Journal of Science and Engineering Applications Vol. 8, No. 2, pp. 89–172, October 2026

Journal of Science and Engineering Applications

Contents are available at https://jsea.iujournals.com

The Deep Learning Shift in Steganalysis: A Modern Review

Hasan Karam. Al-Saeedi[1], Asaad Noori Hashim[2]

[1] Department of Computer Science, College of Computer Science and Mathematics, University of Kufa, Najaf, Iraq

ARTICLE INFORMATION

Received date: 22-04-2026
Revised date: 29-06-2026
Accepted date: 10-07-2026

Keywords

steganalysis
convolutional neural networks
generalization

ABSTRACT

The rapid evolution of adaptive image steganography has made hidden payloads increasingly difficult to detect, particularly when embedding traces are weak, content-dependent, and affected by image-source variation. This review examines the shift in image steganalysis from hand-crafted statistical features to deep learning-based detection, with emphasis on convolutional detectors, residual architectures, hybrid CNN-classifier models, and domain-adaptive learning. Representative networks, including Qian-Net, Xu-Net, Ye-Net, Yedroudj-Net, SRNet, and Zhu-Net, are compared in terms of preprocessing strategy, nonlinearity, classifier design, and benchmark setting. The reviewed evidence shows that deep residual models can achieve strong accuracy on controlled natural-image benchmarks, but their performance often declines under cover-source mismatch, heterogeneous image sources, adversarial perturbations, and computational constraints. The main outcome of this review is a structured synthesis of current methodological trends, persistent limitations, and practical mitigation strategies. It highlights the need for source-diverse benchmarks, reproducible evaluation protocols, lightweight architectures, model compression, knowledge distillation, and domain adaptation to support more robust and deployable steganalysis systems.

1 INTRODUCTION

Digital steganography is the art of concealing secret data within digital multimedia, often image-based, files. Since the steganographic existence of data concealing communications is designed to be secret from unwelcome third parties, most steganographic communications remain undetected. Though, the effort to detect steganographic data is called steganalysis, which also aims to describe how the data was embedded within the object. The substitution of the Least Significant Bits (LSB) was only the beginning of steganographic methods, followed by content-adaptive and learning-based methods. So, the payloads embedded within steganographic cover images left less sharp

* Corresponding authors: Department of Computer Science, College of Computer Science and Mathematics, University of Kufa, Najaf, Iraq
E-mail addresses: hasank.alsaeedi@student.uokufa.edu.iq (Hasan Karam. Al-Saeedi)

traces, and became even more difficult to detect. This advancement of techniques made modeling the traces to devise a detection method almost impossible, especially when the only relied upon tools to build the method were handcrafted statistical features[1,2].

In the last decade, the progress in the field has also witnessed a shift in the development of steganalysis methods that use deep learning (DL) techniques, and methods that rely on the mathematical modeling of the Noise Residual. This new approach offers highly efficient methods for various detection problems, but also poses a number of problems that didn't exist with previous methods. Many of the modern detection methods became highly ineffective if restrained to the same data sets, if the images were captured in significantly different scenes, or if a number of advanced manipulations were performed. Modern detection methods also often require significant computational resources and an extensive training set of data [2,3].

This paper offers an insight into deep learning based methods for image steganalysis. It makes four contributions.

The progression of methods is documented, capturing shifts from handcrafted features to CNNs, residual approaches, hybrids, and adaptive domain detectors. A classification framework based on Preprocessing, Nonlinearity, Classifier Head, and Evaluation Dataset is also provided. Performance on widely used benchmark datasets is aggregated along with the analysis on the generalization gap, and is concentrated on various types of data, both naturally occurring and AI-generated. Labor of unsolved problems as well as potential solutions is also included. The following sections of the paper progress as follows. Section 2 presents the literature review. Section 3 presents AI-based steganalysis. Section 4 presents AI-based detection. Section 5 presents benchmark datasets, and evaluation techniques and protocols. Section 6 presents a synthesis of the available literature pertaining to qualitative comparisons. Section 7 presents the problems relating to generalization and computation. Section 8 presents the conclusions along with the constraints and future directions.

1.1 Defining the Steganalysis Problem

Steganalysis seeks to detect secret data hidden within digital multimedia. It is the analytical counterpart of steganography, which conceals a message so that only the intended recipient is aware of its presence. In practice, the steganalyst must decide whether an observed object is a cover or a stego object, and may also attempt to infer properties of the embedding method [1]. As illustrated in Figures 1 and 2, steganalysis methods can be broadly grouped into generic and specific detectors. Generic detectors are designed to operate across multiple embedding methods and image conditions, whereas specific detectors target a particular embedding algorithm and may achieve higher accuracy only within that restricted setting[2,3] .

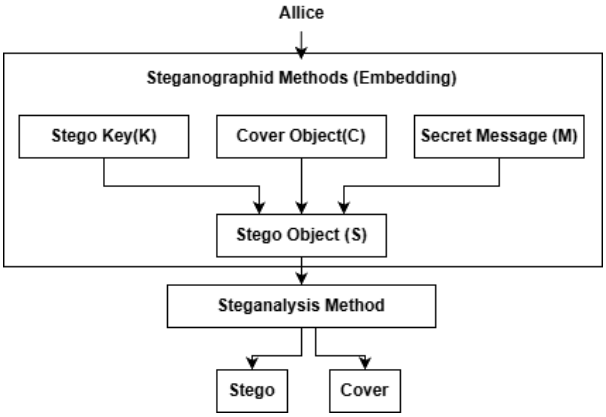


Figure 1. General steganographic communication channel, where a cover object, secret message, and optional stego key are used to produce a stego object that is later inspected by a steganalysis method for cover/stego classification [2].

In Figure 1 is a summary of the communications channel of steganography and the detection phase. Converts are combined with the secret message and an optional stego key during the embedding stage to produce a stego message object. The stego object is then transmitted and later subjected to analysis by the steganalyst.

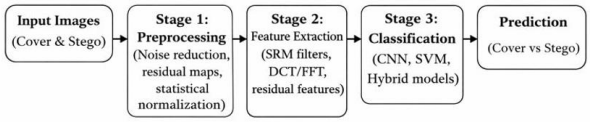


Figure 2. Shows a multi-stage steganalysis pipeline. This pipeline includes a pre-processing stage, a stage for either residual or statistical feature extraction, and a stage for final classification utilizing CNN-based, statistical, or hybrid models[3].

Pipelines for Steganalysis in Contemporary,useThe first few steps of a state-of-the-art steganalysis pipeline include: preprocessing (residual extraction, noise suppression), feature extraction (using learned descriptive residuals, SRM, FFT, etc.), and feature extraction (residuals extraction, noise suppression). Then a classifier is applied and cover content is identified and differentiated from stego content. CNNs, hybrid classifiers and statistical models are used. In areas of research including Digital Forensics and

Cybersecurity, pipelines such as these are researched and developed in order to detect covert communications and find statistically minor changes in image data and features [3,4].

1.2 Adaptive Embedding Techniques

The sophisticated steganographic algorithms that have developed in the years since plain LSB substitution [5] have been developed, have opened many new avenues in steganography. As exemplified in HUGO, WOW, S-UNIWARD, and MiPOD[6],[7–9], modern adaptive techniques tend to improve the optimization of a distortion function that governs the acceptable degradation of the cover. Embedding in the spatial domain often causes a greater loss in embedding efficiency. However, methods that rely on the distortions in the transform domain, such as the Discrete Cosine Transform (DCT)[10]or the Discrete Wavelet Transform (DWT)[11], can embed messages in less visually disturbing places. Perceptual masking also exploits the distortions of the Human Visual System (HVS). Recent advances in this field have moved the focus to the ability of Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) to model low-cost distortion functions that minimize the perceived risk of embedding. The most commonly used techniques include region-aware embedding, which groups embedding changes in less sensitive areas of the image [9], and frequency-adaptive embedding, which alters transform coefficients to embed information[12]. Overall, embedding methods emphasize greater adaptability and learning techniques when embedding volume and perceptibility are of concern.

1.3 Steganalysis Networks in Recent Years

Recent steganalysis networks have branched in three directions. First, early CNN detectors proved that steganographic residuals can be drawn from images. Second, the use of residual architectures made it possible to extract faint embedding artifacts by filtering semantic contents in images and preserving high frequencies. Third, the use of models that are lightweight, hybrid, and domain adaptive was aimed at increasing the robustness and the ease of implementation of the models after a distributional shift. A common design feature of many recent models is high-pass preprocessing, which often is either derived from SRM or SRM-like residual filters, and which enhances noise embedding in the images prior to the feature extraction process for the purpose of steganalysis[13]. Also on the increase hybrid and ensemble designs are used, wherein features of CNNs that are used to learn transferable residual representations are used in combination with discriminative classifiers or adaptation layers in order to increase the robustness of the modeling under mismatch of cover sources. The focus of this review is on image steganalysis; however, text, speech, and video steganalysis also cover the same broad concept[14,15] .

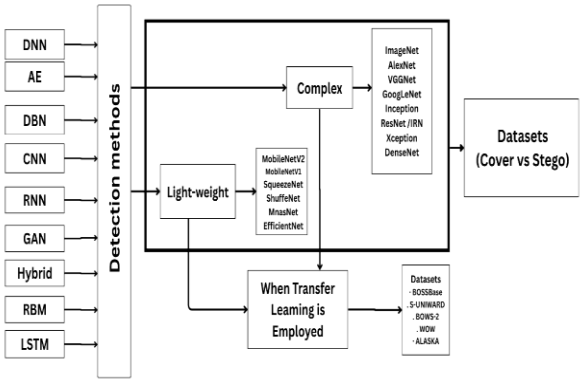


Figure 3. samples of deep learning steganalysis systems. The systems demonstrate high-pass preprocessing, learning residual features, transfer learning methodologies, as well as the classification stage.

In figure 3 Steganalysis deep learning pipelines presented for high-pass preprocessing, residual feature learning, transfer-learning variants, and classification stages[13,14,16].

1.4 GAN-Based Steganography and Detection

GAN-based steganography has become an important direction because it can model the interaction between embedding and detection as an adversarial optimization problem. Early adversarial approaches generated steganographic images through encoder-discriminator training [50], while later work learned pixel-wise change probabilities or embedding costs using GANs and reinforcement learning[17–19]. CNN-based adversarial embedding also demonstrated that covers can be modified to evade target steganalyzers[20]. These methods increase embedding security, but they also raise the computational burden and may generalize poorly when the target detector, dataset source, or deployment condition changes.

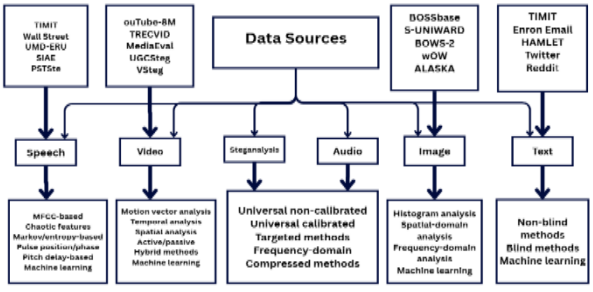


Figure 4. Steganalysis Scope.

Different input types as in figure 4 (image, text, audio and video) each create their own array of embedding artifacts, resulting in different feature representations.[2,21–23],Optimization-Oriented Payload Embedding Techniques.

1.5 Optimization-Oriented Embedding

Another area of growth in modern steganography is the embedding of optimized hiding techniques. Modern steganography hides information inside an image by modifying pixels in a non-uniform manner. These techniques implement a distortion function that calculates the optimal location to make the pixel change. These methods tend to prefer textured regions or regions that are difficult to model, while avoiding the smooth regions of the image, which are easily identifiable by the human eye[6–9,24]. These methods can increase the security of the embedding and the imperceptibility of the changes. However, this also increases the difficulty of the analysis because the embedding artifacts become dependent on the content. Optimization of the methods is especially important in large-scale systems because traditional methods are too slow. For these methods to be useful, the detectors must be able to identify the embedded information without being overly sensitive to a single image or embedding method.

1.6 Adversarial and Specialized Techniques

It is possible to fool CNN-based steganalysis models with the use of adversarial perturbations. However, defenses designed for a particular target detector tend to lack generalizability [5], [49]. These problems are evident in other media as well. Different forms of media require different techniques to detect embedding artifacts because they are not the same as those found in stationary images. Text, speech, and video all require their own methods of steganalysis [16]-[18]. This highlights the importance of robust steganalysis across all possible media.

1.7 AI-Based Steganalysis Methods

The recent progress in image steganalysis owes much to the development of deep neural network architectures, from static feature extraction towards adaptive residual driven learning as well as hierarchical and space-efficient multistage representation. Chapter 5 next divides the notable contributions by methodological paradigms.

2 Related Work

Image steganalysis research has developed across three phases: hand-crafted statistical modeling, deep convolutional detection, and more recently, transfer-learning, adversarial, and domain-adaptive models. This section locates the present work in that research and describes the gap in that research. An in-depth analysis of the architecture of the CNN detectors is reserved for Section 3.

2.1 Hand-Crafted Feature Steganalysis

Early steganalysis detectors were built by modeling the statistics of the embedding residuals. The Subtractive Pixel Adjacency Matrix (SPAM)[25] captured greater order Markov Pixel dependence. The Spatial Rich Model (SRM) extended this idea into a greater number of models and sub-models of residuals[26]. These, and other representations, were paired to ensemble classifiers [27] and selection-channel-aware variants utilizing embedding probability maps[28]. Hand-crafted detectors proved effective under matched conditions, but experienced degradation against content-adaptive embedding schemes, such as HUGO[7] , WOW[8], S-UNIWARD[6], and MiPOD[9], concentrating changes in regions that are most difficult to model or in areas of high texture. Their deterioration when there was a cover-source mismatch helped drive the shift toward learned residual representations.

2.2 Deep Convolutional Steganalysis

Convolutional steganalysis implements detection models with auto-prediction capabilities, as illustrated by Qian et al.[29] , Xu et al.[13], and subsequent models like Ye-Net[30], Yedroudj-Net[31], SRNet[15], and Zhu-Net[12] that leverage high-pass preprocessing, activation truncation, residue learning, and multi-scale pooling. More recent models, such as GBRAS-Net, achieve improved results on steganographic methods HILL and HUGO as well as on the BOSSBase and BOWS2 Corpora by combining filter banks inspired by SRM with depthwise separable convolution and skip connections. Very deep steganalysis networks may be further improved by utilizing shared and bias-free normalization[32].

2.3 Transfer Learning and JPEG-Domain Steganalysis

After the ALASKA challenges, emphasis shifted toward realistic, heterogeneous JPEG images. General-purpose vision backbones, including EfficientNet, MixNet, and ResNet, demonstrated that freezing and/or modifying the early layers of these networks to maintain input resolution and residual detail, could enable JPEG

steganalysis without significant alteration of these networks[33,34]. Transfer learning in this context may promote cost-effective training and improved performance on heterogeneous data.

2.4 GAN-Based and Adversarial Methods

Trends in generative adversarial training have been employed in both steganography and steganalysis. Adversarial play in the context of steganalysis allows ASDL-GAN to learn pixel-level change probabilities and embedding security within a more generalized adversarial embedding framework[17]. Cost of embedding security is the time and resources spent training the model, along with cross-model limitations and heavy training costs[18–20,35] discussed in Section 7.

2.5 Generalization Studies and Surveys

Cover-source mismatch is one of the barriers that obstruct the real-world deployment of image steganalysis. In general, most detectors trained on one source often transfer poorly to another[36]. Several surveys have highlighted the rapid progress and existing challenges of deep steganalysis[37,38], while some recent surveys expand the challenges to various data types and AI-generated data[2,39].

2.6 Position of This Review

Unlike the broad surveys that catalog methods, this review examines the transition from handcrafted steganalysis to deep learning-based image steganalysis. The main contribution of this review is a compact architectural taxonomy, a synthesis of performance analysis in the context of natural and heterogeneous image sources, and a discussion of the efficiency and robustness trade-offs for practical use.

3 Taxonomy of AI-Based Steganalysis Methods

Progress in image steganalysis has moved from static feature extraction toward adaptive, residual-driven learning and hierarchical multi-scale representation. This section organizes representative methods by their methodological paradigm rather than by chronology alone.

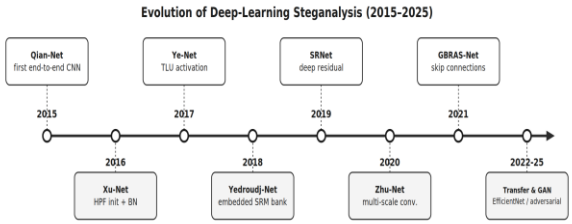


Figure 5. Timeline of the deep-learning shift in image steganalysis from early CNN detectors to residual, transfer-learning, and GAN-influenced systems.

Figure 5 places representative methods on a timeline, showing the progression from Qian-Net as an early end-to-end CNN to residual networks, transfer-learning approaches, and GAN-driven systems.

3.1 Static CNN-Based Steganalysis

Early CNN-based steganalysis showed that learned features could replace fully hand-designed pipelines. Qian et al. proposed one of the first end-to-end CNN detectors, using five convolutional layers, three fully connected layers, a Gaussian-shaped nonlinearity, mean pooling, and ReLU activations for multi-scale feature extraction[10] . Although the architecture was relatively simple, it established a foundation for later CNN-based detectors.

3.2 Embedding-Aware Residual Architectures

Xu et al. were the first to create a CNN structure to perform image steganalysis. The first neural network layer was configured to contain a kernel for high-pass filtering, to help emphasize small embedding changes. Some batch standardization was used to help reduce the fluctuations of the neural network, and the first layer was down-sampled to retain local evidence. Layer-by-layer normalization was used to improve the convergence of the structure. The model was designed to not perform down-sampling too early, which would cause the model to lose very very small signals that were probably steganographic in nature. The model became a reference model for CNN-based steganalysis designed to counter adaptive models, such as the WOW model, and the S-UNIWARD model[13,32].

3.3 Contemporary Multi-Scale Architectures

Later designs focused on multi-scale and residual feature learning. Yedroudj-Net reports competitive performance across datasets and embedding rates, with a filter bank that uses SRM, a truncated linear unit (TLU), and a stack of convolutional layers each followed by a batch normalization and scaling layer ([31,32][22], [24]). SRNet improved on deeper residual learning with the use

of global average pooling. Zhu-Net improved multi-scale representation and efficiency with depthwise separable convolutions, Inception-style modules, and spatial pyramid pooling ([14,40]). These designs point to a more or less constant trade-off of stronger representation, improved detection, and higher training and inference costs.

3.4 Deep Residual Steganalysis (2018–2025)

SRNet represents a major step in deep residual steganalysis. It uses cascaded 3 x 3 convolutions and residual learning blocks, followed by global average pooling to produce a fixed-length representation[14]. Compared with SRM+ensemble classifiers and earlier CNN detectors, SRNet improves detection of faint embedding signals in both fixed-pattern and adaptive schemes, including WOW and S-UNIWARD. Table 1 summarizes the broader architectural trajectory from shallow CNNs to deeper residual and multi-scale models.

Table 1. Analytical comparison of representative CNN-based steganalysis networks.

Model Design principle	High-pass strategy Nonlinearity	Classification/ Evaluation setting Design principle
Qian-Net [10] Initial noise reduction	Fixed 5×5 HPF Gaussian	Single FC + softmax; BOSSBase Initial noise reduction
Xu-Net [12] ABS-based residual enhancement	Fixed 5×5 HPF ABS	FC after pooling; BOSSBase ABS-based residual enhancement
Ye-Net [32] Truncation-aware learning	No explicit HPF TLU	FC; BOSSBase, BOWS2 Truncation-aware learning
Yedroudj-Net [31] SRM-inspired preprocessing	Fixed 5×5 HPF ABS + ReLU	FC after global pooling BOSSBase, BOWS2
SRNet [1] Deep residual learning	Multiple learned HPFs ReLU	Deep FC block BOSSBase, BOWS2, ALASKA
Zhu-Net [12] Multi-scale separable convolutions	5×5 & 3×1 filters ReLU	Dual FC BOSSBase, ALASKA

Table 1 shows how early CNN detectors such as Qian-Net and Xu-Net used a fixed high-pass filter to suppress content in the image and reveal embedding artifacts. However, due to their shallow structures, capacity and robustness were limited. Ye-Net and Yedroudj-Net, among others, attempted to change that by using learned nonlinearities and SRM-style preprocessing to develop a greater sensitivity to weak residuals. Further down the line, SRNet and Zhu-Net introduced multi-scale models to the residuals. While these models are accurate, the cost of computation and the interpretability of the models were affected. The trend, therefore, leans towards multi-scale models combined with residuals. Practical implementations, however, still demand efficiency and robustness.

4 AI-Driven Detection Models

This section summarizes the transition from classical machine learning to deep neural networks and explains how this transition changed steganalysis from hand-crafted feature construction to data-driven representation learning.

4.1 Evolution of Detection Techniques

Classical steganalysis relied on statistical descriptors, including co-occurrence matrices, residual histograms, SPAM, and[41] SRM feature sets[25,27,29]. These descriptors were commonly classified using SVMs or ensemble classifiers. Although such models can perform well in constrained settings, they lose accuracy against adaptive schemes such as WOW and S-UNIWARD and degrade in cross-dataset evaluations [10,42].
CNNs enabled end-to-end feature learning from raw pixels or residual images, reducing dependence on manually designed descriptors[31,43]. On standard benchmarks such as BOSSBase at 0.4 bpp, Yedroudj-Net and SRNet have reported accuracy above 90%, exceeding many hand-crafted baselines by a substantial margin[14,31]. GANs further intensified the steganography-steganalysis arms race by supporting more adaptive embedding strategies for hiders and richer training signals for detectors[38,39].

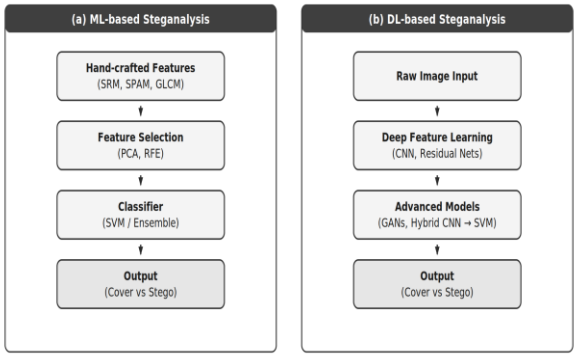


Figure 6. Comparison of two steganalysis paradigms: a hand-crafted feature pathway[26,30].

Figure 6 contrasts the two dominant pathways. The machine-learning route extracts hand-crafted features such as SRM, SPAM, and GLCM, applies feature selection through PCA or RFE, and classifies images using SVMs or ensemble models. The deep-learning route learns residual features from raw or preprocessed

images through CNNs or residual networks and then performs cover/stego classification using neural or hybrid classifier heads.

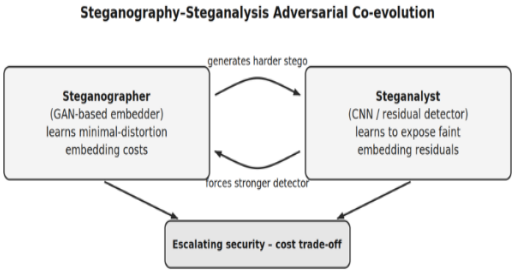


Figure 7. is about embedding methods and steganalysis methods.

With GANs becoming the most popular for steganography and removers using CNN and residual analysis, there is going to be an arms race, and each is going to be even more costly to produce and run.

For steganography, GANs are the primary tool. These algorithms are built to be extremely hard to detect. Steganalysis methods have to be able to detect extremely small residual traces. Each side has made significant advancements, and each side ups the difficulty for the other. It's even more costly to both sides to produce these methods.

4.2 Representative Model Families

CNN-based spatial-artifact detectors capture residuals and structured perturbations that survive embedding.
Residual connections, attention mechanisms, and frequency-aware components are often added to improve sensitivity to weak signals[14,16,41].
Source-aware and robust-training methods aim to reduce dataset-specific bias and improve domain invariance so that performance degrades more gradually under moderate distribution shifts[3,36].
Recurrent and spatio-temporal hybrids combine CNNs with RNN or LSTM modules to capture spatial and temporal dependencies in dynamic media such as video, speech, or sequential communication streams[2,21–23].

5 Benchmark Datasets for Evaluation

This section reviews the datasets commonly used to evaluate steganalysis models and explains their importance for fair and reproducible comparison.

5.1 Commonly Used Datasets

Steganalysis progress and challenges rely on realistic and solid benchmarks. BOSSBase 1.01 provides a collection of uncompressed and un- grayscaled images and facilitates tailored testing. BOSSBase 1.01 is typically used with diverse datasets such as COCO to evaluate existing approaches of steganography[44]. Increasingly, datasets of synthetic or generative images are used alongside datasets of real photographs. Natural photographs provide datasets to help identify selection bias present in benchmarks with a limited set of photographs[6,42]. A combination of datasets of real and synthetic images help expand sample variety and facilitate cross-domain diagnostics and expose behavior of models. As embedding methods become flexible and more adaptive, the diversity of the dataset becomes necessary for the adequate measurement of evolving methods[11,44,45].

5.2 Relevance to AI-Driven Models

AI-driven models have improved detection capability by learning representations that are difficult to capture with predefined descriptors. Traditional detectors based on hand-engineered features remain useful as baselines, but they are limited by the assumptions built into their feature sets[26,27,46]. CNNs, by contrast, can learn discriminative residual patterns directly from image data [41,44]. This shift has moved the field from manual feature construction toward training-protocol design, residual preprocessing, large-scale data handling, and efficient adaptation to new steganographic methods[2,38]. Deep models can also be retrained or fine-tuned when new embedding techniques appear, which helps maintain performance against adaptive attacks.

5.3 Evaluation Protocols Reported in the Literature

Artificial intelligence (AI)-based image steganalysis is typically assessed using specific, regulated benchmark protocols, particularly BOSSBase 1.01, BOWS2, and related datasets. These benchmarks create conditions for controlled payloads, in which the image source and embedding algorithm are known[31,38,47] [24]. Since most studies cast the problem in terms of binary cover/stego discrimination, results are usually presented for testing scenarios that involve embedding algorithms in the spatial and JPEG domains. These algorithms include WOW, S-UNIWARD, HUGO, MiPOD, HILL, and J-UNIWARD [2,31,38,48]. Unlike traditional machine learning, which follows a single standard partitioning of the dataset, studies in this area generally

specify their own training, validation, and test splits for training the discriminators, controlling overfitting, and selecting the model, and then cite the performance for the splits [31,38]. The performance metrics typically described include detection error and detection accuracy. However, in many assessment studies, additional metrics have included precision, recall, F1 score, area under the curve (AUC), receiver operating characteristic (ROC) curves, and confusion matrices, which evaluate the reliability of the classification and the separability of cover and stego images[37,49]. There are a number of conditions disclosed in the literature that can adversely impact the performance of deep steganalyzers, which include a source-cover mismatch, varying image acquisition setups, and several types of post-processing and noise[10,34,36,50,51]. Because of this, there is a strong focus on two key areas. The first is measuring the bias of the source and distributions of data through cross-dataset and cross-source evaluations. The second is designing robust, deployable detectors using transfer learning, data augmentation, domain adaptation, active learning, ensemble methods, and modular, streamlined designs[10,36,50,52].

5.4 CNN-SVM Hybrid Models

Combining convolutional neural networks (CNNs) with support vector machine (SVM) classifiers utilizes the representation ability of CNNs along with the margin-based decision boundary of SVM classifiers. CNN layers model local structures and SVM heads improve cover/stego separation, even when embedding conditions or payload levels change[3,38]. Hybrid models can also be used on audio, text, and video, like CNN-RNN or CNN-LSTM models, because of the presence of temporal dependencies. The combination of local features and temporal dependencies makes hybrid models useful for multimodal steganalysis. However, hybrid models also require dataset design and validation for the best results[2,3,7].

6 Generalization and Computational Constraints

6.1 Cross-Dataset Generalization

A central limitation in steganalysis is cover-source mismatch. Models trained on homogeneous data may learn source-specific statistics and then transfer poorly to heterogeneous sources. As summarized in Table 2, both classical and deep models can overfit when training data are not sufficiently diverse. Larger and more varied datasets, together with domain-adaptation techniques, could transfer more useful inductive bias across domains and produce more realistic benchmarks[2,3,53,54] [1].

Table 2. Reported detection accuracy on natural and heterogeneous data in spatial and JPEG-domain steganalysis [1].

Experiment	Domain	Embedding algorithm (payload)	Dataset composition	Accuracy (%)
Exp-1	Spatial	HUGO (0.4 bpp)	BOSSBase v1.01 (10,000 natural images)	90.64
Exp-2	JPEG	J-UNIWARD (0.4 bpp)	BOSSBase v1.01 (10,000 natural images)	91.75
Exp-3	Spatial	HUGO (0.4 bpp)	Heterogeneous: 5,000 natural + 5,000 AI-generated	85.76
Exp-4	JPEG	J-UNIWARD (0.4 bpp)	Heterogeneous: 5,000 natural + 5,000 AI-generated	88.30

In the underlying study [44], features were compared across spatial and JPEG-domain settings using original and mixed datasets. The experiment and domain columns identify the evaluation setting; the embedding algorithm and payload column specifies the hiding method; the dataset column distinguishes BOSSBase 1.01 natural photographs from a heterogeneous mixture of natural and AI-generated images; and the final column reports detection accuracy.

Table 3. Summary of the generalization gap between natural-only and heterogeneous datasets.

hidden message type	natural photos only	mixed (natural + ai)
HUGO (0.4 bpp)	90.64%	85.76%
J-UNIWARD (0.4 bpp)	91.75%	88.30%

Table 3 summarizes the same trend more compactly. Detection accuracy is above 90% on natural photographs, reaching 90.64% for HUGO and 91.75% for J-UNIWARD. When the same setting is evaluated on a mixed dataset of natural and AI-generated images, accuracy decreases to 85.76% for HUGO and 88.30% for J-UNIWARD. This 3-5% reduction illustrates the generalization gap introduced by source heterogeneity.

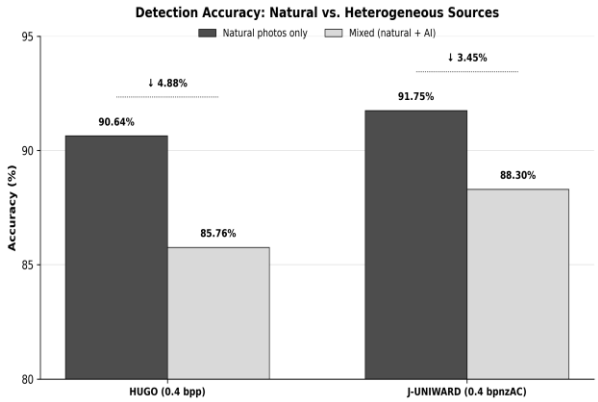


Figure 10. The data shows the HUGO and J-UNIWARD detection accuracy generalization gap (3-5%) discussed in Table 3 with natural-only and datasets combining natural and AI-generated data.

6.2 Computational Constraints

Deep steganalysis handles most tasks with high accuracy, but performing these tasks comes at a cost. Multi-layer networks, such as modern 2D CNNs, rely on large, multi-epoch datasets and expensive hardware. This reliance only grows when hybrid or multi-stage networks are developed, as the time taken to perform an inference may not even be feasible in real-time situations. There are several strategies to achieve a favorable mean of both accuracy and efficiency:

- Model compression reduces the number of parameters and loss in accuracy by removing unnecessary parameters and reducing precision.

- Knowledge distillation uses a larger, more accurate network as a “teacher” to guide a more compact network, or “student,” to approximate the same decisions.

- Mixed-precision inference modifies the precision in some calculations to improve the speed of an inference with an acceptable loss of accuracy.

- Early-exit mechanisms integrate intermediate classifiers to allow samples to exit early if the prediction confidence is adequate to avoid unnecessary computation.

Large variances in the datasets is still a major problem, even with the proposed solutions. The suggested solutions are matching dataset partitions, lightweight modification layers, selective fine-tuning of model blocks tailored to the deployment, and transfer learning from large, efficient models such as EfficientNet, which has performed well with heterogeneous JPEG sources. The end goal is to innovate on robust, efficient steganalysis models that are resource conscious while maintaining accurate detection across heterogeneous real world sources.

7 Conclusion and Future Directions

The focus of this review was to analyze the transition of image steganalysis from hand engineered features to steganalysis methods based on deep learning. The recent techniques have been categorized into the families of CNN based methods, residual based methods, hybrid methods, and domain adaptive methods. Additionally, we have assessed a number of architectures concerning their image processing methods, nonlinearity, classifier design and the benchmarks they evaluated their methods on.

Summary of results. Among the various architectures, deep residual networks and in particular the architectures based on the Structural Residual Networks (SRNet) have proven to be the most successful in a number of image steganalysis benchmarks and have shown to outperform the traditional methods based on SRM and ensemble techniques. Hybrid methods combining Convolutional Neural Networks (CNN) with Support Vector Machines (SVM) and Long Short Term Memory (LSTM) networks have also been shown to be successful in steganalysis of images with high payloads and multiple modalities. The success of these methods is highly dependent on the validation of the methods and the quality of the dataset.

Challenges. The quality of the chosen cover images remains an issue that directly degrade performance. Furthermore, the models' lack of adversarial robustness and high computational costs are also major concerns.

Future directions. Among the domains of primary importance will be the development of adaptive and invariant source models, and the design of lightweight frameworks based on compression and knowledge distillation methods, and new benchmarks combining images stemming from both the natural and artificial domains. The successful completion of these goals will close the gap between the performance of methods in the laboratory and their usage in the field, and will provide more robust, interpretable and practical solutions to the growing problem of advanced steganography.

8 REFERENCES

[1] Huang S, Zhang M, Kong Y, Ke Y, Di F. FACSNet: Forensics aided content selection network for heterogeneous image steganalysis. Sci Rep 2024;14. https://doi.org/10.1038/s41598-024-77552-x.

[2] Kheddar H, Hemis M, Himeur Y, Megías D, Amira A. Deep learning for steganalysis of diverse data types: A review of methods, taxonomy, challenges and future directions. Neurocomputing 2024;581. https://doi.org/10.1016/j.neucom.2024.127528.

[3] Sanjalawe Y, Al-E'mari S, Fraihat S, Abualhaj M, Alzubi E. A deep learning-driven multi-layered steganographic approach for enhanced data security. Sci Rep 2025;15. https://doi.org/10.1038/s41598-025-89189-5.

[4] Wang J, Cheng M, Wu P, Chen B. A Survey on Digital Image Steganography. Journal of Information Hiding and Privacy Protection 2019;1:87–93. https://doi.org/10.32604/jihpp.2019.07189.

[5] Kumar A, Singla P, Yadav A. StegaVision: Enhancing Steganography with Attention Mechanism 2024.

[6] Holub V, Fridrich J, Denemark T. Universal distortion function for steganography in an arbitrary domain. 2014.

[7] Pevný T, Filler T, Bas P. Using High-Dimensional Image Models to Perform Highly Undetectable Steganography. In: Böhme R, Fong PWL, Safavi-Naini R, editors. Information Hiding, Berlin, Heidelberg: Springer Berlin Heidelberg; 2010, p. 161–77.

[8] Holub V, Fridrich J. DESIGNING STEGANOGRAPHIC DISTORTION USING DIRECTIONAL FILTERS. n.d.

[9] Sedighi V, Cogranne R, Fridrich J. Content-adaptive steganography by minimizing statistical detectability. IEEE Transactions on Information Forensics and Security 2016;11:221–34. https://doi.org/10.1109/TIFS.2015.2486744.

[10] Alrusaini OA. Deep learning for steganalysis: evaluating model robustness against image transformations. Front Artif Intell 2025;8. https://doi.org/10.3389/frai.2025.1532895.

[11] You H, Xiang L, Yang C, Shen X. Linguistic steganalysis via multi-task with crossing generative-natural domain. Neurocomputing 2024;603:128260. https://doi.org/https://doi.org/10.1016/j.neucom.2024.128260.

[12] Zhang R, Zhu F, Liu J, Liu G. Efficient feature learning and multi-size image steganalysis based on CNN 2018.

[13] Xu G, Wu HZ, Shi YQ. Structural design of convolutional neural networks for steganalysis. IEEE Signal Process Lett 2016;23:708–12. https://doi.org/10.1109/LSP.2016.2548421.

[14] Boroumand M, Chen M, Fridrich J. Deep residual network for steganalysis of digital images. IEEE Transactions on Information Forensics and Security 2019;14:1181–93. https://doi.org/10.1109/TIFS.2018.2871749.

[15] Subramanian N, Elharrouss O, Al-Maadeed S, Bouridane A. Image Steganography: A Review of the Recent Advances. IEEE Access 2021;9:23409–23. https://doi.org/10.1109/ACCESS.2021.3053998.

[16] Review of Deep Residual Learning for Image Recognition n.d.

[17] Tang W, Tan S, Li B, Huang J. Automatic Steganographic Distortion Learning Using a Generative Adversarial Network. IEEE Signal Process Lett 2017;24:1547–51. https://doi.org/10.1109/LSP.2017.2745572.

[18] Zhang KA, Cuesta-Infante A, Xu L, Veeramachaneni K. SteganoGAN: High Capacity Image Steganography with GANs 2019.

[19] Tang W, Li B, Barni M, Li J, Huang J. An Automatic Cost Learning Framework for Image Steganography Using Deep Reinforcement Learning. IEEE Transactions on Information Forensics and Security 2021;16:952–67. https://doi.org/10.1109/TIFS.2020.3025438.

[20] Tang W, Li B, Tan S, Barni M, Huang J. CNN-Based Adversarial Embedding for Image Steganography. IEEE Transactions on Information Forensics and Security 2019;14:2074–87. https://doi.org/10.1109/TIFS.2019.2891237.

[21] Bouzegza M, Belatreche A, Bouridane A, Tounsi M. A comprehensive review of video steganalysis. IET Image Process 2022;16:3407–25. https://doi.org/10.1049/ipr2.12573.

[22] Yang Y, Zha L, Zhang Z, Wen J. An Overview of Text Steganalysis. Lecture Notes in Electrical Engineering, vol. 813, Springer Science and Business Media Deutschland GmbH; 2022, p. 933–43. https://doi.org/10.1007/978-981-16-6963-7_82.

[23] Wu Z, Guo J, Zhang C, Li C. Steganography and steganalysis in voice over ip: A review. Sensors (Switzerland) 2021;21:1–31. https://doi.org/10.3390/s21041032.

[24] Kuznetsov O, Frontoni E, Chernov K, Kuznetsova K, Shevchuk R, Karpinski M. Enhancing Steganography Detection with AI: Fine-Tuning a Deep Residual Network for Spread Spectrum Image Steganography. Sensors 2024;24. https://doi.org/10.3390/s24237815.

[25] Multimedia A, Workshop S. Steganalysis by Subtractive Pixel Adjacency Matrix. 2009.

[26] Fridrich J, Kodovský J. Rich Models for Steganalysis of Digital Images. 2012.

[27] Kodovský J, Fridrich J, Holub V. Ensemble Classifiers for Steganalysis of Digital Media. 2010.

[28] Denemark T, Sedighi V, Holub V, Cogranne R, Fridrich J. Selection-channel-aware rich model for Steganalysis of digital images. 2014 IEEE International Workshop on Information Forensics and Security, WIFS 2014, Institute of Electrical and Electronics Engineers Inc.; 2014, p. 48–53. https://doi.org/10.1109/WIFS.2014.7084302.

[29] Qian Y, Dong J, Wang W, Tan T. Deep learning for steganalysis via convolutional neural networks. Media Watermarking, Security, and Forensics 2015, vol. 9409, SPIE; 2015, p. 94090J. https://doi.org/10.1117/12.2083479.

[30] Zeng J, Tan S, Li B, Huang J. Large-scale JPEG steganalysis using hybrid deep-learning framework 2017. https://doi.org/10.1109/TIFS.2017.2779446.

[31] Yedroudj M, Comby F, Chaumont M. YEDROUDJ-NET: AN EFFICIENT CNN FOR SPATIAL STEGANALYSIS. 2018.

[32] Wu S, Zhong S, Liu Y. A Novel Convolutional Neural Network for Image Steganalysis with Shared Normalization 2017.

[33] Tan M, Le Q V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. n.d.

[34] Yousfi Y, Butora J, Khvedchenya E, Ai O, Fridrich J. ImageNet Pre-trained CNNs for JPEG Steganalysis. n.d.

[35] Hayes J, Danezis G. Generating steganographic images via adversarial training. n.d.

[36] Mallet A, Beneš M, Cogranne R. Cover-source mismatch in steganalysis: systematic review. EURASIP J Inf Secur 2024;2024. https://doi.org/10.1186/s13635-024-00171-6.

[37] Tabares-Soto R, Raúl RP, Gustavo I. Deep learning applied to steganalysis of digital images: A systematic review. IEEE Access 2019;7:68970–90. https://doi.org/10.1109/ACCESS.2019.2918086.

[38] Croix NJD La, Ahmad T, Han F. Comprehensive survey on image steganalysis using deep learning. Array 2024;22. https://doi.org/10.1016/j.array.2024.100353.

[39] Setiadi DRIM, Ghosal SK, Sahu AK. AI-Powered Steganography: Advances in Image, Linguistic, and 3D Mesh Data Hiding – A Survey. Journal of Future Artificial Intelligence and Technologies 2025;2:1–23. https://doi.org/10.62411/faith.3048-3719-76.

[40] Zhang R, Zhu F, Liu J, Liu G. Efficient feature learning and multi-size image steganalysis based on CNN 2018.

[41] Jin Z, Yang Y, Chen Y, Chen Y. IAS-CNN: Image adaptive steganalysis via convolutional neural network combined with selection channel. Int J Distrib Sens Netw 2020;16:1550147720911002. https://doi.org/10.1177/1550147720911002.

[42] Wang Z, Chen M, Yang Y, Lei M, Dong Z. Joint multi-domain feature learning for image steganalysis based on CNN. EURASIP J Image Video Process 2020;2020. https://doi.org/10.1186/s13640-020-00513-7.

[43] Wu S, Zhong S, Liu Y. A Novel Convolutional Neural Network for Image Steganalysis with Shared Normalization 2017.

[44] Huang S, Zhang M, Kong Y, Ke Y, Di F. FACSNet: Forensics aided content selection network for heterogeneous image steganalysis OPEN n.d. https://doi.org/10.1038/s41598-024-77552-x.

[45] Reinel TS, Brayan AAH, Alejandro BOM, Alejandro MR, Daniel AG, Alejandro AGJ, Buenaventura BJA, Simon OA, Gustavo I, Raul RP. GBRAS-Net: A Convolutional Neural Network Architecture for Spatial Image Steganalysis. IEEE Access 2021;9:14340–50. https://doi.org/10.1109/ACCESS.2021.3052494.

[46] Fridrich J, Kodovský J. Rich Models for Steganalysis of Digital Images. 2012.

[47] Bas P, Filler T, Pevny T, Break TP", Steganographic O. “Break Our Steganographic System”: The Ins and Outs of Organizing BOSS 2011:10. https://doi.org/10.1007/978-3.

[48] Xu G. Deep convolutional neural network to detect J-UNIWARD. IH and MMSec 2017 - Proceedings of the 2017 ACM Workshop on Information Hiding and Multimedia Security, Association for Computing Machinery, Inc; 2017, p. 67–73. https://doi.org/10.1145/3082031.3083236.

[49] Alrusaini OA. Deep learning for steganalysis: evaluating model robustness against image transformations. Front Artif Intell 2025;8. https://doi.org/10.3389/frai.2025.1532895.

[50] Abecidan R, Itier V, Boulanger J, Bas P, Pevný T. Leveraging Data Geometry to Mitigate CSM in Steganalysis 2023.

[51] Lerch-Hostalot D, Megías D. Detection of classifier inconsistencies in image steganalysis. IH and MMSec 2019 - Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, Association for Computing Machinery, Inc; 2019, p. 222–9. https://doi.org/10.1145/3335203.3335738.

[52] Yousfi Y, Butora J, Khvedchenya E, Ai O, Fridrich J. ImageNet Pre-trained CNNs for JPEG Steganalysis. n.d.

[53] Huang S, Zhang M, Kong Y, Ke Y, Di F. FACSNet: Forensics aided content selection network for heterogeneous image steganalysis OPEN n.d. https://doi.org/10.1038/s41598-024-77552-x.

[54] Mallet A, Beneš M, Cogranne R. Cover-source mismatch in steganalysis: systematic review. EURASIP J Inf Secur 2024;2024. https://doi.org/10.1186/s13635-024-00171-6.