existing GAN-based steganography methods. This method treats the steganographic generation network and steganalysis network in GAN steganography as the policy network and environment network, and replaces the neural network-based simulator in GAN with the sampling-based simulator in RL. Experimental results show that this method can enhance the security performance of existing GAN-based steganography methods. JoPoL [150] utilized SPAR-RL to generate initial embedding probabilities, then combines probabilities within neighbouring units with the attention module to capture the correlation of embedding units within blocks, thus generating joint embedding probabilities for non-additive steganography. Subsequently, Tang *et al.* [149] extended the spatial SPAR-RL method to JPEG steganography, which can simultaneously extract both inter-block and intra-block features for DCT coefficients. PICO-RL [91] directly learned embedding costs based on the SPAR-RL structure and applies the optimal probability approximation module to obtain embedding probabilities for different embedding rates, enabling efficient end-to-end training and further improving the model universality. RLAE [104] adjusted the embedding costs of existing steganography methods based on reinforcement learning principles. Both the pre-trained steganalyzer and image residuals are utilized to calculate the reward function. Mo *et al.* [116] proposed MCTSteg, which formulates the process of adjusting embedding costs for different embedding units as MDP, and adjusts their embedding costs by Monte Carlo Tree Search. Subsequently, they propose an A3C-based method called ReLOAD [117], wherein the asymmetric additive distortion is optimized for minimizing the embedding effects on image textures. Results show that this additive method can even outperform non-additive methods.

In Table 3, we provide a summary of representative deep learning-based image steganography methods. Furthermore, we have provided a comparison of the parameters and FLOPs (floating point operations) for typical deep learning-based steganography methods, as shown in Table 4.

3.2.4 Other Related Methods

Unlike the basic framework shown in Figure 9, some other related works utilizing GANs bypass learning embedding probabilities and instead directly focus on learning information embedders and extractors. For instance, Hayes *et al.* [50] first designed steganographic network models with steganographic information encoders, steganographic information decoders, and steganalyzer, allowing the network to perform multitask learning under the goal of accurately reconstructing secret information and adversarial steganalysis, enabling the network to directly output cover images and extract secret information. Zhu *et al.* [226] also used a similar structure to implement image information hiding, introducing noise layers during training to make secret information extraction robust to operations such as Gaussian blur and JPEG compression. Zhang *et al.* [213] proposed ISGAN, hiding grayscale images in color images and generating steganographic images with semantic and color similarity to the cover image and enhancing security through adversarial training. Zheng *et al.* [221] proposed a composite perceptual steganography method, integrating rule-based combination methods and generative adversarial networks to synthesize more natural steganographic images and achieving better performance than ISGAN. Zhang *et al.* [211] proposed SteganoGAN, a high-capacity steganography based on GAN. The encoder proposed residual and dense structures to improve the visual quality of the cover image and the accuracy of information extraction. This method achieves a maximum embedding rate of 4.4 bpp,

Table 3: Summary of representative deep learning-based image steganographic methods

Domain	**Method**	**Year**	**Highlight**	**Categories**
Spatial	ASDL-GAN	2017	The first method to apply GANs to image steganography	GAN
	UT-GAN	2019	U-Net-based generator, TanH-simulator and multiple high-pass filters in the discriminator	GAN
	SPAR-RL	2020	Pixel-level rewards	Reinforcement learning
	MAE	2021	Mixed gradient of the cover and stego	Adversarial example
	Steg-GMAN	2023	Multiple steganalyzers and an adaptive update strategy	GAN
	asym Steg-GMAN	2023	Asymmetric dual-branch generator and an innovative loss function	GAN
	ReLOAD	2023	Minimizing the impact of embedding on image texture	Reinforcement learning
JPEG	JS-GAN	2019	End-to-end training through an IDCT and a simulated embedding	GAN
	JS-GAN (ESI)	2021	Edge information estimation with a CNN	GAN
	JEC-RL	2021	Domain Transformation	Reinforcement learning
Both	ADV-EMB	2019	Division ratio for dividing the embedding units	Adversarial example
	Min-max	2020	Min-max strategy to select the appropriate stego image	Adversarial example
	MCTSteg	2021	Monte carlo tree search	Adversarial example
	Backpack	2022	Update embedding cost with gradient descent	Adversarial example
	USGS	2022	Stego generation	Adversarial example

and the visual quality and extraction accuracy of the cover image exhibit certain generalization performance on different databases. Tan *et al.* [142] introduced channel attention mechanisms into the generator and extractor to guide embedding and extraction on important channels, improving the quality and extraction accuracy of the cover image. Additionally, this method introduces error correction codes to reduce the error rate. Yuan *et al.* [204] adopted a structure similar to SteganoGAN and added an attack module to generate adversarial perturbations using a pre-trained network and added

Table 4: Summary of parameters and FLOPs for representative steganographic methods

Method	Parameters	FLOPs
ASDL-GAN	1.77×10^5	10.80×10^9
SPAR-RL	2.60×10^6	0.42×10^9
UT-GAN	2.60×10^6	0.42×10^9
MCTSteg	4.77×10^6	5.95×10^9
MAE	6.21×10^5	22.99×10^9
USGS	6.21×10^5	390.85×10^9
JoPoL	1.41×10^5	1090×10^9
ReLOAD	7.51×10^6	0.90×10^9
Steg-GMAN	7.64×10^6	208×10^9
Asym Steg-GMAN	1.83×10^7	535×10^9

them to the cover image to achieve the goal of adversarial attack. Additionally, the cover image is embedded with secret information using a per-pixel depth fusion method to improve the visual quality of the cover image. Based on diffusion model, Peng *et al.* [118] proposed a generative image steganography technique based on a denoising diffusion probability model to achieve large-capacity and distribution-preserving secret data hiding, showing outstanding performance in steganography capacity and extraction accuracy. Recently, Cui et al introduced meta learning into deep image hiding and proposed MSM-DIH [23]. Please note that all the methods mentioned above primarily function as forms of information hiding technology. They mainly emphasize enhancing the information embedding capacity and maintaining the visual quality of the stego image, rather than guaranteeing the complete extraction of secret information. Additionally, their security is generally considered weaker than that of traditional steganographic methods under the same embedding payload.

4 Image Steganalysis

As described in Section 2.2, the task of image steganalysis is a specialized binary classification problem aimed at distinguishing visually indistinguishable cover and stego images. Steganalysis detection is based on the principle that steganography inevitably disrupts the inherent statistical properties of the cover image, even though the steganographic modifications to the image embedding units are typically minor in scale and magnitude, thereby introducing potentially detectable artifacts. As discussed in Section 3, different steganographic methods vary significantly in their modes of modification, resulting in distinct statistical artifacts. Early non-content adaptive steganographic methods, such as those based on LSB replacement, introduced obvious tampering traces, enabling the design of effective specialized steganalysis techniques. However, with the development of content-adaptive steganography under the minimal distortion framework, the tampering traces have become relatively weaker, making it challenging to develop effective dedicated detection methods. Consequently, the features used in current steganalysis are often complex and high-dimensional, and they are universal for exploring the artifacts introduced by various steganography modifications. Regarding methodologies, steganalysis mainly includes traditional handcrafted feature-based methods and deep learning-based steganalysis methods. In the following, we will delve into an introduction and summary of these methodologies.

4.1 Handcrafted Feature-based Steganalysis

Since there are no significant visual differences between cover images and stego images (with visuals typically reflecting the mid-to-low frequency features of an image), it can be inferred that the artifacts introduced by steganography are predominantly manifested in the high-frequency components of the image. Therefore, current handcrafted feature-based steganalysis primarily focuses on extracting and analyzing these high-frequency components. The basic process of this analysis is illustrated in Figure 11. Initially, various high-pass filters (such as KV, KB, Sobel filters, etc.) transform the input image from the spatial domain into the image residual domain. This transformation helps suppress image content while retaining subtle steganographic signals, thereby enhancing the signal-to-noise ratio required for steganalysis. Subsequently, the image residuals often undergo quantization and truncation to constrain the range of residual feature values, which helps control the dimensions of the subsequent steganalysis features. The residual images are then subjected to feature extraction for steganalysis, primarily using statistical methods such as co-occurrence matrices, histograms, Markov chains, etc. Finally, machine learning models are employed to classify the effectively extracted steganalysis feature sets. Common classifiers include Fisher linear discriminant (FLD) [35], support vector machine (SVM)[21], ensemble classifiers [66], and others, for binary classification.

Subtractive Pixel Adjacency Matrix (SPAM) [119], introduced in 2010, is a notable steganalysis technique. This method constructs a Markov transition probability matrix using adjacent pixel residuals and employs a SVM to classify the extracted features. It has been particularly effective in detecting LSB matching steganography. In 2012, Fridrich *et al.* first introduced the spatial rich model (SRM) [37], which utilizes a plethora of different high-pass filters to extract high-dimensional noise residual features. The feature space

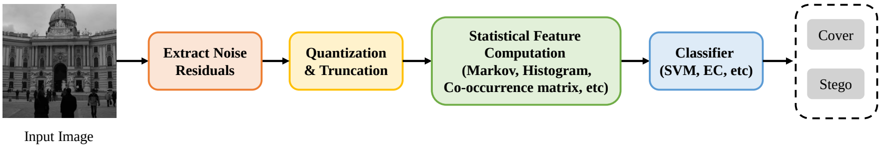

Input Image
Extract Noise Residuals
Quantization & Truncation
Statistical Feature Computation (Markov, Histogram, Co-occurrence matrix, etc)
Classifier (SVM, EC, etc)
Cover
Stego

Figure 11: Flowchart of the general processing steps in handcrafted feature-based steganalysis.

of SRM reaches up to 34,671 dimensions, significantly enhancing the diversity and detection capability of features. The introduction of rich model features has accelerated the development of steganalysis. Subsequently, numerous steganalysis methods based on rich models and methods with similar design principles have been proposed. Based on different embedding domains, manually crafted image steganalysis features are mainly divided into spatial domain features and JPEG domain features. Holub *et al.* [52] proposed the Projection Spatial Rich Model (PSRM), which locally and randomly projects elements in adjacent noise residuals and then calculates their histograms, to reduce the feature dimension. The standard LBP has demonstrated its versatility in performing image classification tasks, including texture analysis, object recognition, and steganalysis. Shi *et al.* [134] proposed LBP-based steganalysis features. Denemark *et al.* [25] introduced the maxSRMd2, a rich model that incorporates selection-channel knowledge. Tang *et al.* [153] proposed an adaptive steganalysis scheme for the WOW method, which focused on analyzing regions with high embedding costs using SRM-based features. Subsequently, they proposed an adaptive steganalysis scheme that assigns different weights to pixels based on their embedding probabilities, focusing on likely modified regions to improve detection accuracy. To address steganography based on synchronizing embedding changes such as CMD [79] and Synch [26], Tan *et al.* [147] proposed the pixel-decimation-assisted steganalysis feature set based on maxSRMd2. It not only reduces the synchronization of embedding changes in SEC (synchronize-Embedding-Changes) steganography but also improves the accuracy of estimating embedding change probabilities. Li *et al.* [77] proposed Threshold LBP (TLBP), which effectively extracts the traces left during steganographic embedding in the binarization process. Zhou *et al.* [222] introduced the αSRM steganalysis method, which uses KV high-pass filters to enhance the estimated embedding change probability. This probability is then weighted and combined with the quantized image residuals' estimated probability to compute co-occurrence features. Wang *et al.* [156] improved TLBP and SRM by using the Fisher discriminant criterion to measure the separability of spatial and frequency domain features. Ma *et al.* [112] addressed the high dimensionality of rich model steganalysis features by transforming it into a feature subset reduction problem, proposing a selection method based on decision rough sets that evaluates and retains features with high attribute separability for dimensionality reduction.

In JPEG domain, a common approach is to first decompress the JPEG image to the spatial domain and then use spatial rich models to extract and classify noise residual features. A few methods combine features extracted from both the spatial and JPEG domains for analysis. Kodovsky *et al.* [72] proposed JPEG Rich Model (JRM) features. Subsequently, they combined spatial and JPEG domain rich models to create the more performant JSRM. Holub *et al.* [54] introduced Discrete Cosine Transform Residual (DCTR) features, which extract residual features from JPEG images using 64 8×8 DCT bases. Later, Holub *et al.* [55] proposed the Phase Aware Projection Model (PHARM), which projects residuals onto vectors to calculate statistical features. Song *et al.* [136] introduced GFR, which use Gabor filters to obtain noise residual features in multiple directions. Qiao *et al.* [123] proposed a steganalysis algorithm based on an adaptive statistical model, building on the DCT channel weighting strategy. Using hypothesis testing theory and the distribution of quantized DCT coefficients, they developed a detector based on statistical models. Feng *et al.* [32] proposed a JPEG steganalysis method based on a cascade of diverse filters. To improve feature extraction speed, the cascade filters with the maximum diversity (MD-CFR) were selected. These chosen filters were convolved with decompressed JPEG images to obtain residuals that capture subtle embedding traces. Despite the excellent detection performance of JPEG phase-aware steganalysis features like DCTR and GFR against adaptive steganography, they use fixed-size DCT or Gabor filters to extract convolutional residuals, limiting their diversity. Xia *et al.* [177] introduced JPEG phase-aware features from residual-difference images, using various convolution filter sizes to generate residuals and calculating features from their differences. They designed symmetry rules to reduce feature dimensionality based on filter type, size, and residual, enhancing feature robustness.

Traditional handcrafted feature-based steganalysis heavily relies on empirical knowledge for feature extraction. Moreover, the independent design of feature extraction and classifiers complicates the synchronization of their optimization. These limitations hinder traditional methods from achieving higher detection performance, necessitating the development of new steganalysis approaches. In Table 5, we provide a summary of representative handcrafted feature-based image steganalysis methods.

4.2 Deap Learning-based Steganalysis

In recent years, the development of deep learning technologies has revolutionized the field of steganalysis. Deep learning methods reduce the need for excessive manual intervention typical in traditional steganalysis by leveraging

Table 5: Summary of representative handcrafted feature-based image steganalysis methods

Domain	Method	Year	Highlight
Spatial	SPAM	2010	Markov transition probability matrix of adjacent pixel residuals
	SRM	2012	A plethora of different high-pass filters
	PHARM	2015	Project residuals onto vectors to calculate statistical features
JPEG	JRM	2012	Correlation within and between blocks of multiple JPEG coefficients
	DCTR	2015	DCT transform based on real JPEG spatial data
	GFR	2015	Constructed based on 2D Gabor filters

data-driven approaches. These methods automatically learn complex patterns directly from data, integrating feature extraction and classification into a seamless, end-to-end process. Consequently, modern deep learning-based steganalysis approaches have significantly surpassed traditional ones, becoming the mainstream direction in steganalysis. This paradigm shift underscores the efficiency and effectiveness of deep learning, particularly in handling large volumes of high-dimensional data, positioning it as the preferred method for tackling contemporary steganalysis challenges. In the following, we first provide an overview of existing steganalysis architectures, and then explore various strategies that can enhance the performance of deep steganalysis frameworks.

4.2.1 Network Architectures for Steganalysis

The existing deep steganalysis framework typically employs hybrid networks that combine deep steganalysis networks with handcrafted features. This approach is depicted in the design paradigm shown in Figure 12. Generally, the paradigm includes three stages: a preprocessing module for extracting noise residual features, a series of convolutional modules for further feature extraction, and one or more fully connected layers for classification. In the context of image steganalysis, it is advantageous to avoid downsampling in the early stages of the feature extraction module to enhance residual feature extraction and improve model performance.

In 2014, Tan and Li [144] first discussed the similarities between deep neural networks and traditional handcrafted-based steganalysis features. They proposed a deep steganalysis architecture based on Stacked Convolutional Auto-Encoders (SCAE), which preprocesses images using 40 5×5 high-pass filters, and achieves detection results comparable to SPAM. This model marked the first application of deep learning to the field of steganalysis. Qian *et al.* [122] introduced a convolutional neural network with Gaussian activation functions, named GNCNN, which utilizes fixed KV high-pass filters for preprocessing images. Its performance surpassed that of the SCAE and SRM. Xu *et al.* [181] designed an absolute value layer for the feature from the first

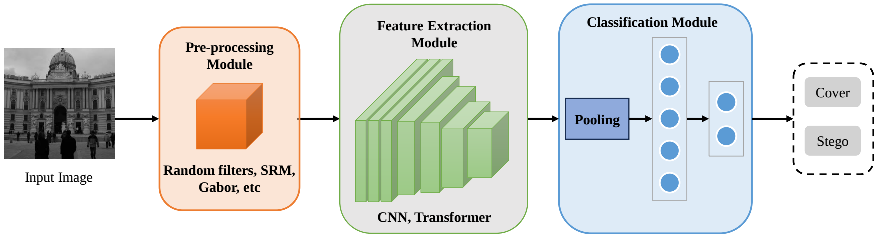


Figure 12: Flowchart of the general processing steps in mainstream deep learning-based steganalysis models.

convolutional layer and proposed a spatial domain steganalysis model based on CNN called XuNet. To prevent overfitting, the authors employed Tanh activation functions in the early layers of the network to constrain the values of the feature maps. Ye *et al.* [191] proposed the use of Thresholded Linear Units (TLU) as activation functions to better capture stego signals and initialized the first layer with 30 5×5 SRM high-pass filters. Yedroudj *et al.* [194] proposed Yedroudj-Net, a CNN-based spatial domain steganalysis model, with the first layer of SRM filters. Li *et al.* [80] proposed a CNN architecture incorporating diverse activation modules, which vary the activation of convolution outputs and then concatenate their outputs for the subsequent layers. Boroumand *et al.*[6] broke new ground in deep learning steganalysis models by proposing an end-to-end approach SRNet based on deep residual networks [51]. The model consists of three parts: the first part extracts noise residuals without pooling operations, the second part performs feature map compactification and dimensionality reduction, and the third part handles classification. Unlike previous deep steganalysis models, this network does not rely on any handcrafted high-pass filters or heuristically initialized preprocessing layer for feature extraction; instead, it fully leverages deep networks for automatic feature learning. The network demonstrates superior detection performance in both spatial and JPEG domain steganalysis tasks compared to all previous models. Deng *et al.* [28] introduced second-order global covariance pooling [84] into deep steganalysis models for spatial domain called CovNet. Subsequently, they introduced 32 additional Gabor filters and shortcut connections into CovNet. This modification ensured better performance in both spatial and JPEG domains [27]. Zhang *et al.* [214] proposed a deep residual multi-scale convolutional network for spatial steganalysis, called DRMCN. To extract features of different dimensions, the authors designed three different scale convolutions, thereby enhancing the detection performance of convolutional neural network-based steganalysis models. Wu *et al.* [174] introduced shared normalization, a novel technique addressing the challenge of generalization in well-trained CNN models with multiple batch normalization layers by sharing consistent statistics across training samples. Zhang *et al.* [212] introduced ZhuNet, which employs SRM-initialized preprocessing layer with trainable weights during training. They replaced traditional convolutional layers with depthwise separable convolutions and utilized Spatial Pyramid Pooling (SPP) for multi-level feature aggregation. You *et al.* [195] introduced a novel deep steganalysis model based on Siamese CNN, called SiaStegNet, capable of handling images of arbitrary sizes without retraining. The model consists of two symmetric subnetworks sharing parameters and weights, aggregating at the end to compute feature similarity and output results. Liu *et al.* [97] introduced a feature enhancement passing module facilitate the transfer of shallow features to deeper layers and an attention downsampling module to perform attention to downsample features while preserving information through the integration of channel attention. Weng *et al.* [169] proposed LWENet, a lightweight deep steganalysis network with less than 400,000 parameters, which enhances performance and reduces parameter count by incorporating lightweight bottleneck residual blocks, depthwise separable convolution layers, and multi-view global pooling. Besides the aforementioned models, researchers have also proposed deep steganalysis models that incorporate self-attention mechanisms to enhance feature extraction capabilities. Luo *et al.* [103] introduced a Convolutional Vision Transformer [172] for spatial steganalysis (CVTStegNet), marking the first application of transformer architecture to grayscale image steganalysis tasks, and achieved performance competitive with SRNet. Weng *et al.* [170] proposed a Swin Transformer-based [101] steganalysis network (SwT-SN) to enhance detection accuracy for arbitrary-sized images. which introduced directional difference adaptive combination (DDAC) [157] followed by a three-layer residual structure, convolutional spatial pyramid pooling equipped with size-independent detector (SID) [155] (CSPP-SID). Xie *et al.* [179] proposed ERANet, integrating an enhanced residual block from Res2Net [41] and an Enhanced Low-level Feature Representation Module (ELLFRM) based on self-attention. This module not only enhances the extraction of complex features but also can serves as a plug-and-play to boost the performance of other steganalysis networks. Li *et al.* [83] proposed IMCoatNet, leveraging CoatNet [24] as the backbone, suitable for spatial and JPEG domains. They incorporated the SKAttention structure [94] for fine-grained feature extraction, followed by a combination of MBConv, Transformer, and PSA layers [208] to extract multi-scale features. Similar to [103], CVTStego-Net [7] incorporates a Convolutional Vision Transformer (CVT) module before the classification stage. Additionally, a bifurcation of trainable and untrainable SRM is employed in the preprocessing stage of CVTStego-Net. From the design of the above models, we can infer that current deep steganalysis networks based on attention are typically combinations of convolutional modules in shallow layers and Transformer modules in deep layers. However, developing a fully attention-based network remains a significant challenge. There is evidence showing that CNNs can effectively utilize locally detectable embedding artifacts [196]. Effective extraction of noise residual features in the shallow layers of the network significantly improves the signal-to-noise ratio of the stego signal to image content, laying the groundwork for successful global feature extraction using attention modules later on.

Given that JPEG images are the predominant image format in social networks, developing deep steganalysis models tailored for JPEG images is of great significance. Apart from the aforementioned SRNet, US-CovNet and IMCoatNet, which can be applied for JPEG steganalysis directily, researchers have also delved into analyzing the characteristics of JPEG images and the statistical perturbations caused by JPEG steganography. This exploration has led to the development of specialized deep steganalysis models for JPEG domain. In 2017, Xu [182] proposed a JPEG steganalysis network, J-XuNet, consisting of 20 layers of deep residual convolutions and initialized the preprocessing layer with fixed 16 4×4 DCT kernels. Inspired by DenseNet, Yang *et al.* [184] proposed a 32-layer CNNs for JPEG image steganalysis, with feature reuse by concatenating all features from preceding layers. This approach facilitates gradient and information propagation, and the shared features and bottleneck layers in the proposed CNN model further reduce the number of model parameters. Zheng *et al.* [220] analyzed the impact of using different high-pass filters in the preprocessing layer of J-XuNet, including DCT filters and Gabor filters. They found that both Gabor and DCT filters exhibited good performance, regardless of whether the filters in the preprocessing layer were fixed or trainable. However, Gabor filters showed superior performance compared to DCT filters. Performance was further improved when the parameters of the preprocessing layer were trainable rather than fixed. Additionally, the authors observed that removing the absolute value layer from J-XuNet effectively enhanced the model's detection performance. Zeng *et al.* proposed a hybrid deep learning framework for JPEG steganalysis [205], called Zeng-Net. The model initializes the preprocessing layer with 25 5×5 DCT base filters, integrates quantization and truncation into deep steganalysis, and then employs a compound deep neural network consisting of multiple subnets. Specifically, the authors validated the effectiveness of this model on a dataset established from ImageNet. Su *et al.* [137] introduced a CNN architecture for JPEG steganalysis, named RXGNet, which employs Gauss partial derivative (GPD) filters as the preprocessing layer and constructs a deep residual network based on ResNeXt [180] blocks. Yousfi and Fridrich [199] introduced one-hot encoding into CNN-based deep learning networks to flexibly compute higher-order statistics of DCT coefficients, thereby enhancing the performance of deep steganalysis networks based on the JPEG domain. Butora *et al.* [11] introduced a novel JPEG steganalysis method called reverse JPEG compatibility attack (RJCA), which involves introducing the statitic of the rounding error in the spatial domain after decompressing the JPEG image. This method is applicable to both color and grayscale JPEG images saved with QF 99 and 100. Subsequently, Butora *et al.* [9] extended the RJCA method by analyzing the logits from CNN detectors on cover images, enabling the establishment of accurate cover logit distributions to determine theoretical thresholds for any desired false positive rate, thus facilitating steganalysis across images of varying sizes without CNN retraining.

For color images, Zeng *et al.* [206] proposed WISERNet, which follows a strategy of feature separation and aggregation. The authors theoretically demonstrated that the summation operation in conventional convolutions acts as a Linear Collusion Attack, preserving strongly correlated patterns while suppressing unrelated noise. At the bottom layers of the network, independent convolutional layers are applied to each color channel to extract features and suppress irrelevant image content. At the higher layers, all channels are aggregated to enhance the information extracted from the bottom layers through convolution operations. Wei *et al.* [167] proposed a universal deep steganalysis network for color images applicable to both spatial and JPEG domains. It initializes a fixed preprocessing layer with 30 SRM kernels and 32 Gabor kernels, separately preprocessing the three color channels of the color image to obtain 186-dimensional noise residual features. At the higher layers of the network, it combines residual convolutions with depthwise separable convolutions, resulting in a steganalysis network with fewer parameters and high performance. Then, Wei *et al.* [166] introduced a Transformer module and global covariance pooling (GCP) into the UCNet, resulting in an effective color spatial steganalysis network. Furthermore, Wei *et al.* [165] introduced a multi-stage neural network for color image steganalysis. It's worth mentioning that in the KAGGLE ALASKA II competition [17], a large number of computer vision deep networks pretrained on the ImageNet dataset have been transferred to color JPEG image steganalysis tasks. Researchers have gradually realized the good transferability of large-scale pretrained models on datasets like ImageNet or JIN in the steganalysis field. These models such as EfficientNet [143], enabling models to converge quickly on more complex steganalysis tasks and achieve quite impressive performance [198, 12]. Introducing certain domain knowledge into these models, such as setting the convolutional stride of the stem layer to one, can further enhance the performance of the network [197].

In Table 6, we provide a summary of representative deep learning-based image steganalysis architectures. In addition, we have provided a comparison of the parameters and FLOPs for typical deep learning-based steganalysis methods, as shown in Table 7. For the sake of fair comparison, we do not present a comprehensive comparison of the performance numbers of various steganalysis methods, due to different studies may use different datasets or adopt different partitioning methods on the same dataset, etc.

Table 6: Summary of representative deep learning-based image steganalysis architecutures

Domain	Method	Year	Highlight	Filter used
Spatial	SCAE	2014	Stacked Convolutional Auto-Encoders	SRM
	GNCNN	2015	Gaussian activation layer	KV
	XuNet	2016	Absolute value layer, TanH activation layer	KV
	YeNet	2017	Thresholded Linear Units	SRM
	Yedroudj-Net	2018	Clever fusion of effective modules	SRM
	Rest-Net	2018	Diverse activation	SRM&Gabor
	DRMCN	2019	Deep residual multi-scale CNN	SRM
	WISERNet	2019	Separation-reunion for color image	SRM
	CovNet	2019	Global covariance pooling layer	SRM
	ZhuNet	2020	Separable convolution, spatial pyramid pooling	SRM
	SiaStegNet	2021	Siamese CNN	SRM
	FPNet	2022	Feature enhancement passing, attention downsampling	SRM
	CVTStegNet	2022	Convolutional Vision Transformer	SRM
JPEG	J-XuNet	2017	DCT Kernels	DCT
	ZengNet	2018	A compound deep neural network consisting of multiple subnets	DCT
	Yang *et al.*,	2018	Feature reuse	DCT
	RXGNet	2021	ResNeXt, Gauss partial derivative filters	Gauss partial derivative
	One-hot	2020	One-hot encoding	Random
Both	SRNet	2019	Clean end-to-end design	None
	UCNet	2022	Universal deep steganalysis network for color images	SRM&Gabor

4.2.2 Improving Performance of Deep Steganalysis Architectures

In addition to designing effective network architectures for steganalysis, implementing appropriate model optimization, data augmentation, and channel awareness strategies is essential for enhancing the performance of deep steganalysis models. In the following, we will describe some related works on these topics.

Model Optimization: In the pursuit of higher accuracy, neural network tend to evolve into increasingly intricate structures, leading to concerns about excessive model size and redundancy. To mitigate model redundancy without significantly impacting performance, compressing deep steganalysis models emerges as a straightforward solution.

Table 7: Summary of parameters and FLOPs for representative steganalysis methods

Method	Parameters	FLOPs
XuNet	0.03×10^6	0.16×10^9
YeNet	0.11×10^6	3.87×10^9
Yedrouj-Net	0.54×10^6	6.63×10^9
SRNet	4.78×10^6	12.10×10^9
CovNet	0.68×10^6	6.70×10^9
ZhuNet	2.87×10^6	2.42×10^9
SiaStegNet	0.71×10^6	14.40×10^9
UCNet	1.12×10^6	14.3×10^9
CALPA-Net	0.07×10^6	3.94×10^9

Li *et al.* [86] first introduced non-structured pruning method (weight level) for deep steganalysis networks. This method significantly sparsified the model weights with minimal accuracy loss following the traditional three-stage model pruning process: training, pruning, and fine-tuning. However, non-structured pruning methods typically require specialized algorithms or hardware for accelerating network inference, imposing certain constraints on achieving acceleration. To overcome the limitations of non-structured network pruning methods, Tan *et al.* [146] combined proposed a structured pruning methods (channel level) for deep residual steganalysis network, named CALPA-NET. This approach significantly reduces the model's parameter count and floating-point operations while maintaining performance comparable to the original model. Subsequently, Tan *et al.* [145] proposed STD-NET based on tensor decomposition for compressing steganalysis models. In comparison to CALPA-NET, this method is not constrained by residual connections and achieves better compression results.

With the advancement of neural architecture search (NAS) techniques across various research domains, researchers have integrated the characteristics of existing state-of-the-art deep learning-based steganalysis networks to propose a series of image steganalysis methods based on neural architecture search. Yang *et al.* [186] first introduced neural architecture search algorithm for JPEG image steganalysis, termed JS-NAG. They proposed a Q-learning-based [163] approach where an agent is trained to continuously select high-performance structures to generate the architecture. This method integrates multiple searched networks to address the issue of unstable performance in individual network. Deng *et al.* [29] proposed a NAS method for spatial image steganalysis based on PC-DARTS [183], which primarily focuses on exploring suitable architectures from a cell-based search space, and achieves competitive performance with SRNet.

Data Augmentation: An ideal deep steganalysis framework relies not only on the design of the network architecture, but also on a suitable data augmentation strategy. Existing data augmentation methods for image steganalysis can be categorized into augmentation before and after steganographic embedding. We refer to them as 'pre-embedding' and 'post-embedding' repectively, as summarized in Table 8. A common and easy-to-implement data augmentation method for steganalysis is flips and rotations (D4) [198]. But this method is not sufficient to improve the generalization of deep steganalysis models. Yedroudj *et al.* [193] proposed a data augmentation method called pixels-off, which randomly selects a small number of pixels from the cover and sets them to zero to obtain a new cover. However, this approach inevitably alters the original cover distribution from the perspective of steganalysis. Subsequently, they further introduced adaptive-pixels-off by combining embedding probability. Yu *et al.* [201] proposed Bitmix, which mixes random patches in a cover and stego image pair and generates a soft label with the ratio of the number of modified pixels in the swapped patch. Itzhaki *et al.* [65] explored various data augmentation techniques for JPEG steganalysis and concluded that StegoSampling and Dropout-style augmentations are beneficial for JPEG steganalysis. Zhang *et al.* [210] proposed a cover augmentation network based on the principle of preserving data distribution, which automatically adds noise to the original cover images to generate new covers. Subsequently, the authors proposed a differentiable augmentation network trained adversarially with steganalyzer to augment cover and stego images by intelligently adding noises [209]. This method not only improves the performance of existing steganalysis networks but also further enhances performance when combined with existing cover enrichment methods.

Table 8: Summary of data augmentation methods for image steganalysis

Method	Category	Domain
D4 [198]	post-embedding	Spatial&JPEG
Pixels-off [193]	pre-embedding	Spatial
BitMix [201]	post-embedding	Spatial
StegoSampling [65]	post-embedding	JPEG
DPAA [210]	pre-embedding	Spatial
AAS [209]	post-embedding	Spatial&JPEG

Selection Channel awareness: Content-adaptive steganography adaptively embeds information into the cover image, guided by embedding probabilities. These probabilities can also be utilized by the steganalyzer as a selection channel, thereby enhancing the capability of existing steganalysis networks. For instance, Huang *et al.* [63] extended YeNet to the JPEG domain and proposed a selection-channel-aware CNN for JPEG steganalysis. Ren *et al.* [129] proposed a method that incorporates selection channel awareness. This approach consists of two main components: a selection channel network and a steganalysis network. These networks are trained concurrently; the selection channel network learns and outputs selection channels for the steganalysis network, which then uses these learned channels to predict whether digital media contains hidden messages. To fully utilize the embedding probability, Li *et al.* [85] proposed an embedding probability guided module to adaptively enhance the feature extraction capability at different depths of the network. Han *et al.* [49] introduced non-local operations and multi-channel convolution modules after the preprocessing layer to enhance noise residual features. Wu *et al.* [175] proposed a method for deep learning-based image steganalysis that automatically learns selection channels in a progressive manner and integrates them into a steganalysis network, significantly improving detection accuracy without requiring prior knowledge. Wei *et al.* [164] proposed a residual guided coordinate attention for selection channel aware image steganalysis, and achieves better performance with CovNet and J-YeNet. Zhang *et al.* [216] proposed a dual attention fusion network, incorporating a Sobel spatial attention module and a channel attention module based on DCT coefficients.

4.3 Exploring Steganalysis for Real-world Scenarios

In this section, we will explore how steganalysis methods have adapted to address the challenges posed by real-world scenarios. Specifically, we will divide our discussions into two main parts: research on mismatched steganalysis scenarios and adversarially robust steganalysis. The former focuses on the challenges posed by mismatched data sources, while the latter addresses the vulnerability of deep steganalysis models to image steganography based on adversarial examples. Through these discussions, we aim to provide insights into the ongoing efforts to develop steganalysis methods that are both effective and adaptable to real-world challenges.

4.3.1 Research on Mismatched Steganalysis Scenarios

In practice, mismatched data sources are often encountered, where the detector is usually unaware of the origin of the target under examination. Discrepancies in the feature distributions between training and testing sets can significantly degrade the performance of steganalysis. As shown in Figure 13, the typical

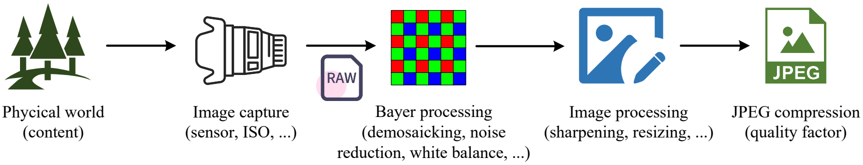


Figure 13: Flowchart of the image processing pipeline using a standard digital camera to convert real-world visuals into JPEG format.

process of capturing an image involves capturing a raw image from the physical world through the lens and sensor, followed by performing pre-processing (also called Bayer processing, such as demosaicking, noise reduction, white balance) to produce full-color images. Image processing steps like sharpening and resizing are then applied. Finally, for the sake of convenient transmission and storage, the image may undergo JPEG compression. Each step of this process affects the distribution of the final image, which is the main reason for encountering cover source mismatches (CSM) issues in the real world. CSM exist both in the spatial and JPEG domains. Enhancing the similarity between the feature distributions of the training and testing sets is a highly effective approach to address the mismatch problem.

In 2014, Kodovský *et al.* [74] proposed two simple strategies to mitigate this problem: training a single classifier on a mixture of sources and training a bank of detectors on multiple different sources and then testing on the one with the closest source. Hu *et al.* [57] investigated the interaction between CSM and texture complexity. They proposed a texture complexity measuring method based on average filter and introduced two-way analysis of variance to analyze the interaction between the two factors. To address the problem of detecting diversified stego sources, where the steganalyst is unaware of the steganographic method used by the steganographer, Butora *et al.* [10] found that the multi-class detector was the most effective approach. Zhang *et al.* [217] proposed J-Net, where the features of the source and target are aligned by minimizing the JMMD [102] distance at the fully connected layer. Quentin *et al.* [128] conducted a study on the impact of various source attributes (including camera, ISO, processing pipeline, and content) on deep learning methods and observed that the holistic strategy leverages the good generalization properties of deep learning to mitigate the CSM with a relatively small number of training samples. To address payload mismatch, Yu *et al.* [202] proposed an adaptive multi-teacher softened relational knowledge distillation framework. Subsequently, they [203] introduced the RCDD (Reliable Steganalysis Labeling-based Contrastive Domain Discrepancy) method to generate reliable labeled target images, thereby achieving domain alignment. Megias *et al.* [113] introduced a "directionality" property arising from subsequent embedding, indicating that additional data embeddings distort the image features in the same direction in the feature space. Through theoretical analysis and experimentation, they demonstrated that this strategy effectively enhances the detection performance of steganalysis models in scenarios with mismatched data sources.

For JPEG mismatched steganalysis, Jia *et al.* [69] proposed the THFSL (transferable heterogeneous feature subspace learning) algorithm that captures local information by imposing low-rank constraints on domain-independent features and avoids negative transfer with a sparse matrix for domain-related features. Yang *et al.* [189] proposed a transfer subspace learning method based on structure preservation, which learns a discriminant projection matrix to map training and test data into a common low-dimensional subspace. To address the CSM issue caused by JPEG quality variations, Yousfi *et al.* [200] found that CNN-based detectors trained on mixed quality factors do not significantly lose performance compared to those trained for specific quality factors. Additionally, under the same strategy, CNN-based detectors show better robustness than feature-based detectors. Jia *et al.* [68] proposed an effective imbalanced JPEG steganalysis scheme based on adaptive cost-sensitive feature learning. Subsequently, they [67] proposed MPSA (multiperspective progressive structure adaptation) scheme based on active progressive learning for JPEG mismatched steganalysis.

4.3.2 Adversarially Robust Steganalysis

In recent years, deep learning steganalyzers have advanced rapidly. However, similar to how deep learning models are vulnerable to adversarial attacks, deep steganalysis models struggle to detect adversarial steganography. In image classification tasks, common methods for defending against adversarial examples include preprocessing [45, 95] and adversarial retraining [44, 131]. However, these methods are not suitable for defending against adversarial steganography. Preprocessing methods, such as transformations[45] or denoising[95], can easily disrupt the information in stego images, while adversarial retraining methods, which augment the training set with adversarial stego images, are not always effective in practice. Consequently, researchers have proposed a series of new methods to address this issue. A naive defense against adversarial steganography is to train the model with adversarial steganographic images included in the training set. The limitation of this approach is that it always fails to address unseen and more sophisticated adversarial steganography. Qin *et al.* [125] proposed Patch Steganalysis, a method that samples image patches based on predicted modification probabilities, followed by deep feature extraction and ensemble classification to enhance the robustness of deep steganalysis. However, sampling small patches is not sufficiently effective, as such patches may not contain enough information. Subsequently, they introduced an adversarial training method to filter out adversarial steganography by combining handcrafted steganalysis features [124]. Hu *et al.* [58] proposed TStegNet, a two-stream CNN steganalysis network against adversarial steganography. This method integrates a confidence loss function to capture the gradient stream that amplify the confidence artifacts, and uses the feature similarity function to reduce the impact of adversarial perturbation.

5 Challenges and Future Research Prospects

In Sections 3 and 4, we categorized and provided an overview of current image steganographic and steganalysis methods, and briefly touched on the core concepts of some pivotal algorithms. In this section, our aim is to analyze the limitations of current research and highlight the gaps between theoretical advancements and real-world applications. Subsequently, we will identify several prevalent future directions in the field of image steganography and steganalysis.

5.1 Challenges

Currently, the majority of research in steganography and steganalysis is conducted under controlled laboratory conditions. While these studies demonstrate promising performance in academic papers, applying these research outcomes directly to real-world environments poses various challenges, so that existing methods may encounter significant algorithmic performance degradation or even failure. The main challenges primarily include:

- **Regarding steganographic issues:** Steganographic research currently faces challenges in three main areas: security assessment of steganographic methods, robustness within social media environments, and the types of covers used. 1) **Security Assessment.** The security performance of current steganographic methods is primarily evaluated using existing steganalysis techniques. This evaluation mechanism lacks scientific rigor for two reasons. First, the security of steganographic methods against unknown, more advanced steganalysis techniques cannot be directly inferred from existing technologies, posing significant security risks. Second, performing security analysis (i.e., detection error rates in steganalysis) on different image databases reveals considerable variations, leading to inconsistent security performance across the same steganographic methods; 2) **Robustness.** When transmitting image data on social media, it often undergoes lossy operations such as image scaling, lossy compression, and the addition of visible watermarks. However, as described in Section 2.1, current models assume a lossless transmission channel, which renders these methods ineffective at ensuring correct extraction of secret information even with minor losses. Although some efforts have been made to explore "robust steganography," these methods often do not prioritize resistance to sophisticated steganalysis techniques as their primary criterion. Instead, they focus on robustness metrics such as the rate of accurate secret information extraction following attacks. Developing models and evaluation metrics tailored to real-world scenarios, departing from the traditional passive-warden model, requires more thorough investigation; 3) **Types of covers.** On various social media platforms, most images are color JPEGs, while grayscale images are increasingly rare. However, current steganographic and steganalysis research predominantly focuses on grayscale images in the spatial domain. Due to significant statistical differences between grayscale and color images, and between spatial and DCT embedding domains, most existing steganographic techniques are ineffective when applied to color JPEG images. This mismatch highlights the need for more targeted research into how modifications within different color channels and JPEG's lossy compression affect the security of steganography.

- **Regarding steganalysis issues:** Steganalysis research currently faces two major challenges: significant performance drop in real, uncontrolled scenarios and the inability to detect small-capacity steganographic embedding. 1) **Uncontrolled scenarios.** In current steganalysis research, experimental setups often require that the training data and test data follow relatively consistent distributions. This consistency is not only reflected in the image processing history of image databases, as mentioned earlier, but also in various important factors related to data embedding, such as the steganography algorithm used and the embedding rate. Although current steganalysis methods, particularly data-driven deep learning methods, can model standardized databases well and achieve relatively optimal test performance, this is largely due to our unrealistic experimental setups. In real-world scenarios, however, it is difficult or impossible to predict these factors for a suspected image. As a result, the assumed consistency between training data and test data cannot be guaranteed. Therefore, steganalysis models trained on laboratory data often struggle to effectively handle the diversity of images encountered in practical scenarios. Existing research has shown that under mismatched conditions, the performance of current steganalysis methods can significantly degrade or even drop to the level of random guessing. 2) **Small-capacity Embedding.** With the rapid increase in internet speed, the transmission of large-capacity carriers has become commonplace and inexpensive. Existing research indicates that even with other factors being equal, reducing the embedding payload significantly degrades the performance of steganalysis methods. Therefore, a simple and effective way to enhance steganographic security is by trading bandwidth for security, reducing the embedding rate to a level that current steganalysis methods find difficult to detect—for example, 0.01bpp/bpnzac or less. Based on the above analysis, achieving practical steganalysis remains a challenge that requires more research efforts.

- **Regarding standardized databases:** Currently, the databases commonly used in digital image steganography and steganalysis exhibit two major issues that result in a significant gap between their utility and real-world application scenarios: 1) **Volume and Diversity.** Digital image steganography and steganalysis primarily target internet applications, where image data is abundant and shows vast variations in content, resolution, and quality. However, as indicated in Table 1, the largest datasets available for these fields contain no more than 210,000 cover samples. This number is markedly small compared to the vast scale of the internet. Additionally, this quantity is typically inadequate for training advanced neural networks, which necessitate large-scale datasets to avoid overfitting and guarantee consistent, robust performance. As a result, while models may perform well on controlled datasets, these performances do not necessarily translate to effectiveness in practical, varied internet environments; 2) **Uniformity in Artifacts.** As discussed in Section 2.3, the construction of image databases for research typically involves standardized post-processing steps applied to raw data from cameras. These steps include image demosaicking, conversion to 8-bit grayscale, downsampling, and center-cropping, all of which standardize the image resolution. This uniform processing introduces specific statistical characteristics to the images, potentially undermining the diversity needed for real-world applications. Crucially, data-driven deep learning approaches may end up modeling and learning these artificial statistical features, rather than the inherent statistical characteristics of the image content. Consequently, this can lead to algorithms that perform well on processed datasets but falter with the more diverse and varied images encountered in real internet scenarios.

5.2 Future Research Prospects

To address the current limitations in steganography and steganalysis, and to align research more closely with practical application scenarios, the subsequent research should first focus on expanding the existing databases. This expansion should not merely increase the quantity of images, but also enhance their diversity to more closely mirror real-world scenarios. Based on the construction of a large and diverse image dataset, the following research directions merit further in-depth exploration and study:

- **Exploring more effective generative deep learning for steganography:** Driven by big data, generative deep learning offers a promising new avenue for enhancing the security of existing steganographic methods and improving robustness in social media contexts. Although various generative deep learning techniques, such as variational autoencoders, GANs, and diffusion models, have shown exceptional performance in fields like image generation, enhancement, and style transfer, their application in image steganography is still nascent with only a few related works. Image steganography is markedly different from other image processing tasks. It involves embedding information into an image through subtle modifications to its embedding units, ensuring that these alterations do not compromise the image's visual quality. Additionally, these modifications must remain undetectable to steganalysis techniques and robust against post-processing on social networks. Due to these specialized requirements, existing generative models cannot be directly applied to steganography. As deep learning technologies continue to evolve, the development of more effective models for various tasks including steganography is anticipated. This raises the significant challenge of how to adapt these advanced models to meet the requirements of practical steganographic scenarios. This adaptation involves a thorough analysis of the core frameworks of existing models, comparing the application scenarios of these technologies with those required for steganography, and making targeted adjustments to their network architectures. Furthermore, special processing measures tailored for steganography might be necessary. These could include designing differentiable simulated embedding functions and adapting to the lossy processes typical on social networks. Additionally, the design of loss functions should prioritize enhancing steganographic security over improving the visual quality, which is a common focus in most conventional image processing tasks. Moreover, proposing effective update strategies to ensure model convergence and stability during steganographic training is essential.

- **Exploring more promising deep steganalysis models:** The next generation of deep learning steganalysis models should be able to adapt well to real, uncontrolled scenarios, as well as effectively detect small-capacity steganographic embeddings. With the rapid development of deep learning, a variety of innovative modules and structures have emerged, demonstrating effectiveness across different fields. It is therefore essential to explore how these advancements can be adapted and integrated into steganalysis tasks. For instance, recent research has developed CNN-Transformer hybrid architectures that have shown good performance in spatial steganalysis tasks by leveraging the local pattern recognition capabilities of CNNs combined with the global context understanding of Transformers. These hybrid models have demonstrated potential, but there remains significant room for improvement and innovation. Given the unique nature of steganalysis modifications, which typically occur in complex image textures while simpler areas remain unchanged, it is essential to incorporate corresponding deep modules to better trace these steganographic artifacts. For instance, attention mechanisms can be particularly effective, as they enable steganalyzers to focus on subtle irregularities indicative of steganographic activities. These mechanisms enhance the model's sensitivity to minimize anomalies in textured regions, where steganographic content is likely hidden. Additionally, incorporating Graph Neural Networks (GNNs) [173] can enrich the model's understanding of data structures by capturing the relationships and interactions between different image regions, essential for identifying complex and dispersed steganographic patterns. Capsule Networks [130] also play a significant role by preserving hierarchical relationships within the image data, thus recognizing both simple and intricate patterns effectively. Green Steganalyzer [227] aims to explore a novel learning solution to image steganalysis based on the green learning paradigm, which has lower computational complexity and smaller model size. Exploring these advanced modules, experimenting with various combinations, optimizing their interactions, and validating their effectiveness across diverse datasets and steganographic techniques are key steps towards developing more proficient and robust steganalysis models. Addressing these challenges will enable the creation of advanced systems capable of detecting small-capacity steganographic embeddings in real, uncontrolled scenarios.

- **Exploring effective utilization of large models in steganaography and steganalysis:** Existing literature demonstrates that large pre-trained models for artificial intelligence-generated content (AIGC), such as ChatGPT and Stable Diffusion, are effective in various downstream tasks. However, there are currently no reported successful applications of these models in the fields of steganography and steganalysis. The primary reasons for this are the complexity of large models and the significant differences between their original training domains and the specialized requirements of steganography and steganalysis tasks. To overcome these challenges, innovative techniques such as prompt-tuning, side networks, and adaptors have been developed as essential tools in the fine-tuning process, effectively bridging the gap between the models' initial training domains (e.g., language understanding and image generation) and the specific needs of steganographic and steganalysis applications. Compared to existing deep learning methods that utilize smaller models, the potential benefits of employing large AIGC models in steganography and steganalysis are immense. These models have the capacity to learn extensive general knowledge and features during their pre-training on vast datasets, which enables them to discern and understand complex patterns and relationships. This adaptability makes them well-suited for fine-tuning to specific tasks across various domains. Harnessing this deep understanding could significantly enhance both the embedding of secret information in steganography and the detection of covert channels in steganalysis, potentially leading to groundbreaking advancements in both fields. For example, large models could be used to generate or identify image regions with enhanced security and robustness for steganography. In steganalysis, these models could help address the performance degradation often seen in deep learning models under mismatched conditions and improve the detection capabilities against low embedding rate steganography in high-capacity carriers.

- **Exploring open-set adversarial learning framework for steganography and steganalysis:** The current state of adversarial interactions between steganography and steganalysis techniques remains largely confined within a rather simplistic, closed set environment. This limitation has significant implications for the performance and adaptability of steganalysis models. Specifically, models trained within this restricted environment tend to exhibit substantial performance degradation when confronted with common real-world challenges. Similarly, steganography models, developed through adversarial training with steganalysis models within this closed set environment, often suffer from overfitting. They become overly specialized to counter the specific steganalysis models they were trained against. Addressing this issue is a significant challenge, and the academic community has yet to propose an effective solution. One potential approach under consideration involves constructing an adversarial learning and gaming framework for image steganography and steganalysis within an open set environment. This framework would incorporate third-party steganalysis attacks, enabling the steganography and steganalysis parties to engage in iterative optimization at the pixel level. The hope is that this approach will overcome the performance limitations of deep learning steganography and steganalysis models in real-world settings, thereby enhancing their robustness. By stepping outside the confines of the closed set environment and embracing a more open and realistic adversarial context, it is expected that both steganography and steganalysis models can better adapt to the complexities and unpredictability of real-world scenarios. For steganography, it can therefore avoid overfitting to specific steganalysis techniques and specific training cover datasets, enhancing the resistance to unknown steganalysis techniques. For steganalysis, it can therefore effectively learn the universal intrinsic features of steganalysis that are not highly coupled with a specific benchmark database, steganographic algorithm, or embedding capacity, enabling the model to adapt to the rapidly evolving steganalysis detection tasks in real-world application scenarios, achieving a significant step forward in the field by providing reliable and resilient tools for secure data communication.

6 Conclusion

This survey provides a detailed overview of the development and current state of steganography and steganalysis in digital communications, emphasizing their vital roles in the secure transmission of covert information. It starts with the passive-warden scenario, examining its importance and foundational concepts, and progresses to the evolution from traditional handcrafted methods to sophisticated deep learning techniques that have emerged over the past two decades. The survey categorizes existing technologies, describes the main ideas and performance of typical algorithms within each category, and offers a summary of these findings. Finally, the survey identifies existing challenges in the fields of steganography and steganalysis and suggests promising directions for future in-depth research.

References

[1] P. Bas, T. Filler, and T. Pevný, "Break Our Steganographic System: The Ins and Outs of Organizing BOSS", in *Proceedings of the International Workshop on Information Hiding*, May 2011, 59–70.

[2] P. Bas and T. Furon, "BOWS2", in *http://bows2.ec-lille.fr*, 2007.

[3] S. Bernard, P. Bas, J. Klein, and T. Pevny, "Explicit optimization of min max steganographic game", *IEEE Transactions on Information Forensics and Security*, 16, 2020, 812–23.

[4] S. Bernard, P. Bas, J. Klein, and T. Pevný, "Backpack: A backpropagable adversarial embedding scheme", *IEEE Transactions on Information Forensics and Security*, 17, 2022, 3539–54.

[5] S. Bernard, T. Pevný, P. Bas, and J. Klein, "Exploiting adversarial embeddings for better steganography", in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2019, 216–21.

[6] M. Boroumand, M. Chen, and J. Fridrich, “Deep residual network for steganalysis of digital images”, *IEEE Transactions on Information Forensics and Security*, 14(5), 2019, 1181–93.

[7] M. A. Bravo-Ortiz, E. Mercado-Ruiz, J. P. Villa-Pulgarin, C. A. Hormaza-Cardona, S. Quiñones-Arredondo, H. B. Arteaga-Arteaga, S. Orozco-Arias, O. Cardona-Morales, and R. Tabares-Soto, “CVTStego-Net: A convolutional vision transformer architecture for spatial image steganalysis”, *Journal of Information Security and Applications*, 81, 2024, 103695.

[8] J. Butora and P. Bas, “Side-informed steganography for jpeg images by modeling decompressed images”, *IEEE Transactions on Information Forensics and Security*, 2023.

[9] J. Butora and P. Bas, “Size-Independent Reliable CNN for RJCA Steganalysis”, *IEEE Transactions on Information Forensics and Security*, 2024, 1–1.

[10] J. Butora and J. Fridrich, “Detection of Diversified Stego Sources with CNNs”, *Electronic Imaging*, 31(5), 2019, 534-1–534-1.

[11] J. Butora and J. Fridrich, “Reverse JPEG Compatibility Attack”, *IEEE Transactions on Information Forensics and Security*, 15, 2020, 1444–54.

[12] J. Butora, Y. Yousfi, and J. Fridrich, “How to Pretrain for Steganalysis”, in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2021, 143–8.

[13] B. Chen, W. Luo, and P. Zheng, “Enhancing Steganography via Stego Post-processing by Reducing Image Residual Difference”, in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2019, 63–8.

[14] B. Chen, W. Luo, P. Zheng, and J. Huang, “Universal stego post-processing for enhancing image steganography”, *Journal of Information Security and Applications*, 55, 2020, 102664.

[15] K. Chen, W. Zhang, H. Zhou, N. Yu, and G. Feng, “Defining cost functions for adaptive steganography at the microscale”, in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2016, 1–6.

[16] K. Chen, H. Zhou, W. Zhou, W. Zhang, and N. Yu, “Defining cost functions for adaptive JPEG steganography at the microscale”, *IEEE Transactions on Information Forensics and Security*, 14(4), 2018, 1052–66.

[17] R. Cogranne, Q. Giboulot, and P. Bas, “ALASKA#2: Challenging Academic Research on Steganalysis with Realistic Images”, in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2020, 1–5.

[18] R. Cogranne, Q. Giboulot, and P. Bas, “Efficient steganography in JPEG images by minimizing performance of optimal detector”, *IEEE Transactions on Information Forensics and Security*, 17, 2021, 1328–43.

[19] R. Cogranne, Q. Giboulot, and P. Bas, “Steganography by minimizing statistical detectability: The cases of JPEG and color images”, in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2020, 161–7.

[20] R. Cogranne, Q. Giboulot, and P. Bas, “The ALASKA steganalysis challenge: A first step towards steganalysis "into the wild"”, in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2019, 125–37.

[21] C. Cortes and V. Vapnik, “Support-vector networks”, *Machine Learning*, 20(3), 1995, 273–97.

[22] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, *Digital watermarking and steganography*, 2007.

[23] Q. Cui, W. Tang, Z. Zhou, R. Meng, G. Nan, and Y. Shi, “Meta Security Metric Learning for Secure Deep Image Hiding”, *IEEE Transactions on Dependable and Secure Computing*, (01), 2024, 1–14, ISSN: 1941-0018.

[24] Z. Dai, H. Liu, Q. V. Le, and M. Tan, “Coatnet: Marrying convolution and attention for all data sizes”, *Proceedings of the Advances in Neural Information Processing Systems*, 34, 2021, 3965–77.

[25] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, and J. Fridrich, “Selection-channel-aware rich model for Steganalysis of digital images”, in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2014, 48–53.

[26] T. Denemark and J. Fridrich, “Improving steganographic security by synchronizing the selection channel”, in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2015, 5–14.

[27] X.-Q. Deng, B.-L. Chen, W.-Q. Luo, and D. Luo, “Universal Image Steganalysis Based on Convolutional Neural Network with Global Covariance Pooling”, *Journal of Computer Science and Technology*, 37(5), 2022, 1134–45, ISSN: 1860-4749.

[28] X. Deng, B. Chen, W. Luo, and D. Luo, “Fast and Effective Global Covariance Pooling Network for Image Steganalysis”, in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2019, 230–4.

[29] X. Deng, W. Luo, and Y. Fang, “Spatial Steganalysis Based on Gradient-Based Neural Architecture Search”, in *Proceedings of the Provable and Practical Security*, 2021, 365–75.

[30] S. Dumitrescu, X. Wu, and Z. Wang, “Detection of LSB steganography via sample pair analysis”, *IEEE Transactions on Signal Processing*, 51(7), 2003, 1995–2007.

[31] Z. Fan, K. Chen, C. Qin, K. Zeng, W. Zhang, and N. Yu, "Image Adversarial Steganography Based on Joint Distortion", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, 1–5.

[32] G. Feng, X. Zhang, Y. Ren, Z. Qian, and S. Li, "Diversity-Based Cascade Filters for JPEG Steganalysis", *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2), 2020, 376–86.

[33] T. Filler and J. Fridrich, "Gibbs construction in steganography", *IEEE Transactions on Information Forensics and Security*, 5(4), 2010, 705–20.

[34] T. Filler, J. Judas, and J. Fridrich, "Minimizing additive distortion in steganography using syndrome-trellis codes", *IEEE Transactions on Information Forensics and Security*, 6(3), 2011, 920–35.

[35] R. A. Fisher, "The use of multiple measurements in taxonomic problems", *Annals of eugenics*, 7(2), 1936, 179–88.

[36] J. Fridrich, M. Goljan, and R. Du, "Detecting LSB steganography in color, and gray-scale images", *IEEE MultiMedia*, 8(4), 2001, 22–8.

[37] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images", *IEEE Transactions on Information Forensics and Security*, 7(3), 2012, 868–82.

[38] J. Fridrich, *Steganography in digital media: principles, algorithms, and applications*, Cambridge University Press, 2009.

[39] J. Fridrich and T. Filler, "Practical methods for minimizing embedding impact in steganography", in *Security, Steganography, and Watermarking of Multimedia Contents IX*, Vol. 6505, SPIE, 2007, 13–27.

[40] J. Fridrich and J. Kodovskŷ, "Steganalysis of LSB replacement using parity-aware features", in *Proceedings of the International Workshop on Information Hiding*, 2012, 31–45.

[41] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A New Multi-Scale Backbone Architecture", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2), 2021, 652–62.

[42] M. Goljan, J. J. Fridrich, and T. Holotyak, "New blind steganalysis and its implications", in *Proceeding of the SPIE on Security, Forensics, Steganography and Watermarking of Multimedia*, Vol. 6072, 2006, 1–13.

[43] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", *Proceedings of the Advances in Neural Information Processing Systems*, 27, 2014.

[44] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples", in *Proceedings of the International Conference on Learning Representations*, 2015.

[45] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering Adversarial Images using Input Transformations", in *Proceedings of the International Conference on Learning Representations*, 2018.

[46] L. Guo, J. Ni, and Y. Q. Shi, "Uniform embedding for efficient JPEG steganography", *IEEE transactions on Information Forensics and Security*, 9(5), 2014, 814–25.

[47] L. Guo, J. Ni, W. Su, C. Tang, and Y.-Q. Shi, "Using statistical image model for JPEG steganography: Uniform embedding revisited", *IEEE Transactions on Information Forensics and Security*, 10(12), 2015, 2669–80.

[48] B. T. Hammad, I. T. Ahmed, and N. Jamil, "A steganalysis classification algorithm based on distinctive texture features", *Symmetry*, 14(2), 2022, 236.

[49] X. Han and T. Zhang, "Spatial Steganalysis Based on Non-Local Block and Multi-Channel Convolutional Networks", *IEEE Access*, 10, 2022, 87241–53.

[50] J. Hayes and G. Danezis, "Generating steganographic images via adversarial training", *Proceedings of the Advances in Neural Information Processing Systems*, 30, 2017.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[52] V. Holub and J. Fridrich, "Random Projections of Residuals for Digital Image Steganalysis", *IEEE Transactions on Information Forensics and Security*, 8(12), 2013, 1996–2006.

[53] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters", in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2012, 234–9.

[54] V. Holub and J. Fridrich, "Low-Complexity Features for JPEG Steganalysis Using Undecimated DCT", *IEEE Transactions on Information Forensics and Security*, 10(2), 2015, 219–28.

[55] V. Holub and J. Fridrich, "Phase-aware projection model for steganalysis of JPEG images", in *Media Watermarking, Security, and Forensics*, Vol. 9409, 2015, 94090T.

[56] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain", *EURASIP Journal on Information Security*, 2014, 2014, 1–13.

[57] D. Hu, Z. Ma, Y. Fan, S. Zheng, D. Ye, and L. Wang, "Study on the interaction between the cover source mismatch and texture complexity in steganalysis", *Multimedia Tools and Applications*, 78(6), 2019, 7643–66.

[58] M. Hu and H. Wang, “Image Steganalysis Against Adversarial Steganography by Combining Confidence and Pixel Artifacts”, *IEEE Signal Processing Letters*, 30, 2023, 987–91.

[59] X. Hu, H. Chen, J. Ni, and W. Su, “A novel steganography scheme based on asymmetric embedding model”, in *Proceedings of the International Conference on Cloud Computing and Security*, Springer, 2018, 183–94.

[60] D. Huang, W. Luo, M. Liu, W. Tang, and J. Huang, “Steganography Embedding Cost Learning with Generative Multi-Adversarial Network”, *IEEE Transactions on Information Forensics and Security*, 2023.

[61] D. Huang, W. Luo, P. Zheng, and J. Huang, “Automatic Asymmetric Embedding Cost Learning via Generative Adversarial Networks”, in *Proceedings of the ACM International Conference on Multimedia*, 2023, 8316–26.

[62] F. Huang, B. Li, and J. Huang, “Attack LSB Matching Steganography by Counting Alteration Rate of the Number of Neighbourhood Gray Levels”, in *Proceedings of the IEEE International Conference on Image Processing*, Vol. 1, 2007, I - 401-I –404.

[63] J. Huang, J. Ni, L. Wan, and J. Yan, “A Customized Convolutional Neural Network with Low Model Complexity for JPEG Steganalysis”, in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2019, 198–203.

[64] S. Huang, M. Zhang, Y. Ke, and X. Bi, “Image steganalysis method based on saliency detection”, *Journal of Computer Applications*, 41(2), 2021, 441.

[65] T. Itzhaki, Y. Yousfi, and J. Fridrich, “Data Augmentation for JPEG Steganalysis”, in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2021, 1–6.

[66] J. Kodovský and J. Fridrich, “Ensemble classifiers for steganalysis of digital media”, *IEEE Transactions on Information Forensics and Security*, 7(2), 2012, 432–44.

[67] J. Jia, M. Luo, J. Liu, W. Ren, and L. Wang, “Multiperspective Progressive Structure Adaptation for JPEG Steganography Detection Across Domains”, *IEEE Transactions on Neural Networks and Learning Systems*, 33(8), 2022, 3660–74.

[68] J. Jia, L. Zhai, W. Ren, L. Wang, and Y. Ren, “An Effective Imbalanced JPEG Steganalysis Scheme Based on Adaptive Cost-Sensitive Feature Learning”, *IEEE Transactions on Knowledge and Data Engineering*, 34(3), 2022, 1038–52.

[69] J. Jia, L. Zhai, W. Ren, L. Wang, Y. Ren, and L. Zhang, “Transferable heterogeneous feature subspace learning for JPEG mismatched steganalysis”, *Pattern Recognition*, 100, 2020, 107105.

[70] A. Ker, “Steganalysis of LSB matching in grayscale images”, *IEEE Signal Processing Letters*, 12(6), 2005, 441–4.

[71] A. D. Ker, “A General Framework for Structural Steganalysis of LSB Replacement”, in *Information Hiding*, 2005, 296–311.

[72] J. Kodovský and J. Fridrich, “Steganalysis of JPEG images using rich models”, in *Media Watermarking, Security, and Forensics*, Vol. 8303, 2012, 83030A.

[73] J. Kodovskŷ and J. Fridrich, “Effect of image downsampling on steganographic security”, *IEEE Transactions on Information Forensics and Security*, 9(5), 2014, 752–62.

[74] J. Kodovský, V. Sedighi, and J. Fridrich, “Study of cover source mismatch in steganalysis and ways to mitigate its impact”, in *Media Watermarking, Security, and Forensics*, Vol. 9028, 2014, 90280J.

[75] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning”, *Nature*, 521(7553), 2015, 436–44.

[76] B. Li, J. Huang, and Y. Q. Shi, “Textural features based universal steganalysis”, in *Proceedings of the SPIE on Security, Forensics, Steganography and Watermarking of Multimedia*, Vol. 6819, 2008, 681912.

[77] B. Li, Z. Li, S. Zhou, S. Tan, and X. Zhang, “New steganalytic features for spatial image steganography based on derivative filters and threshold LBP operator”, *IEEE Transactions on Information Forensics and Security*, 13(5), 2018, 1242–57.

[78] B. Li, M. Wang, J. Huang, and X. Li, “A new cost function for spatial image steganography”, in *Proceedings of the IEEE International Conference on Image Processing*, IEEE, 2014, 4206–10.

[79] B. Li, M. Wang, X. Li, S. Tan, and J. Huang, “A strategy of clustering modification directions in spatial image steganography”, *IEEE Transactions on Information Forensics and Security*, 10(9), 2015, 1905–17.

[80] B. Li, W. Wei, A. Ferreira, and S. Tan, “ReST-Net: Diverse Activation Modules and Parallel Subnets-Based CNN for Spatial Image Steganalysis”, *IEEE Signal Processing Letters*, 25(5), 2018, 650–4.

[81] F. Li, Z. Yu, and C. Qin, “GAN-based spatial image steganography with cross feedback mechanism”, *Signal Processing*, 190, 2022, 108341.

[82] F. Li, Y. Zeng, X. Zhang, and C. Qin, “Ensemble stego selection for enhancing image steganography”, *IEEE Signal Processing Letters*, 29, 2022, 702–6.

[83] H. Li, X. Luo, and Y. Zhang, “Improving CoatNet for Spatial and JPEG Domain Steganalysis”, in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2023, 1241–6.

[84] P. Li, J. Xie, Q. Wang, and Z. Gao, “Towards Faster Training of Global Covariance Pooling Networks by Iterative Matrix Square Root Normalization”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[85] Q. Li, G. Feng, Y. Ren, and X. Zhang, "Embedding Probability Guided Network for Image Steganalysis", *IEEE Signal Processing Letters*, 28, 2021, 1095–9.

[86] Q. Li, Z. Shao, S. Tan, J. Zeng, and B. Li, "Non-structured Pruning for Deep-learning based Steganalytic Frameworks", in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2019, 1735–9.

[87] S. Li, Z. Wang, X. Zhang, and X. Zhang, "Robust Image Steganography Against General Downsampling Operations With Lossless Secret Recovery", *IEEE Transactions on Dependable and Secure Computing*, 21(1), 2024, 340–52.

[88] W. Li, H. Wang, and Y. Chen, "From Cover to Immucover: Adversarial Steganography via Immunized Cover Construction", *IEEE Transactions on Fuzzy Systems*, 2023.

[89] W. Li, H. Wang, Y. Chen, S. M. Abdullahi, and J. Luo, "Constructing immunized stego-image for secure steganography via artificial immune system", *IEEE Transactions on Multimedia*, 2023.

[90] W. Li, B. Li, W. Zhang, and S. Zhang, "Quaternary Quantized Gaussian Modulation with Optimal Polarity Map Selection for JPEG Steganography", *IEEE Transactions on Information Forensics and Security*, 2023.

[91] W. Li, S. Wu, B. Li, W. Tang, and X. Zhang, "Payload-independent direct cost learning for image steganography", *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[92] W. Li, W. Zhang, K. Chen, W. Zhou, and N. Yu, "Defining joint distortion for JPEG steganography", in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2018, 5–16.

[93] W. Li, W. Zhang, L. Li, H. Zhou, and N. Yu, "Designing Near-Optimal Steganographic Codes in Practice Based on Polar Codes", *IEEE Transactions on Communications*, 68(7), 2020, 3948–62.

[94] X. Li, W. Wang, X. Hu, and J. Yang, "Selective Kernel Networks", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019.

[95] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, 1778–87.

[96] X. Liao, Y. Yu, B. Li, Z. Li, and Z. Qin, "A new payload partition strategy in color image steganography", *IEEE Transactions on Circuits and Systems for Video Technology*, 30(3), 2019, 685–96.

[97] J. Liu, G. Jiao, and X. Sun, "Feature Passing Learning for Image Steganalysis", *IEEE Signal Processing Letters*, 29, 2022, 2233–7.

[98] M. Liu, W. Luo, P. Zheng, and J. Huang, “A new adversarial embedding method for enhancing image steganography”, *IEEE Transactions on Information Forensics and Security*, 16, 2021, 4621–34.

[99] M. Liu, T. Song, W. Luo, P. Zheng, and J. Huang, “Adversarial steganography embedding via stego generation and selection”, *IEEE Transactions on Dependable and Secure Computing*, 2022.

[100] T. Liu, Y. Chen, and W. Gu, “Deniable Diffusion Generative Steganography”, in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2023, 67–71.

[101] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021, 10012–22.

[102] M. Long, H. Zhu, J. Wang, and M. I. Jordan, “Deep transfer learning with joint adaptation networks”, in *Proceedings of the International Conference on Machine Learning*, 2017, 2208–17.

[103] G. Luo, P. Wei, S. Zhu, X. Zhang, Z. Qian, and S. Li, “Image Steganalysis with Convolutional Vision Transformer”, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, 3089–93.

[104] J. Luo, P. He, J. Liu, H. Wang, C. Wu, Y. Chen, W. Li, and J. Li, “Content-adaptive Adversarial Embedding for Image Steganography Using Deep Reinforcement Learning”, in *Proceedings of the IEEE International Conference on Multimedia and Expo*, IEEE, 2023, 49–54.

[105] J. Luo, P. He, J. Liu, H. Wang, C. Wu, C. Yuan, and Q. Xia, “Improving security for image steganography using content-adaptive adversarial perturbations”, *Applied Intelligence*, 53(12), 2023, 16059–76.

[106] W. Luo, F. Huang, and J. Huang, “Edge Adaptive Image Steganography Based on LSB Matching Revisited”, *IEEE Transactions on Information Forensics and Security*, 5(2), 2010, 201–14.

[107] W. Luo, Y. Wang, and J. Huang, “Security Analysis on Spatial ±1 Steganography for JPEG Decompressed Images”, *IEEE Signal Processing Letters*, 18(1), 2011, 39–42.

[108] Y. Luo, J. Du, K. Yan, and S. Ding, “LaRE^ 2: Latent Reconstruction Error Based Method for Diffusion-Generated Image Detection”, *arXiv preprint arXiv:2403.17465*, 2024.

[109] S. Ma, Q. Guan, X. Zhao, and Y. Liu, “Adaptive spatial steganography based on probability-controlled adversarial examples”, *arXiv preprint arXiv:1804.02691*, 2018.

[110] S. Ma, Q. Guan, X. Zhao, and Y. Liu, “Weakening the detecting capability of CNN-based steganalysis”, *arXiv preprint arXiv:1803.10889*, 2018.

[111] S. Ma, X. Zhao, and Y. Liu, “Adaptive spatial steganography based on adversarial examples”, *Multimedia Tools and Applications*, 78(22), 2019, 32503–22.

[112] Y. Ma, X. Luo, X. Li, Z. Bao, and Y. Zhang, “Selection of Rich Model Steganalysis Features Based on Decision Rough Set α -Positive Region Reduction”, *IEEE Transactions on Circuits and Systems for Video Technology*, 29(2), 2019, 336–50.

[113] D. Megías and D. Lerch-Hostalot, “Subsequent Embedding in Targeted Image Steganalysis: Theoretical Framework and Practical Applications”, *IEEE Transactions on Dependable and Secure Computing*, 20(2), 2023, 1403–21.

[114] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, “Asynchronous methods for deep reinforcement learning”, in *Proceedings of the International Conference on Machine Learning*, 2016, 1928–37.

[115] H. Mo, T. Song, B. Chen, W. Luo, and J. Huang, “Enhancing JPEG steganography using iterative adversarial examples”, in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, IEEE, 2019, 1–6.

[116] X. Mo, S. Tan, B. Li, and J. Huang, “MCTSteg: A Monte Carlo tree search-based reinforcement learning framework for universal non-additive steganography”, *IEEE Transactions on Information Forensics and Security*, 16, 2021, 4306–20.

[117] X. Mo, S. Tan, W. Tang, B. Li, and J. Huang, “ReLOAD: Using reinforcement learning to optimize asymmetric distortion for additive steganography”, *IEEE Transactions on Information Forensics and Security*, 18, 2023, 1524–38.

[118] Y. Peng, D. Hu, Y. Wang, K. Chen, G. Pei, and W. Zhang, “StegaDDPM: Generative Image Steganography based on Denoising Diffusion Probabilistic Model”, in *Proceedings of the ACM International Conference on Multimedia*, 2023, 7143–51.

[119] T. Pevny, P. Bas, and J. Fridrich, “Steganalysis by Subtractive Pixel Adjacency Matrix”, *IEEE Transactions on Information Forensics and Security*, 5(2), 2010, 215–24.

[120] T. Pevnỳ, T. Filler, and P. Bas, “Using high-dimensional image models to perform highly undetectable steganography”, in *Proceedings of the International Conference on Information Hiding*, Springer, 2010, 161–77.

[121] C. V. Priscilla and V. HemaMalini, “Effective Analysis of Real World Stego Images through Deep Learning Techniques”, in *International Conference on Applied Artificial Intelligence and Computing*, 2024, 780–5.

[122] Y. Qian, J. Dong, W. Wang, and T. Tan, "Deep learning for steganalysis via convolutional neural networks", in *Proceedings of IS&T/SPIE Electronic Imaging 2015 (Media Watermarking, Security, and Forensics)*, 2015, 94090J-1–94090J-10.

[123] T. Qiao, X. Luo, T. Wu, M. Xu, and Z. Qian, "Adaptive Steganalysis Based on Statistical Model of Quantized DCT Coefficients for JPEG Images", *IEEE Transactions on Dependable and Secure Computing*, 18(6), 2021, 2736–51.

[124] C. Qin, W. Zhang, H. Zhou, J. Liu, Y. He, and N. Yu, "Robustness enhancement against adversarial steganography via steganalyzer outputs", *Journal of Information Security and Applications*, 68, 2022, 103252.

[125] C. Qin, N. Zhao, W. Zhang, and N. Yu, "Patch Steganalysis: A Sampling Based Defense Against Adversarial Steganography", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, 3079–83.

[126] X. Qin, B. Li, S. Tan, W. Tang, and J. Huang, "Gradually enhanced adversarial perturbations on color pixel vectors for image steganography", *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8), 2022, 5110–23.

[127] X. Qin, S. Tan, W. Tang, B. Li, and J. Huang, "Image steganography based on iterative adversarial perturbations onto a synchronized-directions sub-image", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, 2705–9.

[128] G. Quentin, B. Patrick, C. Rémi, and B. Dirk, "The Cover Source Mismatch Problem in Deep-Learning Steganalysis", in *Proceedings of the European Signal Processing Conference*, 2022, 1032–6.

[129] W. Ren, L. Zhai, J. Jia, L. Wang, and L. Zhang, "Learning selection channels for image steganalysis in spatial domain", *Neurocomputing*, 401, 2020, 78–90, ISSN: 0925-2312.

[130] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules", in *Proceedings of the International Conference on Neural Information Processing Systems*, 2017, 3859–69.

[131] S. Sankaranarayanan, A. Jain, R. Chellappa, and S. N. Lim, "Regularizing Deep Networks Using Efficient Layerwise Adversarial Training", *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

[132] V. Sedighi, R. Cogranne, and J. Fridrich, "Content-adaptive steganography by minimizing statistical detectability", *IEEE Transactions on Information Forensics and Security*, 11(2), 2015, 221–34.

[133] Y. Shi, G. Xuan, D. Zou, J. Gao, C. Yang, Z. Zhang, P. Chai, W. Chen, and C. Chen, "Image steganalysis based on moments of characteristic functions using wavelet decomposition, prediction-error image, and

neural network", in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2005, 4.

[134] Y. Q. Shi, P. Sutthiwan, and L. Chen, "Textural Features for Steganalysis", in *Information Hiding*, Springer Berlin Heidelberg, 2013, 63–77.

[135] T. Song, M. Liu, W. Luo, and P. Zheng, "Enhancing image steganography via stego generation and selection", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2021, 2695–9.

[136] X. Song, F. Liu, C. Yang, X. Luo, and Y. Zhang, "Steganalysis of Adaptive JPEG Steganography Using 2D Gabor Filters", in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2015, 15–23.

[137] A. Su, X. He, and X. Zhao, "JPEG steganalysis based on ResNeXt with Gauss partial derivative filters", *Multimedia Tools and Applications*, 80(3), 2021, 3349–66.

[138] W. Su, J. Ni, X. Hu, and J. Fridrich, "Image steganography with symmetric embedding using Gaussian Markov random field model", *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3), 2020, 1001–15.

[139] W. Su, J. Ni, X. Hu, and J. Huang, "New design paradigm of distortion cost function for efficient JPEG steganography", *Signal Processing*, 190, 2022, 108319.

[140] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation", in *Proceedings of the International Conference on Neural Information Processing Systems*, 1999, 1057–63.

[141] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks", *arXiv preprint arXiv:1312.6199*, 2013.

[142] J. Tan, X. Liao, J. Liu, Y. Cao, and H. Jiang, "Channel attention image steganography with generative adversarial networks", *IEEE Transactions on Network Science and Engineering*, 9(2), 2021, 888–903.

[143] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks", in *Proceedings of the International Conference on Machine Learning*, PMLR, 2019, 6105–14.

[144] S. Tan and B. Li, "Stacked convolutional auto-encoders for steganalysis of digital images", in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2014, 1–4.

[145] S. Tan, Q. Li, L. Li, B. Li, and J. Huang, "STD-NET: Search of Image Steganalytic Deep-Learning Architecture via Hierarchical Tensor Decomposition", *IEEE Transactions on Dependable and Secure Computing*, 20(3), 2023, 2657–73.

[146] S. Tan, W. Wu, Z. Shao, Q. Li, B. Li, and J. Huang, "CALPA-NET: Channel-Pruning-Assisted Deep Residual Network for Steganalysis of Digital Images", *IEEE Transactions on Information Forensics and Security*, 16, 2021, 131–46.

[147] S. Tan, H. Zhang, B. Li, and J. Huang, "Pixel-decimation-assisted steganalysis of synchronize-embedding-changes steganography", *IEEE Transactions on Information Forensics and Security*, 12(7), 2017, 1658–70.

[148] W. Tang, B. Li, M. Barni, J. Li, and J. Huang, "An automatic cost learning framework for image steganography using deep reinforcement learning", *IEEE Transactions on Information Forensics and Security*, 16, 2020, 952–67.

[149] W. Tang, B. Li, M. Barni, J. Li, and J. Huang, "Improving cost learning for JPEG steganography by exploiting JPEG domain knowledge", *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6), 2021, 4081–95.

[150] W. Tang, B. Li, W. Li, Y. Wang, and J. Huang, "Reinforcement learning of non-additive joint steganographic embedding costs with attention mechanism", *Science China Information Sciences*, 66(3), 2023, 132305.

[151] W. Tang, B. Li, W. Luo, and J. Huang, "Clustering steganographic modification directions for color components", *IEEE Signal Processing Letters*, 23(2), 2015, 197–201.

[152] W. Tang, B. Li, S. Tan, M. Barni, and J. Huang, "CNN-based adversarial embedding for image steganography", *IEEE Transactions on Information Forensics and Security*, 14(8), 2019, 2074–87.

[153] W. Tang, H. Li, W. Luo, and J. Huang, "Adaptive steganalysis against WOW embedding algorithm", in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2014, 91–6.

[154] W. Tang, S. Tan, B. Li, and J. Huang, "Automatic steganographic distortion learning using a generative adversarial network", *IEEE Signal Processing Letters*, 24(10), 2017, 1547–51.

[155] C. F. Tsang and J. Fridrich, "Steganalyzing Images of Arbitrary Size with CNNs", *Electronic Imaging*, 30(7), 2018, 121-1–121-1.

[156] P. Wang, F. Liu, and C. Yang, "Towards feature representation for steganalysis of spatial steganography", *Signal Processing*, 169, 2020, 107422.

[157] X. Wang, J. Li, and Y. Song, "DDAC: a feature extraction method for model of image steganalysis based on convolutional neural network", *Journal on Communications*, 43(5), 68, 2022, 68–81.

[158] Y. Wang, W. Zhang, W. Li, and N. Yu, "Non-additive cost functions for JPEG steganography based on block boundary maintenance", *IEEE Transactions on Information Forensics and Security*, 16, 2020, 1117–30.

[159] Y. Wang, W. Zhang, W. Li, X. Yu, and N. Yu, "Non-additive cost functions for color image steganography based on inter-channel correlations and differences", *IEEE Transactions on Information Forensics and Security*, 15, 2019, 2081–95.

[160] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "DIRE for Diffusion-Generated Image Detection", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, 22388–98.

[161] Z. Wang, G. Feng, Z. Qian, and X. Zhang, "JPEG steganography with content similarity evaluation", *IEEE Transactions on Cybernetics*, 2022.

[162] Z. Wang, X. Zhang, and Z. Yin, "Hybrid distortion function for JPEG steganography", *Journal of Electronic Imaging*, 25(5), 2016, 50501–1.

[163] C. J. Watkins and P. Dayan, "Q-learning", *Machine learning*, 8, 1992, 279–92.

[164] K. Wei, W. Luo, M. Liu, and M. Ye, "Residual guided coordinate attention for selection channel aware image steganalysis", *Multimedia Systems*, 2023, 1–11.

[165] K. Wei, W. Luo, and L. Minglin, "Spatial Color Image Steganalysis Based on Central Difference Convolution and Attention", *Journal of Software*, 2024, 1.

[166] K. Wei, W. Luo, S. Tan, and J. Huang, "CTNet: A Convolutional Transformer Network for Color Image Steganalysis", *Journal of Computer Science and Technology*, 2023.

[167] K. Wei, W. Luo, S. Tan, and J. Huang, "Universal Deep Network for Steganalysis of Color Image Based on Channel Representation", *IEEE Transactions on Information Forensics and Security*, 17, 2022, 3022–36.

[168] Q. Wei, Z. Yin, Z. Wang, and X. Zhang, "Distortion function based on residual blocks for JPEG steganography", *Multimedia Tools and Applications*, 77, 2018, 17875–88.

[169] S. Weng, M. Chen, L. Yu, and S. Sun, "Lightweight and Effective Deep Image Steganalysis Network", *IEEE Signal Processing Letters*, 29, 2022, 1888–92.

[170] S. Weng, S. Sun, and L. Yu, "Fast SwT-Based Deep Steganalysis Network for Arbitrary-Sized Images", *IEEE Signal Processing Letters*, 30, 2023, 1782–6.

[171] H. Wu, F. Li, X. Zhang, and K. Wu, "GAN-based steganography with the concatenation of multiple feature maps", in *Proceedings of the International Workshop of Digital Forensics and Watermarking*, Springer, 2020, 3–17.

[172] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing Convolutions to Vision Transformers", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021, 22–31.

[173] L. Wu, P. Cui, J. Pei, L. Zhao, and X. Guo, "Graph Neural Networks: Foundation, Frontiers and Applications", in *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*, 2022, 4840–1.

[174] S. Wu, S.-h. Zhong, and Y. Liu, "A Novel Convolutional Neural Network for Image Steganalysis With Shared Normalization", *IEEE Transactions on Multimedia*, 22(1), 2020, 256–70.

[175] T. Wu, L. Wang, L. Zhai, C. Fang, and M. Zhang, "Progressive selection-channel networks for image steganalysis", *International Journal of Intelligent Systems*, 37(10), 2022, 7444–58.

[176] Z. Xi, W. Huang, K. Wei, W. Luo, and P. Zheng, "AI-Generated Image Detection using a Cross-Attention Enhanced Dual-Stream Network", in *Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2023, 1463–70.

[177] C. Xia, Q. Guan, X. Zhao, and K. Wu, "Improved JPEG Phase-Aware Steganalysis Features Using Multiple Filter Sizes and Difference Images", *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11), 2020, 4100–13.

[178] G. Xie, J. Ren, S. Marshall, H. Zhao, and R. Li, "A novel gradient-guided post-processing method for adaptive image steganography", *Signal Processing*, 203, 2023, 108813.

[179] G. Xie, J. Ren, S. Marshall, H. Zhao, R. Li, and R. Chen, "Self-attention enhanced deep residual network for spatial image steganalysis", *Digital Signal Processing*, 139, 2023, 104063.

[180] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[181] G. Xu, H. Z. Wu, and Y. Q. Shi, "Structural Design of Convolutional Neural Networks for Steganalysis", *IEEE Signal Processing Letters*, 23(5), 2016, 708–12.

[182] G. Xu, "Deep Convolutional Neural Network to Detect J-UNIWARD", in *Proceedings of ACM Information Hiding and Multimedia Security Workshop*, 2017, 67–73.

[183] Y. Xu, L. Xie, X. Zhang, X. Chen, G.-J. Qi, Q. Tian, and H. Xiong, "PC-DARTS: Partial Channel Connections for Memory-Efficient Architecture Search", in *Proceedings of the International Conference on Learning Representations*, 2020.

[184] J. Yang, X. Kang, E. K. Wong, and Y.-Q. Shi, "Deep Learning with Feature Reuse for JPEG Image Steganalysis", in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2018, 533–8.

[185] J. Yang, Y. Liao, F. Shang, X. Kang, and Y.-Q. Shi, "JPEG Steganography with Embedding Cost Learning and Side-Information Estimation", *arXiv preprint arXiv:2107.13151*, 2021.

[186] J. Yang, B. Lu, L. Xiao, X. Kang, and Y.-Q. Shi, "Reinforcement Learning Aided Network Architecture Generation for JPEG Image Steganalysis", in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2020, 23–32.

[187] J. Yang, D. Ruan, J. Huang, X. Kang, and Y.-Q. Shi, "An embedding cost learning framework using GAN", *IEEE Transactions on Information Forensics and Security*, 15, 2019, 839–51.

[188] J. Yang, D. Ruan, X. Kang, and Y.-Q. Shi, "Towards automatic embedding cost learning for JPEG steganography", in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2019, 37–46.

[189] L. Yang, M. Men, Y. Xue, J. Wen, and P. Zhong, "Transfer subspace learning based on structure preservation for JPEG image mismatched steganalysis", *Signal Processing: Image Communication*, 90, 2021, 116052.

[190] Z. Yang, K. Wang, S. Ma, Y. Huang, X. Kang, and X. Zhao, "IStego100K: Large-scale Image Steganalysis Dataset", in *Proceedings of the International Workshop on Digital Watermarking*, 2019.

[191] J. Ye, J. Ni, and Y. Yi, "Deep Learning Hierarchical Representations for Image Steganalysis", *IEEE Transactions on Information Forensics and Security*, 12(11), 2017, 2545–57.

[192] M. Ye, D. Huang, K. Wei, and W. Luo, "A Novel Residual-Guided Learning Method for Image Steganography", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, 4565–9.

[193] M. Yedroudj, M. Chaumont, F. Comby, A. Oulad Amara, and P. Bas, "Pixels-off: Data-augmentation Complementary Solution for Deep-learning Steganalysis", in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2020, 39–48.

[194] M. Yedroudj, F. Comby, and M. Chaumont, "Yedroudj-Net: An Efficient CNN for Spatial Steganalysis", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, 2092–6.

[195] W. You, H. Zhang, and X. Zhao, "A Siamese CNN for Image Steganalysis", *IEEE Transactions on Information Forensics and Security*, 16, 2021, 291–306.

[196] Y. Yousfi, J. Butora, and J. Fridrich, “CNN Steganalyzers Leverage Local Embedding Artifacts”, in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2021, 1–6.

[197] Y. Yousfi, J. Butora, J. Fridrich, and C. Fuji Tsang, “Improving EfficientNet for JPEG Steganalysis”, in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2021, 149–57.

[198] Y. Yousfi, J. Butora, E. Khvedchenya, and J. Fridrich, “ImageNet Pre-trained CNNs for JPEG Steganalysis”, in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2020, 1–6.

[199] Y. Yousfi and J. Fridrich, “An Intriguing Struggle of CNNs in JPEG Steganalysis and the OneHot Solution”, *IEEE Signal Processing Letters*, 27, 2020, 830–4.

[200] Y. Yousfi and J. Fridrich, “JPEG Steganalysis Detectors Scalable With Respect to Compression Quality”, *Electronic Imaging*, 32(4), 2020, 75-1–75-1.

[201] I.-J. Yu, W. Ahn, S.-H. Nam, and H.-K. Lee, “BitMix: data augmentation for image steganalysis”, *Electronics Letters*, 56(24), 2020, 1311–4.

[202] L. Yu, Y. Li, S. Weng, H. Tian, and J. Liu, “Adaptive multi-teacher softened relational knowledge distillation framework for payload mismatch in image steganalysis”, *Journal of Visual Communication and Image Representation*, 95, 2023, 103900.

[203] L. Yu, S. Weng, M. Chen, and Y. Wei, “RCDD: Contrastive domain discrepancy with reliable steganalysis labeling for cover source mismatch”, *Expert Systems with Applications*, 237, 2024, 121543.

[204] C. Yuan, H. Wang, P. He, J. Luo, and B. Li, “GAN-based image steganography for enhancing security via adversarial attack and pixel-wise deep fusion”, *Multimedia Tools and Applications*, 81(5), 2022, 6681–701.

[205] J. Zeng, S. Tan, B. Li, and J. Huang, “Large-scale JPEG steganalysis using hybrid deep-learning framework”, *IEEE Transactions on Information Forensics and Security*, 13(5), 2018, 1242–57.

[206] J. Zeng, S. Tan, G. Liu, B. Li, and J. Huang, “WISERNet: Wider separate-then-reunion network for steganalysis of color images”, *IEEE Transactions on Information Forensics and Security*, 14(10), 2019, 2735–48.

[207] K. Zeng, K. Chen, W. Zhang, and Y. Wang, “Upward Robust Steganography Based on Overflow Alleviation”, *IEEE Transactions on Multimedia*, 26, 2024, 299–312.

[208] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, “Epsanet: An efficient pyramid split attention block on convolutional neural network”, *arXiv preprint arXiv:2105.14447*, 2021.

[209] J. Zhang, K. Chen, C. Qin, W. Zhang, and N. Yu, “AAS: Automatic Virtual Data Augmentation for Deep Image Steganalysis”, *IEEE Transactions on Dependable and Secure Computing*, 2023, 1–13.

[210] J. Zhang, K. Chen, C. Qin, W. Zhang, and N. Yu, “Distribution-Preserving-Based Automatic Data Augmentation for Deep Image Steganalysis”, *IEEE Transactions on Multimedia*, 24, 2022, 4538–50.

[211] K. A. Zhang, A. Cuesta-Infante, L. Xu, and K. Veeramachaneni, “SteganoGAN: High capacity image steganography with GANs”, *arXiv preprint arXiv:1901.03892*, 2019.

[212] R. Zhang, F. Zhu, J. Liu, and G. Liu, “Depth-Wise Separable Convolutions and Multi-Level Pooling for an Efficient Spatial CNN-Based Steganalysis”, *IEEE Transactions on Information Forensics and Security*, 15, 2020, 1138–50.

[213] R. Zhang, S. Dong, and J. Liu, “Invisible steganography via generative adversarial networks”, *Multimedia Tools and Applications*, 78(7), 2019, 8559–75.

[214] S. Zhang, H. Zhang, X. Zhao, and H. Yu, “A Deep Residual Multi-scale Convolutional Network for Spatial Steganalysis”, in *Digital Forensics and Watermarking*, 2019, 40–52.

[215] W. Zhang, Z. Zhang, L. Zhang, H. Li, and N. Yu, “Decomposing joint distortion for adaptive steganography”, *IEEE Transactions on Circuits and Systems for Video Technology*, 27(10), 2016, 2274–80.

[216] X. Zhang, X. Zhang, and G. Feng, “Image Steganalysis Network Based on Dual-Attention Mechanism”, *IEEE Signal Processing Letters*, 30, 2023, 1287–91.

[217] X. Zhang, X. Kong, P. Wang, and B. Wang, “Cover-Source Mismatch in Deep Spatial Steganalysis”, in *Digital Forensics and Watermarking*, 2020, 71–83.

[218] Y. Zhang, X. Luo, J. Wang, W. Lu, C. Yang, and F. Liu, “Research progress on digital image robust steganography”, *Journal of Image and Graphics*, 27(1), 2022.

[219] Y. Zhang, W. Zhang, K. Chen, J. Liu, Y. Liu, and N. Yu, “Adversarial examples against deep neural network based steganalysis”, in *Proceedings of the ACM Workshop on information hiding and multimedia security*, 2018, 67–72.

[220] H. Zheng, X. Li, D. Ruan, X. Kang, and Y.-Q. Shi, “Comparison of DCT and Gabor filters in residual extraction of CNN based JPEG steganalysis”, in *Proceedings of the International Workshop on Digital Watermarking*, 2019, 29–39.

[221] Z. Zheng, Y. Hu, Y. Bin, X. Xu, Y. Yang, and H. T. Shen, “Composition-Aware Image Steganography Through Adversarial Self-Generated Supervision”, *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[222] S. Zhou, W. Tang, S. Tan, and B. Li, "Content-adaptive steganalysis via augmented utilization of selection-channel information", in *Proceedings of the International Workshop on Digital Watermarking*, Springer, 2018, 261–74.

[223] W. Zhou, W. Zhang, and N. Yu, "A new rule for cost reassignment in adaptive steganography", *IEEE Transactions on Information Forensics and Security*, 12(11), 2017, 2654–67.

[224] Z. Zhou, C. Ding, J. Li, F. Peng, and X. Zhang, "Research on Generative Steganography", *Chinese Journal of Computers*, 46(9), 2023.

[225] Z. Zhou, Y. Su, J. Li, K. Yu, Q. M. J. Wu, Z. Fu, and Y. Shi, "Secret-to-Image Reversible Transformation for Generative Steganography", *IEEE Transactions on Dependable and Secure Computing*, 20(5), 2023, 4118–34.

[226] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks", in *Proceedings of the European Conference on Computer Vision*, 2018, 657–72.

[227] Y. Zhu, X. Wang, H.-S. Chen, R. Salloum, C.-C. J. Kuo, *et al.*, "Green steganalyzer: A green learning approach to image steganalysis", *APSIPA Transactions on Signal and Information Processing*, 12(1), 2023.