Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Deep Learning-Based Image Manipulation Detection for Cybersecurity: A Comprehensive Survey and Future Directions

Kawthar Hamza[1*], Zainalabideen Rasheed[2]

[1]College of Education, University of Kufa, Najaf, 54001, Iraq. E-mail: kawthara.alsaabry@student.uokufa.edu.iq
[2]College of Education, University of Kufa, Najaf, 54001, Iraq. E-mail: zainalabideena.alsaraf@uokufa.edu.iq

Article's Information

Received: 14.09.2026
Accepted: 17.10.2026
Published: 23.11.2026

Keywords:
Deep Learning
Image Manipulation Detection
Digital Image Forensics
Deepfake Detection
Image Steganalysis
Document Forgery Detection
Cybersecurity

Abstract

Fast development of deep learning and Generative AI has resulted in the emergence of highly convincing manipulated images, thus creating threats to cybersecurity and integrity of digital media. Image manipulation includes deepfake synthesis, steganographic manipulation, and document forgery and requires accurate methods of forensic identification.

The present paper reviews image manipulation detection methods using deep learning within cybersecurity across three domains: deepfake image manipulation detection, image steganographic manipulation detection, and digital document forgery detection. Applying a systematic review approach, we selected and analyzed 49 articles and preprints published between 2014 and 2026 using a unified framework. Methods are grouped in a taxonomy involving convolutional neural networks (CNNs), transformers, attention mechanisms, and hybrids. The methods are compared by the benchmark datasets used, evaluation metrics and experimental setup, performance, computational cost, and limitations. Cross-domain synthesis reveals some common issues: discrepancy between the within-dataset and cross-dataset performance as well as lack of standardized metrics and cost measurement.

Open problems, including limitations on generalization, vulnerability to adversarial attacks and realistic transformations, high computational cost, and others are discussed. The directions of future research include multi-modal learning, lightweight architectures and adaptation methods, self-supervision, and foundation models.

https://doi.org/10.46649/fjiece.v5.2.4a.23.11.2026

*Corresponding author: kawthara.alsaabry@student.uokufa.edu.iq

1. INTRODUCTION

Recent advancements in visual media technology have enabled the creation of increasingly lifelike images, accompanied by the emergence of new tools for image manipulation [1]–[3]. Developments in deep learning and generative artificial intelligence (Generative AI) have further hastened the process of image creation, complicating the task of image verification [4], [5].

Image manipulation encompasses a broad spectrum of activities, such as the creation of deepfakes, image steganography, and document falsification [1]–[3]. These manipulative practices pose risks to cybersecurity, digital forensics, identity verification, journalism, and legal processes [6]–[8]. As manipulated images become more convincing, traditional detection methods have proven inadequate, prompting a transition to deep learning techniques [9], [10]. Currently, deep learning is considered the most effective strategy for image verification, as it extracts distinguishing features directly from the data [9], [11]. Techniques such as convolutional neural networks (CNNs), transformers, attention models, and hybrid architectures have demonstrated significant success in detecting deepfakes, identifying document forgery, and performing steganalysis [12]–[16].

ATU-FJIECE, Volume: 5, Issue: 2 , 11,23 , 2026, © 2020 FJIECE, All Rights Reserved

Despite advancements, several issues persist, such as limited generalization across datasets, susceptibility to adversarial attacks, subpar performance on compressed images, significant computational demands, and challenges in deploying in real-world scenarios [17]–[19]. This survey explores the latest progress in detecting image manipulation through deep learning for cybersecurity, focusing on three areas: deepfake detection, image steganalysis, and digital document forgery detection. The studies reviewed are categorized based on a taxonomy of deep learning techniques, facilitating a comparison of their benchmark datasets, metrics, methods, performance, and limitations. Additionally, the survey highlights the primary shortcomings of the current state of the art and suggests potential avenues for future research. Earlier surveys have covered parts of this field. Tolosana et al. [1] reviewed face manipulation and fake detection, Verdoliva [20] gave an overview of media forensics with a focus on deepfakes, Zanardelli et al. [21] reviewed deep learning approaches to copy-move, splicing, and deepfake forgery detection, De La Croix et al. [2] surveyed deep learning-based image steganalysis, and Hassan and Fadlalla [3] surveyed techniques for detecting identity-document forgery. Table 1 summarizes these surveys. None of them covers image steganalysis, deepfake detection, and document forgery detection together, and none compares the three domains within a single framework. The present survey differs from them in three respects: (i) the 49 studies from the three domains are examined with the same set of attributes, namely dataset, architecture, metric, reported performance, and limitations; (ii) the benchmark datasets of the three domains are compared in a single table; and (iii) Section 6.2 shows that cover-source mismatch in steganalysis, failure on unseen generators in deepfake detection, and failure on unseen templates in document forgery detection are different forms of the same generalization problem.

Table 1. Comparison of This Survey with Existing Surveys

Survey	Year	Deepfake detection	Image steganalysis	Document forgery detection	Main focus
Hassan & Fadlalla [3]	2017	No	No	Yes	Techniques for detecting forged identity documents
Tolosana et al. [1]	2020	Yes	No	No	Face manipulation types, databases, and detection methods
Verdoliva [20]	2020	Yes	No	No	Image and video forensics with an emphasis on deepfakes
Zanardelli et al. [21]	2023	Partly	No	No	Deep learning detection of copy-move, splicing, and deepfake forgeries
De La Croix et al. [2]	2024	No	Yes	No	Deep learning-based image steganalysis
This survey	2026	Yes	Yes	Yes	49 studies (2014–2026) from the three domains analyzed with a common framework, with a cross-domain synthesis

2. SURVEY METHODOLOGY

The SLR approach is used. The protocol follows Kitchenham and Charters [22] and Snyder [23] with search and selection process reported according to PRISMA 2020 [24]. The corpus of this review contains journal papers, conference papers, and two preprints published between 2014 and 2026. The key period of consideration is 2018-2026, when deep learning becomes the prevailing solution within all three domains. Earlier works (before 2018) were included if (1) they presented the architecture or training strategy which has been used by other works, and (2) they reported the results on the publicly available benchmark.

Seven papers satisfy both conditions [25], [26], [8], [9], [27], [28], [29]. All of them are image steganalysis studies, because deep learning was applied to steganalysis from 2014 onwards, several years before deep learning-based deepfake and document forgery detection appeared. They are listed in Table 3(a) to trace the evolution from early CNN steganalyzers to residual, attention-based, and transformer-based detectors.

The literature search was carried out in four databases: IEEE Xplore, ScienceDirect, SpringerLink, and Google Scholar. The keywords "image manipulation detection," "deepfake detection," "image steganalysis," "document forgery detection," and "digital image forensics" were each combined with "deep learning" using the Boolean operator AND, giving the following general query: ("image manipulation detection" OR "deepfake detection" OR "image steganalysis" OR "document forgery detection" OR "digital image forensics") AND ("deep learning"). In IEEE Xplore, ScienceDirect, and SpringerLink, the query was applied to the title, abstract, and keyword fields. Because Google Scholar does not support field-restricted search, it was used as a complementary source to find relevant papers not indexed in the other three databases. The search was last updated in September 2026. Explicit inclusion and exclusion criteria were then applied. A paper was included if it (i) proposed a deep learning-based approach to image or video manipulation detection or experimentally evaluated an existing one; (ii) addressed visual image or video manipulation (papers dealing only with audio or non-visual multimodal content were excluded); (iii) reported at least one quantitative evaluation metric, such as Accuracy, Precision, Recall, F1-score, Area Under the Curve (AUC), average precision (AP or PR-AUC), equal error rate (EER), detection error (P_e), identification rate, or the Dice coefficient for localization, or, for quantitative steganalysis, the error of the estimated payload (MAE or MSE); and (iv) was published in a recognized peer-reviewed journal or conference proceeding [22]–[24], with the two exceptions described next. Two included studies, Zhu-Net [30] and FLiD [31], were available only as preprints at the time of the final search; they were kept because they report complete experiments on public benchmarks, and they are marked with † in Tables 3(a) and 3(c). One included steganalysis study, USteg-DSE, estimates the embedded payload instead of performing binary detection; it is retained as the only quantitative-steganalysis method and is reported with its own error metrics in Table 3(a). Dataset, benchmark, and challenge papers [32]–[45] are cited to describe the benchmarks in Table 2 and are not counted among the reviewed studies, except FaceForensics++ [34] and DeeperForensics-1.0 [46], which also propose and evaluate detection methods. Surveys, review papers, editorials, tutorials, duplicate publications, papers without experiments, papers limited to forgery generation, and papers unrelated to image manipulation detection were excluded from the comparative analysis.

Figure 1 describes the selection process of the literature review according to PRISMA 2020 [24]. After removing duplicates and clearly irrelevant articles by their titles and abstracts, 120 potentially relevant papers are left for full-text screening. Out of 120, 62 papers do not correspond to the criteria described above, and only 58 papers are retained for further analysis. A second, stricter eligibility check was then applied to the 58 retained studies: each study had to propose or evaluate a deep learning-based detector and report quantitative results that could be placed in the common analytical framework. This check removed 16 studies: seven dataset, benchmark, or challenge reports; three methods without a deep learning-based detector; two studies whose contribution concerned training optimization or preprocessing rather than the detector itself, one of which overlaps with a retained study by the same authors; one study on forgery generation only; one robustness evaluation of existing models, which is used as supporting evidence in Section 7; one study without quantitative results usable for comparison; and one method based on physiological signals extracted from video, which lies outside the visual-artifact scope of this review. The search update of September 2026 added seven document forgery detection studies that met the same criteria. The final set therefore contains 49 studies: 17 on image steganalysis, 18 on deepfake detection, and 14 on digital document forgery detection; of the latter, 12 are document-specific detectors and two (TruFor and a lightweight ViT) are general image-forgery detectors included as reference methods and marked with § in

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Table 3(c). The number of records initially returned by each database, the number of duplicates, the number of records removed at title and abstract screening, and the individual reasons for the 62 exclusions in the first full-text assessment were not recorded separately; this is a limitation of the present review. To allow the search to be repeated, the databases, query, search fields, and inclusion and exclusion criteria are reported in full above. The selected papers are analyzed according to their topic (manipulation detection), methodological contribution, experimental evaluation, and importance for this field. Papers are divided into three categories (manipulation domains) and further subdivided by the type of the architecture into convolutional neural networks (CNNs), transformers, attention mechanism-based and hybrid methods. Analytical framework is designed to unify the comparison across all papers, including the domain of the image manipulation, benchmark datasets, architecture characteristics, evaluation metric and setup, performance, computational cost, limitations, and the key results and innovations. This framework will help in detecting the common issues, tendencies, and prospective research areas in deep learning-based image manipulation detection within cybersecurity.

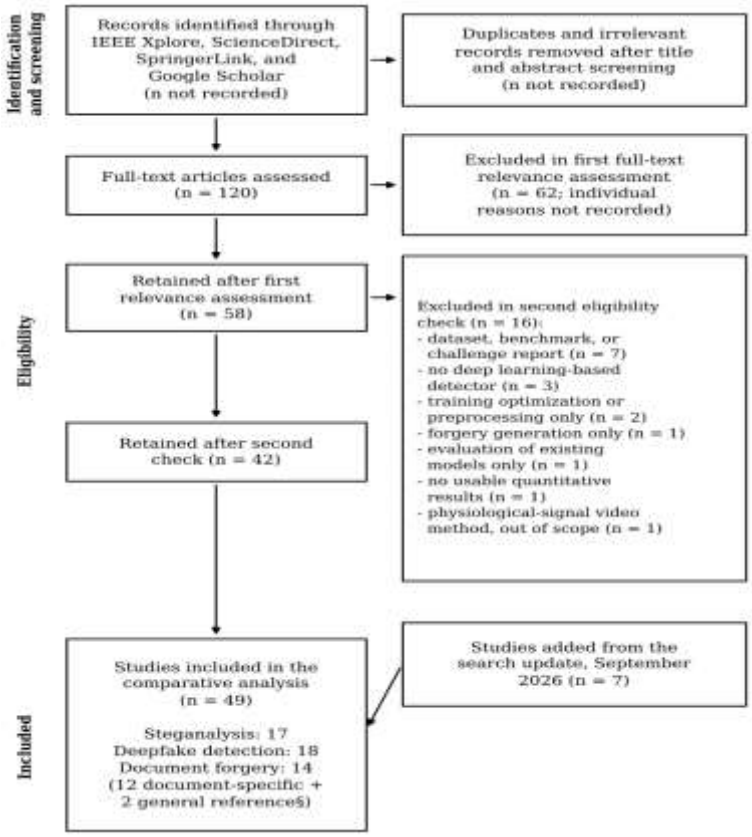

Fig. 1. PRISMA 2020 flow diagram of the study selection process.

Scope with respect to video. This survey focuses on still images, but deepfake detection cannot be fully separated from video, because the main deepfake benchmarks (FaceForensics++, Celeb-DF, and DFDC) are released as videos and most detectors are trained and tested on frames extracted from them. The inclusion criteria therefore accept both image and video manipulation detection, and the reviewed deepfake methods fall into two groups. Fourteen methods operate on single images or individual frames and can be applied directly to still images. Four methods, namely CNN+LSTM [47], LipForensics [48], STIL [49], and RealForensics [50], require a sequence of frames because they model temporal or audio-visual cues. They were retained because they are widely used reference points for cross-dataset generalization in deepfake

detection, although they cannot be applied to a single image. They are marked with ‡ in Table 3(b), and their results are discussed separately from those of the frame-level methods in Section 6.2.

3. TAXONOMY OF IMAGE MANIPULATION DETECTION

Image manipulation detection has become a multidisciplinary research area because image generation and manipulation techniques have developed and diversified rapidly. To give structure to this area, this survey classifies image manipulation into three major categories: deepfake image manipulation, image steganographic manipulation, and digital document forgery. These categories cover the main lines of research on deep learning-based image manipulation detection. The proposed taxonomy is shown in Fig. 2, which illustrates the relationships among the three domains.

3.1. DEEPFAKE IMAGE MANIPULATION

Deepfake image manipulation is defined as generation or modification of images with use of deep learning-based algorithms: GANs, autoencoders, and diffusion models [1], [4], [5]. They create highly plausible content including face-swap, facial attribute modification, face reenactment, and fully synthesized identity [1], which makes the modern deepfake detection challenging. The vast majority of detectors search for inconsistencies in the spatial domain, the frequency domain, biological signals, and textures [1], [6], [13]. Modern detectors use transformers and attention mechanisms to extract the dependencies which are long-range in order to achieve better generalization and robustness [13].

3.2. IMAGE STEGANOGRAPHIC MANIPULATION

Image steganography involves hiding the information inside an image while preserving its visible part and making the hidden part invisible [2]. Contemporary adaptive steganography raises many issues to the conventional steganalysis methods which use handcrafted features [2], [8]. That is why the contemporary deep learning-based steganalysis learns features which correspond to residual patterns and embedding artifacts [8]–[10]. Various CNNs, attention-based and hybrid architectures demonstrate remarkable results in identifying the spatial domain and transform domain steganographic manipulation with different conditions of embedding the information [10]–[12].

3.3. DIGITAL DOCUMENT FORGERY DETECTION

Document forgery includes malicious modification of identity or official records with the use of image processing tools or deep generative models [3], [7]. It may include modifications of texts, faces, numbers, signatures, and creation of fully synthetic documents. Detection of forgeries requires analysis of inconsistencies in the content and layout of a document. The latest approaches involve integration of deep convolutional networks with transformer-based architectures for improving localization, verification and classification of the documents [14].

This taxonomy provides unifying perspective on domains where image manipulation detection is researched. It will be used for comparison of similarity and difference of various deep learning solutions used in different manipulation domains and forms a basis for detailed analysis in this paper.

 © 2020 FJIECE, All Rights Reserved

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

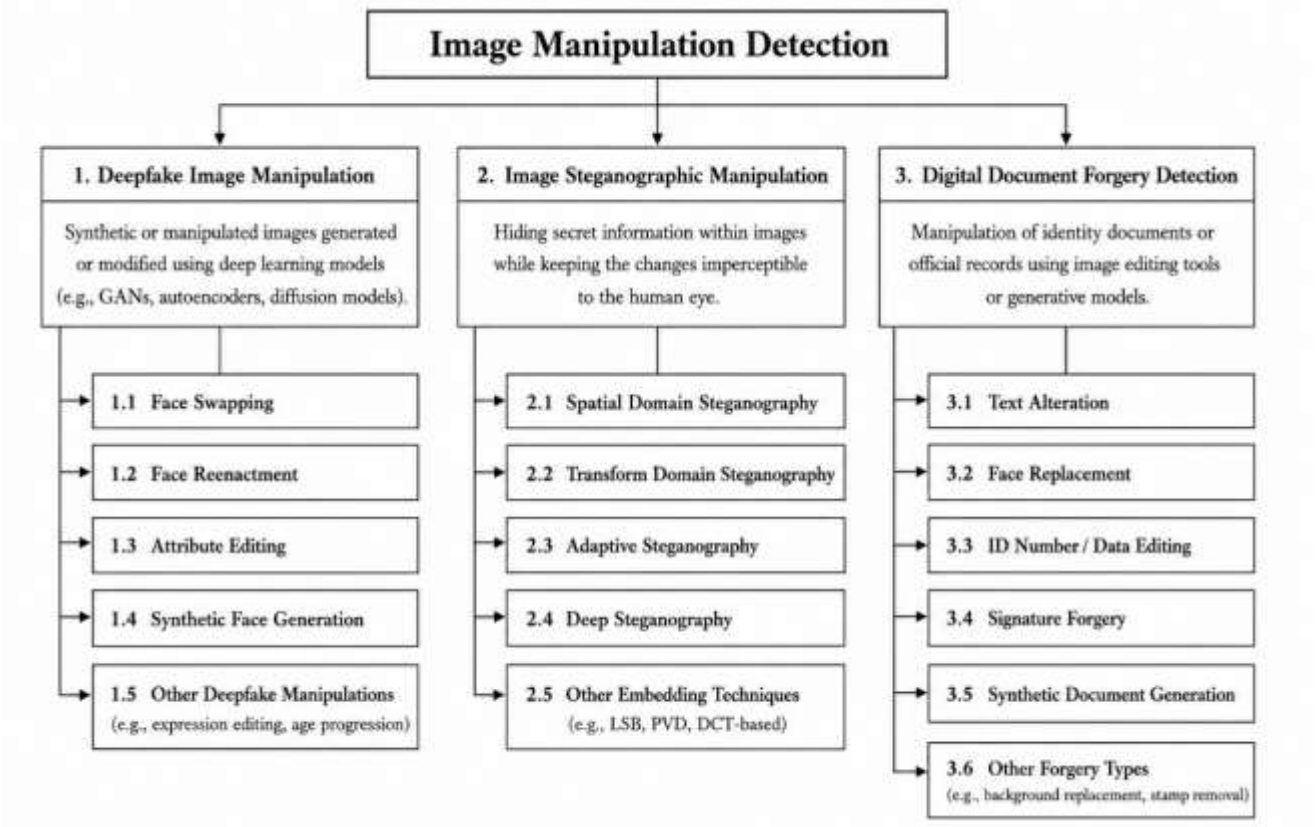

Fig. 2. Taxonomy of deep learning-based image manipulation detection, indicating three primary domains: deepfake image manipulation, image steganographic manipulation, and digital document forgery detection.

4. DEEP LEARNING ARCHITECTURES FOR IMAGE MANIPULATION DETECTION

Deep learning-based approaches are the dominant paradigm for image manipulation detection owing to their capability to learn complex features from the visual data. In the last decade, many architectures have been introduced for different manipulated images. Each of them is characterized by different feature extraction mechanisms, receptive fields, complexity, and other properties and belongs to one of four classes of architectures: CNN, transformer, attention and hybrid. Figure 3 demonstrates the prominent architectures and their complementary characteristics.

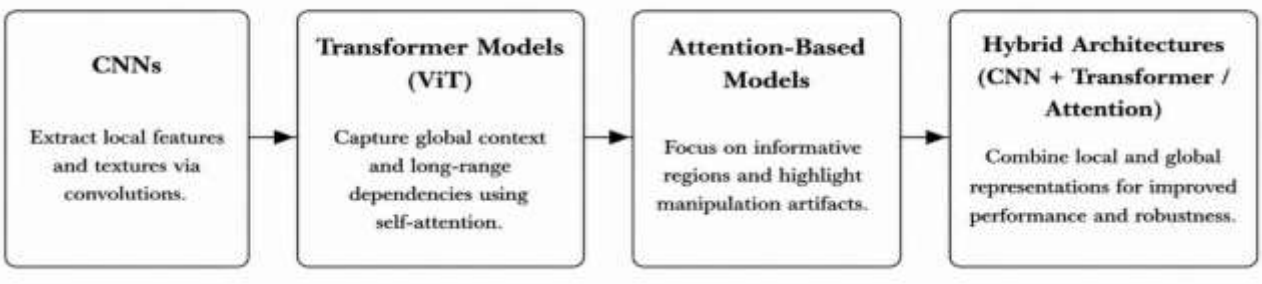

Fig. 3. Notable Deep Learning Architectures for Image Manipulation Detection and Their Complementary Properties

4.1. CONVOLUTIONAL NEURAL NETWORKS (CNNS)

CNNs form the basis of most image manipulation detection methods because their convolutional structure efficiently extracts local features, textures, and the artifacts left by image editing or synthesis. CNN-based

models have proved particularly effective in detecting subtle forensic traces that handcrafted features fail to capture, and they have been used mainly in image steganalysis and deepfake detection [8], [10], [13]. More advanced CNN-based models combine high detection performance with computational efficiency [6], [10], [13].

4.2. TRANSFORMER-BASED MODELS

Transformer architectures have recently emerged as an alternative to CNNs because self-attention allows them to model long-range dependencies. Whereas CNNs analyze local receptive fields, Vision Transformers (ViTs) model the global context of an image. In manipulation detection, a ViT divides the image into a sequence of patches and uses self-attention to relate distant regions, which allows the detector to capture global inconsistencies that local features alone may miss. This property is particularly useful for manipulations that preserve local consistency but alter the global image structure [11], [12]. Transformer-based models have shown strong performance in image manipulation detection, especially in tasks that require modeling both local and global image characteristics [11], [12], [14].

4.3. ATTENTION-BASED MODELS

Attention mechanisms have become an essential part of deep learning architectures because they allow a model to focus selectively on the most informative image regions. Spatial, channel, and self-attention modules improve feature representation by emphasizing manipulation artifacts and suppressing irrelevant context. As a result, attention-based models localize manipulated regions more accurately and maintain high detection performance even when the manipulation is subtle or imperceptible [13], [14]. These mechanisms are now used in many forensic models for different types of image manipulation [13], [14].

4.4. HYBRID ARCHITECTURES

Hybrid architectures are a further stage in this evolution because they combine the strengths of several approaches, typically CNNs for local feature extraction and transformer or attention modules for modeling the global image structure. Such architectures can therefore capture both fine-grained forensic artifacts and long-range semantic relationships within an image, which improves the performance, robustness, and generalization of the detector. However, hybrid models are generally described as requiring more computational resources and having higher model complexity [11], [12], [14], although few studies report these costs in a comparable form (Section 6.2). Overall, the evolution of deep learning architectures for manipulation detection can be seen as a shift from purely local feature extraction towards integrated local–global feature representations, a shift that has substantially improved detection performance.

5. BENCHMARK DATASETS FOR IMAGE MANIPULATION DETECTION

Benchmark datasets are essential for developing, training, and evaluating deep learning models for image manipulation detection. The performance, robustness, and generalization of a detector depend strongly on the quality, size, diversity, and realism of its training and test data. Over the last decade, several public benchmark datasets have been released for deepfake detection, image steganalysis, and digital document forgery detection [7], [32]–[38]. Table 2 summarizes the representative public datasets used by the reviewed studies, grouped by domain. General-purpose image collections used only as a source of real or cover images (e.g., ImageNet, CelebA, and LSUN), as well as private or single-study datasets (e.g., [51]–[53]), are not included.

Table 2. Benchmark Datasets for Image Manipulation Detection

Dataset	Category	Data Type	Main Characteristics	Limitations
FakeIDet2 [7]	Identity Document Forgery	Identity Documents	Privacy-aware benchmark containing realistic forged identity documents and modern evaluation protocols.	Limited coverage of identity-document types and acquisition conditions.
FantasyID [32]	Identity Document Forgery	Identity Documents	Dataset containing multilingual document templates and multiple forgery scenarios for identity-document analysis.	Cross-template and cross-domain generalization remains challenging.
MIDV-2020 [33]	Identity Document Forgery	Identity Documents	Dataset for identity-document analysis under varied acquisition conditions, including different layouts, lighting conditions, and occlusions.	Primarily designed for document analysis rather than explicit forgery detection.
SIDTD [45]	Identity Document Forgery	Identity Documents	Synthetic identity and travel documents with automatically generated forged versions.	Forgeries are synthetic and may not reflect the traces of manually forged documents.
DocTamper [44]	Document Tampering	Document Images	Large-scale set of document images with tampered text and pixel-level annotations, including cross-domain test subsets.	Covers text tampering only; photo substitution and layout-level forgeries are not included.
CASIA v1.0/v2.0 [40]	General Image Forgery	Images	Splicing and copy-move forgeries on natural images; one of the most frequently used forgery benchmarks.	Small, low-resolution images; ground-truth masks were not part of the original release.
Columbia [41]	General Image Forgery	Images	Uncompressed spliced and authentic images captured with different cameras.	Small dataset containing splicing only.
IMD2020 [42]	General Image Forgery	Images	Real-life manipulated images collected from the Internet with manually created masks, together with a large set of images from many camera models.	Heterogeneous image quality and post-processing make results difficult to compare across subsets.
DEFACTO [43]	General Image Forgery	Images	Automatically generated splicing, copy-move, object-removal, and face-morphing forgeries based on MS-COCO images.	Automatic generation may not reflect the traces of manual forgeries.
FaceForensics++ [34]	Deepfake Detection	Videos	Large-scale dataset containing multiple facial manipulation techniques, including DeepFakes, Face2Face, FaceSwap, and NeuralTextures.	Performance can be sensitive to compression artifacts and generalization to unseen manipulations.
Celeb-DF [35]	Deepfake Detection	Videos	High-quality DeepFake videos with fewer visible artifacts, providing a challenging benchmark for detection.	Focuses primarily on celebrity faces, which may limit generalization to non-celebrity identities
DFDC [36]	Deepfake Detection	Videos	Large-scale dataset created for the DeepFake Detection Challenge, containing diverse identities and manipulation scenarios.	Large dataset size may increase computational and storage requirements for training and evaluation.
DeeperForensics-1.0 [46]	Deepfake Detection	Videos	Face-swap videos with real-world perturbations such as compression, noise, and blur, and a hidden test set.	Contains face-swap manipulation only.
BOSSBase 1.01 [37]	Image Steganalysis	Images	Standard benchmark for spatial-domain steganalysis with controlled embedding processes.	Limited content diversity and relatively small dataset size.
BOWS2 [38]	Image Steganalysis	Images	Supplementary benchmark commonly combined with BOSSBase to increase the available training data.	Limited variability compared with large real-world image collections.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Dataset	Category	Data Type	Main Characteristics	Limitations
ALASKA II [39]	Image Steganalysis	Images (JPEG)	JPEG images from many camera models and processing pipelines, with stego images produced by modern JPEG steganographic schemes; designed to be closer to real-world conditions.	JPEG domain only; large size and high diversity increase training cost.

6. COMPARATIVE ANALYSIS OF EXISTING METHODS

This section compares original deep learning-based methods for image manipulation detection in three research domains: image steganalysis, deepfake detection, and digital document forgery detection. The analysis is restricted to original detection methods evaluated on benchmark datasets. For each method, the target manipulation, benchmark datasets, architecture, main contribution, reported performance, evaluation metric, evaluation setting, computational cost, and limitations are recorded. The methods are organized in three comparative tables according to the type of manipulation detection: Table 3(a) for image steganalysis, Table 3(b) for deepfake detection, Table 3(c) for digital document forgery detection. The performance metrics used in comparison conform to the metrics used in original studies (Accuracy, Precision, Recall, F1-score, AUC, AP, PR-AUC, EER, Dice coefficient, or the error probability P_e). Metric and Evaluation setting columns include metrics along with test conditions (payload, JPEG quality, compression, and cross-dataset testing, for example) as it has been done in the corresponding article. The performance values are provided by original articles and should be compared within each row but not between the rows. Cost values are also taken from the original studies where available; otherwise, they come from explicitly cited later comparative studies [54], [55], [56] or from standard backbone specifications [57], [58], [59], [60], and they are labeled accordingly. Because the studies differ in datasets, payloads or compression levels, metrics, and training and testing protocols, two values can be compared only when the dataset, metric, and protocol are the same. For example, the accuracy of 99.81% reported for MAT [13] on raw FaceForensics++ and the AUC of 72.42% reported for SBI [61] on DFDC do not show that MAT is the better detector, because the first is a within-dataset result and the second a cross-dataset result. Table 4 groups selected results reported under similar evaluation settings; even these are shown as trends rather than as a ranking.

Table 3(a). Comparative Analysis of Deep Learning-Based Image Steganalysis Methods

No	Author (Year)	Dataset / Embedding Scheme	Method	Key Contributions	Reported Performance	Metric	Evaluation setting	Cost*	Limitations
1	Tan & Li (2014) [25]	BOSSBase 1.01; HUGO	SCAE	Stacked convolutional autoencoders for CNN pretraining.	P_e = 31.0% (HUGO, 0.4 bpp), compared with 14.0% for SRM in the reported experiment.	P_e	Spatial; HUGO; 0.4 bpp	SCAE autoencoder; size not stated [25]	Requires unsupervised pretraining; lower than SRM in the reported setting.
2	Qian, Dong & Wang (2015) [26]	BOSSBase 1.01; ImageNet	QianNet	Early CNN-based learning of steganalytic representations.	BOSSBase, 0.4 bpp: P_e = 28.9% (HUGO); 29.3% (WOW); 30.9% (S-UNIWARD). ImageNet, 0.4 bpp: P_e = 33.6%, 34.1%, 34.7%, respectively.	P_e	Spatial; 0.4 bpp (BOSSBase, ImageNet)	Shallow CNN; size not stated [26]	Relatively shallow architecture.

No	Author (Year)	Dataset / Embedding Scheme	Method	Key Contributions	Reported Performance	Metric	Evaluation setting	Cost*	Limitations
3	Xu, Wu & Shi (2016) [8]	BOSSBase 1.01	Xu-Net	Constrained high-pass filtering, ABS activation, and truncation.	P_e = 32.4% (WOW, 0.2 bpp); 20.7% (WOW, 0.4 bpp); 39.1% (S-UNIWARD, 0.2 bpp); 27.2% (S-UNIWARD, 0.4 bpp).	P_e	Spatial; 0.2 and 0.4 bpp	≈87.8 K params; 2.14 GFLOPs [56]	Mainly spatial-domain steganalysis.
4	Qian, Dong, Wang & Tan (2016) [9]	BOSSBase 1.01	Transfer CNN	Transfers representations learned at higher payloads to lower-payload detection.	P_e = 30.78% (WOW, 0.2 bpp); 34.38% (S-UNIWARD, 0.2 bpp).	P_e	Spatial; 0.2 bpp	Transfer CNN; count depends on source network [9]	Depends on source/target payload similarity.
5	Ye, Ni & Yi (2017) [27]	BOSSBase 1.01; WOW; S-UNIWARD; HILL	Ye-Net	Uses SRM-based filtering, TLU, and hierarchical feature learning.	P_e = 33.1% (WOW, 0.2 bpp); 40.0% (S-UNIWARD, 0.2 bpp).	P_e	Spatial; 0.2 bpp	≈88.6 K params; 5.77 GFLOPs [56]	Performance varies with payload and embedding algorithm.
6	Wu, Zhong & Liu (2017) [28]	BOSSBase 1.01; WOW; S-UNIWARD; HILL; MiPOD	Residual CNN with Adaptive Content Suppression	Integrates adaptive content suppression with residual learning.	Reported substantially lower detection error than previous CNN methods across the tested steganographic algorithms.	P_e	Spatial; WOW, S-UNIWARD, HILL, and MiPOD; payloads as reported in the source	Residual CNN + content-suppression subnet; count not stated [28]	Mainly spatial-domain steganalysis.
7	Xu (2017) [29]	JPEG BOSSBase; ImageNet CLS-LOC	20-layer Deep CNN	Uses deep CNN depth/pooling analysis for J-UNIWARD detection.	Detection error reduced by approximately 35% compared with previous JPEG CNN methods.	P_e (relative)	JPEG; J-UNIWARD	20-layer CNN; params/FLOPs not stated [29]	Designed specifically for JPEG/J-UNIWARD.
8	Yedroudj, Comby & Chaumont (2018) [62]	BOSSBase 1.01; BOWS2	Yedroudj-Net	Combines high-pass preprocessing, TLU, batch normalization, and CNN learning.	P_e = 27.8% (WOW, 0.2 bpp); 36.7% (S-UNIWARD, 0.2 bpp); 14.1% and 22.8% at 0.4 bpp.	P_e	Spatial; 0.2 and 0.4 bpp	≈252.5 K params; 12.51 GFLOPs [56]	Requires preprocessing and substantial training data.
9	Zhang, Zhu, Liu & Liu (2018) [30]	BOSSBase 1.01; BOWS2	Zhu-Net†	Multi-size CNN with spatial pyramid pooling.	Accuracy = 76.9% (WOW, 0.2 bpp); 88.1% (WOW, 0.4 bpp); 71.4% (S-UNIWARD, 0.2 bpp); 84.5% (S-UNIWARD, 0.4 bpp).	Acc	Spatial; 0.2 and 0.4 bpp	≈10.23 M params; 3.07 GFLOPs [56]	Higher model complexity.

No	Author (Year)	Dataset / Embedding Scheme	Method	Key Contributions	Reported Performance	Metric	Evaluation setting	Cost*	Limitations
10	Li, Wei, Ferreira & Tan (2018) [63]	BOSSBase 1.01; BOWS2	ReST-Net	Uses parallel CNN sub-networks and diverse activation modules.	Accuracy = 71.35% (S-UNIWARD); 70.64% (HILL); 65.14% (CMD-HILL) at 0.2 bpp.	Acc	Spatial; 0.2 bpp	Three parallel CNN subnets; total not stated [63]	Higher memory and training cost.
11	Zeng, Tan, Li & Huang (2018) [64]	ImageNet JPEG datasets	Hybrid Deep-Learning Framework	Combines handcrafted JPEG-domain processing with deep CNNs.	P_e = 48.2%, 43.7%, 40.2%, and 35.5% at 0.1, 0.2, 0.3, and 0.4 bpnzAC.	P_e	JPEG; 0.1–0.4 bpnzAC	JPEG preprocessing + CNN; total not stated [64]	JPEG-specific; high computational cost.
12	Boroumand, Chen & Fridrich (2019) [10]	BOSSBase 1.01; BOWS2; JPEG datasets	SRNet	Deep residual CNN for spatial and JPEG steganalysis.	P_e = 16.76% (WOW); 20.90% (S-UNIWARD); 23.53% (HILL) at 0.2 payload.	P_e	Spatial; 0.2 bpp	≈4.87 M params; 134.77 GFLOPs [56]	High computational and memory requirements.
13	Hu et al. (2019) [65]	BOSSBase 1.01; JPEG J-UNIWARD	VA-DRL	Combines visual attention with deep reinforcement learning for region selection.	P_e = 0.4422 at 0.4 bpnzAC, QF = 75.	P_e	JPEG; QF = 75; 0.4 bpnzAC	CNN + visual attention/deep RL; total not stated [65]	Increased computational complexity and training time.
14	Yu et al. (2024) [16]	ImageNet; MIRFlickr-25K; BOSSBase	FME-Net	Uses source-supervised and target-unsupervised adaptation for cover-source mismatch.	Accuracy improvement = 2.27–9.18 percentage points over J-Net.	Acc (Δ)	Cross-source (source → target)	Source/target adaptation network; total not stated [16]	Requires labeled source and unlabeled target data.
15	Singhal & Bedi (2024) [66]	BOSSBase; transform-domain datasets	USteg-DSE	Combines DenseNet with Squeeze-and-Excitation for universal quantitative steganalysis.	MAE and MSE were lower than the compared universal quantitative steganalysis methods.	MAE / MSE	Transform domain; payload estimation	DenseNet + SE regression network; variant/total not stated [66]	Estimates payload rather than performing binary classification.
16	Wei, Luo, Tan & Huang (2025) [11]	ALASKA II; BOSSBase; BOWS2	CTNet	CNN–Transformer architecture for local and global steganalytic features.	Accuracy = 82.75%, 80.00%, 74.75% (CMD-C-SUNIWARD) and 86.60%, 83.73%, 80.21% (CMD-C-HILL) at 0.4, 0.3, and 0.2 bpc, respectively.	Acc	Spatial colour steganography (CMD-C) on ALASKA II images; 0.4, 0.3, and 0.2 bpc	CNN + Transformer groups, GCP and FC head; total not stated [11]	Higher computational and GPU-memory cost.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

No	Author (Year)	Dataset / Embedding Scheme	Method	Key Contributions	Reported Performance	Metric	Evaluation setting	Cost*	Limitations
17	Wang, Yan & Gu (2025) [12]	BOSSBase 1.01; BOWS2; ALASKA II	CTS-Net	Combines multi-scale residual extraction with CNN–Transformer feature fusion.	Best detection accuracy at 0.1 bpp on BOSSBase; superior detection accuracy to SRNet, SiaStegNet, CvT-Net, and CVTS on ALASKA 2.	Acc	Spatial; 0.1 bpp; JPEG (ALASKA II)	Multi-scale residual CNN–Transformer; total not stated [12]	Higher model complexity

Note: † preprint at the time of the final search. Performance values are not directly comparable across rows because datasets, payloads, metrics, and protocols differ. P_e = probability of error; Acc = accuracy. *Cost is reported as parameters or FLOPs, or described by architecture when a verified total is unavailable. The numerical results for Xu-Net, Ye-Net, Yedroudj-Net, Zhu-Net, and SRNet are provided by the same comparative implementation [56]; another comparison of the same architectures provides different totals for some of them [55]; thus configuration differences should be taken into account when interpreting the numbers. The values are shown in the format of each original source: fractions in the range [0, 1] or percentages.

Table 3(b). Comparative Analysis of Deep Learning-Based Deepfake Detection Methods

No	Author (Year)	Dataset	Method	Key Contributions	Reported Performance	Metric	Evaluation setting	Cost*	Limitations
1	Afchar, Nozick, Yamagishi & Echizen (2018) [67]	DeepFake; FaceForensics	MesoNet	Compact CNN based on mesoscopic facial features.	Accuracy = 98% (DeepFake); 95% (Face2Face).	Acc	DeepFake/Face2Face video-frame tests.	≈28 K params (Meso-4) [67]	Sensitive to heavy compression and unseen manipulations.
2	Güera & Delp (2018) [47]	600 online DeepFake videos	CNN + LSTM‡	Combines CNN spatial features with LSTM temporal modeling.	Accuracy > 97%.	Acc	Video sequences	CNN + LSTM; exact total depends on backbone/head [47]	Requires complete video sequences; unsuitable for single-image detection.
3	Nataraj et al. (2019) [68]	CycleGAN; StarGAN	Co-occurrence Matrices + CNN	Uses RGB co-occurrence features for GAN-generated image detection.	Accuracy > 99% on both CycleGAN and StarGAN datasets.	Acc	CycleGAN/StarGAN; within/cross-dataset tests.	Co-occurrence-matrix CNN; total not stated [68]	Mainly evaluated on early GAN architectures.
4	Nguyen, Yamagishi & Echizen (2019) [69]	DeepFake; FaceForensics; REPLAY-ATTACK	Capsule-Forensics	Capsule network for forged-image/video detection.	Accuracy = 97.69% (DeepFake); 99.23% (DeepFake + noise); 99.37% (Face2Face).	Acc	Clean and noise-added (DeepFake)	≈2.797 M params (old released implementation; version-specific) [70]	Higher computational complexity than conventional CNNs.
5	Xuan, Peng,	CelebA-HQ; PGGAN;	CNN + Image	Suppresses image-	Accuracy = 95.45%	Acc	128×128; PGGAN training;	CNN + preprocessing	Performance decreases

 © 2020 FJIECE, All Rights Reserved

	Wang & Dong (2019) [71]	DCGAN; WGAN-GP	Preprocessing	content cues to emphasize GAN forensic features.	(PGGAN); 64.62% (WGAN-GP); 60.55% (DCGAN).		DCGAN/WGAN-GP tests.	; total not stated [71]	on unseen GAN architectures.
6	Rössler et al. (2019) [34]	FaceForensics++	XceptionNet	Xception-based detection of manipulated facial images.	Accuracy = 99.26% (Raw); 95.73% (HQ); 81.00% (LQ).	Acc	FF++ Raw / HQ / LQ	≈20.86 M params (Xception feature extractor; top excluded) [58]	Performance decreases under compression and unseen manipulations.
7	Li et al. (2020) [6]	FaceForensics++	Face X-Ray	Detects blending-boundary artifacts for generalized face-forgery detection.	AUC = 99.17% (DF); 98.57% (F2F); 98.21% (FS); 98.13% (NT).	AUC	FF++ (DF, F2F, FS, NT)	HRNet + blending-boundary head; total not stated [6]	Depends on blending artifacts.
8	Dang et al. (2020) [72]	DFFD; UADFV; Celeb-DF	Attention-based Multi-task CNN	Uses attention to improve classification and localization of manipulated regions.	AUC 99.64% (DFFD); 71.2% (Celeb-DF; UADFV+DFFD train).	AUC; ACC; EER; TDR; localization	DFFD split; cross-dataset UADFV/Celeb-DF tests.	Multi-task attention CNN; total not stated [72]	Requires manipulation-region supervision for localization.
9	Frank et al. (2020) [73]	FFHQ/StyleGAN; LSUN; CelebA	Frequency Analysis	Exploits frequency-domain artifacts caused by GAN upsampling.	100% test accuracy (FFHQ vs. StyleGAN; DCT/ridge).	Test-set accuracy	FFHQ/StyleGAN; 10k train, 1k val, 5k test/class.	Frequency-based detector; neural count not established [73]	Sensitive to changes in generators and post-processing.
10	Durall, Keuper & Keuper (2020) [74]	CelebA; LSUN; ProGAN; StyleGAN; BigGAN; CycleGAN	Spectral Distribution Analysis	Exploits differences between real and generated spectral distributions.	Accuracy up to 100%.	Acc	CelebA/LSUN; multiple GANs; spectral tests.	Fourier/spectral-feature detector; neural count not established [74]	Frequency artifacts may diminish in newer generators.
11	Jiang et al. (2020) [46]	DeeperForensics-1.0	XceptionNet / Deep Transfer Learning	Evaluates detectors under realistic perturbations and hidden test conditions.	Accuracy = 100.00% (standard); 77.00% (hidden).	Acc	Standard vs. hidden test set	≈20.86 M params (Xception feature extractor; top excluded) [58]	Large standard-to-hidden performance gap.
12	Haliassos, Vougioukas, Petridis & Pantic (2021) [48]	FaceForensics++; Celeb-DF-v2; DFDC	LipForensics‡	Learns fine-grained lip-motion representations.	AUC = 82.4% (Celeb-DF-v2); 73.5% (DFDC); 97.1% (FaceShifter); 97.6% (DeeperForensics).	AUC	Cross-dataset; video	ResNet-18 ≈11.7 M params + MS-TCN; temporal head excluded [48], [60]	Limited to talking-face videos.

13	Zhao et al. (2021) [13]	FaceForensics ++; Celeb-DF	MAT	Uses multiple spatial attention heads, texture enhancement, and regional-independence loss.	Accuracy = 99.81% (FF++ Raw); 90.37% (Celeb-DF).	Acc	FF++ Raw; Celeb-DF	Multi-attention heads + texture-enhancement blocks; total not stated [13]	Performance decreases under compression and unseen manipulations.
14	Gu et al. (2021) [49]	FaceForensics ++; Celeb-DF; DFDC	STIL‡	Joint spatial and temporal inconsistency learning.	Accuracy = 91.78% (NeuralTextures/C40); AUC = 75.58% (Celeb-DF, trained on FF++ c40).	Acc / AUC	FF++ C40; cross-dataset; video	≈23.5 M params (ResNet-50 backbone; classifier top and temporal head excluded) [49], [60]	Requires video sequences and higher computational cost.
15	Shiohara & Yamasaki (2022) [61]	FaceForensics ++; Celeb-DF; DFD; DFDC; DFDCP; FFIW	SBI	Uses synthetic self-blended samples for manipulation-general learning.	AUC = 93.18% (CDF); 97.56% (DFD); 72.42% (DFDC); 86.15% (DFDCP); 84.83% (FFIW).	AUC	Cross-dataset (train FF++); video-level, 32 frames	≈17.55 M params (EfficientNet-B4 feature extractor; top excluded) [59]	Depends on synthetic training quality.
16	Chen et al. (2022) [19]	FaceForensics ++; DFDC; Celeb-DF; DeeperForensics	SLADD	Uses adversarial forgery synthesis with self-supervised configuration prediction.	AUC = 0.772 (DF); 0.730 (F2F); 0.742 (FS); 0.787 (NT).	AUC	Per manipulation (DF, F2F, FS, NT)	≈20.86 M params (Xception feature extractor; top excluded) [58]	Higher training complexity.
17	Haliassos, Mira, Petridis & Pantic (2022) [50]	FaceForensics ++; Celeb-DF-v2; DFDC; FaceShifter; DeeperForensics	RealForensics‡	Self-supervised audio-visual learning from real talking-face videos.	AUC = 86.9% (Celeb-DF-v2); 75.9% (DFDC); 99.7% (FaceShifter); 99.3% (DeeperForensics).	AUC	Cross-dataset; video	Audio-visual temporal model; full count not stated [50]	Requires talking-face videos and large-scale pretraining.
18	Yan et al. (2023) [18]	FaceForensics ++; Celeb-DF; DFD; DFDC; DeeperForensics	UCF	Disentangles common and manipulation-specific forgery features.	AUC = 82.4% (Celeb-DF); 94.5% (DFD); 80.5% (DFDC).	AUC	Cross-dataset	≈20.86 M params (Xception feature extractor; top excluded) [58]	Multi-stage optimization increases complexity.

Note: ‡ video-based method requiring a sequence of frames. Performance values are not directly comparable across rows because datasets, compression levels, metrics, and protocols differ. *Costs marked as backbone-only are estimates from standard model specifications, not counts of the complete forensic detector; for temporal and audio-visual models, the cell states when the temporal or multimodal component is not included. Values are given in the form used by each source, either as fractions between 0 and 1 or as percentages.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Table 3(c). Comparative Analysis of Representative Digital Document Forgery Detection Studies (§ General Image-Forgery Reference Detectors)

No	Author (Year)	Dataset	Method	Key Contributions	Reported Performance	Metric	Evaluation setting	Cost*	Limitations
1	Bibi et al. (2022) [51]	Documents from 20 printers	CNN-based Source Printer Identification	Uses deep CNN features from document patches and characters for source-printer analysis.	Identification rate = 95.52% (patches); 98.06% (characters).	Identification rate	Patch and character level	Patch/character CNN; total not stated [51]	Focuses on printer-source identification rather than direct forgery localization.
2	Lim & Cho (2022) [52]	GAN-generated Hangul and numeric document images	Reinforced CNN Forensic Discriminator	Uses CNN optimization with Gaussian and salt-and-pepper preprocessing for forged text and numeric images.	Accuracy improvement = 20% for Hangul text and 5% for numbers compared with existing state-of-the-art methods.	Acc (Δ)	GAN-generated documents	Reinforced CNN discriminator; total not stated [52]	Focused on GAN-generated text and numeric document forgeries.
3	Jaiswal, Sharma & Yadav (2022) [53]	UWA writing-ink hyperspectral dataset	CAE-LR	Uses convolutional autoencoder features and logistic regression for unsupervised ink-mismatch detection.	Classification accuracy improvement = 4.85% for black ink and 0.13% for blue ink over state-of-the-art results.	Acc (Δ)	UWA hyperspectral ink; black/blue; 2–5 similar types.	Convolutional autoencoder + logistic regression; total not stated [53]	Requires hyperspectral imaging.
4	Guillaro et al. (2023) [75]	Training: CASIA v2; FantasticReality; IMD2020; tampered COCO/RAISE images. Testing: see Evaluation setting.	TruFor§	Fuses RGB and noise-based forensic evidence for forgery detection and localization.	Image-level AUC / balanced Acc: Columbia 0.996 / 0.984; CASIA v1+ 0.916 / 0.813; Coverage 0.770 / 0.680; average over seven test sets 0.857 / 0.781.	AUC / balanced Acc	Image-level detection; tested on CASIA v1+, Coverage, Columbia, NIST16, DSO-1, VIPP, and CocoGlide	68.7 M params; ≈1.17 s per 3.2-MP image [75]	Performance may degrade under strong compression and post-processing.
5	Rouis et al. (2024) [76]	PRImA; Private Sets	Residual CNN	Uses residual representations for document-image manipulation and copy-paste forgery detection.	Accuracy = 95.26% using 30 SRM kernels as initialization.	Acc	PRImA: 64×64 gray patches, 6 classes; private copy-paste tests.	Residual CNN; 30 SRM initialization kernels; total not stated [76]	Broader cross-domain generalization remains insufficiently validated.
6	Diarra, Bissyandé & Poda (2024) [77]	Document Forgery Dataset	Doc-Patch	Uses unsupervised patch-based representation learning and retrieval for document forgery detection.	AP = 96.51%; ROC-AUC = 93.69%.	AP / AUC	Unsupervised patches; FAISS k-NN; AP/AUC.	Patch representation + retrieval; encoder-specific [77]	Depends on reference-database quality and coverage.

7	Aryan et al. (2025) [78]	CASIA v1; DEFACTO; MICC-F2000; Columbia	Lightweight Patch-Based ViT§	Uses patch embedding and multi-head self-attention in a lightweight ViT.	CASIA: AUC = 0.9712, F1 = 0.9587; DEFACTO: 0.9993/0.9914; Columbia: 0.8491/0.8319; MICC-F2000: 0.9804/0.9097.	AUC / F1	Per dataset	69,793 trainable head params; full ViT total not verified [78]	Mainly evaluated on general image-forgery datasets.
8	Li et al. (2025) [15]	StaVer; SCUT-EnsExam	SFMS-Net	Combines spatial-frequency features, HRNet, attention, and multi-scale supervision.	F1 improvement ≈ 5.73%.	F1 (Δ)	StaVer/SCUT-EnsExam; tampering detection/localization/type classification.	HRNet + spatial-frequency attention/multi-scale heads; total not stated [15]	Higher computational complexity.
9	George & Marcel (2025) [14]	FantasyID	EdgeDoc	Combines CNN–Transformer features with NoisePrint for ID-forgery detection and localization.	Weighted F1 = 1.00 on public FantasyID validation; ICCV 2025 challenge F1 = 0.43 (FantasyID), 0.66 (private), 0.59 (aggregate) [14].	Weighted F1 / challenge F1	Public validation and ICCV 2025 challenge; separate protocols	≈1.33 M params (EdgeNeXt-XXS backbone only) [57]	Small manipulated regions remain challenging.
10	Wang, Li & Zhao (2025) [79]	Identity-document tampering datasets	Multistream Network	Fuses texture, frequency, and noise-distribution features for identity-document tampering detection.	F1 = 92.4% (FCD), 82.7% (SCD), and 67.9% (TextTamper).	F1	Per subset (FCD, SCD, TextTamper)	Multi-stream texture/frequency/noise network; total not stated [79]	Mainly focuses on secondary-editing attacks.
11	Bae, Cho & Jung (2025) [80]	DocTamper; Receipt; MIDV-2020	Edge-Focused Deep Learning	Uses edge-aware features to capture subtle document-forgery artifacts.	Accuracy = 0.73 (DenseNet); 0.93 (ViT).	Acc	Per dataset and architecture	Backbones only: DenseNet121 ≈7.0 M; ResNet50 ≈23.5 M; ViT-B/16 ≈85.8 M; EA/EC excluded [60], [80]	Performance varies across datasets and architectures.
12	Li et al. (2026) [81]	DocTD	DCLNet	Uses a ConvNeXt encoder-decoder, global-context attention, and contrastive learning.	Accuracy = 94.9%; F1 = 0.826.	Acc / F1	Desensitized documents	ConvNeXt encoder–decoder + global-context attention; variant/total not stated [81]	Evaluated in the desensitized-document scenario.
13	Díaz Rodríguez et al.	FantasyID	DocVerify	Jointly optimizes identity-document	PR-AUC = 0.9921 ± 0.0058; Dice = 0.856 ±	PR-AUC / Dice / F1	In-dataset; zero-shot on SIDTD	Joint detection/localization model; total not stated [82]	Zero-shot SIDTD performance drops to

 © 2020 FJIECE, All Rights Reserved

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

	(2026) [82]			forgery detection and localization using multi-objective HPO.	0.030; Detection F1 = 0.969 ± 0.014; Localization F1 = 0.807 ± 0.096.				PR-AUC = 0.555.
14	Kumar et al. (2026) [31]	FantasyID	FLiD†	Localizes facial and textual fields and uses lightweight MobileNetV3-Small representations for field-level forgery detection.	AUC = 0.834 (face), 0.926 (text), and 0.837 (combined); EER reduced by 28–29 percentage points versus the full-document baseline.	AUC / EER	Field-level attacks (face, text, combined)	2.55 M params (full-document baseline); 191 K head params; 119 M FLOPs/field [31]	Mainly evaluated on FantasyID and field-level attacks.

Note: † preprint at the time of the final search. Performance metrics cannot be compared in different rows because of different datasets, evaluation criteria, manipulation types, and experiments. *The Cost column distinguishes full models, the estimates for only backbones, heads, and architectural descriptions. The EdgeDoc values are provided only for its backbone, EdgeNeXt-XXS [57], and the FLiD architecture provides a full-document baseline and trainable head for classifier [31]. Entries that give only the architecture description are not efficiency measurements and cannot be compared with the number of parameters and FLOP counts. Values are provided according to notation of each source: either fractions in the range [0, 1] or percentages. § TruFor [75] and the lightweight ViT [78] are general image-forgery detectors whose results were obtained on natural-image benchmarks (e.g., CASIA and Columbia), not on document images. TruFor is included because it is used as a reference detector in ID-document forgery benchmarks, including FantasyID [32]; the lightweight ViT is included as a representative compact transformer-based forgery detector. Their values should not be read as document-forgery performance.

6.1. TRENDS ACROSS ARCHITECTURES

Tables 3(a)–3(c) reveal several trends. Early detectors were mostly CNN-based and captured local spatial, residual, and texture-related forensic artifacts, especially in image steganalysis and facial forgery detection [8]–[10], [34]. Later methods added residual learning, attention mechanisms, feature fusion, and frequency-domain representations to detect subtler manipulation traces [6], [13], [73]. More recently, transformer-based and hybrid CNN–Transformer models have been proposed to combine local feature extraction with global dependency modeling, with promising results in steganalysis and document forgery detection [11], [12], [14], [15].

Although all three domains report clear progress, several problems remain. Many detectors generalize poorly across datasets or manipulation types, especially when tested on unseen datasets, manipulation techniques, or document templates [16], [18], [19], [34]. Transformer-based and hybrid architectures are also generally described as more demanding in memory and computation than conventional CNNs [11], [12], [14]; however, as the Cost* column of Tables 3(a)–3(c) shows, parameter counts and run times are rarely reported in a comparable form, so this statement rests mainly on the qualitative remarks of the original studies (see Section 6.2, item 5). Realistic transformations such as compression, rescaling and addition of noise may reduce the performance of the models [17], [34], [46]. Moreover, different datasets, evaluation criteria, manipulation types and experimental protocols are used in different studies, which prevents cross-domain comparison of deep learning solutions. Deep learning greatly improves the image manipulation detection; however, no existing model provides accuracy on all datasets and transformations, cross-dataset generalization and efficiency at once. These challenges motivate the development of unified and adaptive forensic solutions that use different forensic representations and are efficient. Despite similarities in detection goals of deepfakes and document forgery, these manipulations use different evidence. The deepfake evidence includes facial appearance, temporal or biological inconsistencies and synthesis artifacts

and is exploited by XceptionNet, FaceXray, attention-based methods and self-supervised learning [6], [13], [19], [34], [50]. Document forgery detection, in contrast, relies on textual, layout, edge, frequency, noise, and field-level inconsistencies and uses CNN, transformer, and patch-based architectures for detection and localization [14], [15], [76], [78], [79], [81], [82]. Deepfake detectors are therefore more sensitive to unseen manipulation techniques, compression, and video variations, whereas document forgery detectors are more vulnerable to template variation, small manipulated regions, and cross-domain or zero-shot settings. These differences suggest that a unified forensic framework should retain domain-specific representations while allowing shared learning across manipulation types.

6.2. CROSS-DOMAIN SYNTHESIS

1) Performance over time under similar settings. Table 4 lists results reported under similar evaluation settings in different years. In spatial steganalysis on BOSSBase at 0.2 bpp, the detection error for WOW fell from 32.4% for Xu-Net (2016) [8] to 27.8% for Yedroudj-Net (2018) [62] and 16.76% for SRNet (2019) [10], and the same trend holds for S-UNIWARD (39.1%, 36.7%, and 20.90%). The largest step came with SRNet, which avoids pooling in its early layers and does not rely on fixed high-pass preprocessing, so that the weak stego signal is not suppressed before the network learns it [10]. In deepfake detection, within-dataset accuracy on FaceForensics++ has been close to saturation since 2019 (99.26% for XceptionNet on raw videos [34] and 99.81% for MAT [13]), so progress is better measured by cross-dataset AUC on Celeb-DF. The values reported between 2021 and 2023 range from 75.58% for the video-based STIL [49] to 93.18% [61] and do not increase steadily with publication year. In document forgery detection, no dataset is used by more than three of the fourteen studies in Table 3(c) (FantasyID is used in [14], [82], and [31]), so no comparable time series can be built; this fragmentation is itself one of the findings of the review.

2) Design choices associated with better generalization. Across the three tables, the methods with the best cross-dataset results are not those with the largest backbones but those that change what the network is trained to detect. In deepfake detection, Face X-ray [6] and SBI [61] learn blending artifacts from synthetic training samples instead of the artifacts of a particular generator, SLADD [19] and RealForensics [50] rely on self-supervision, and UCF [18] separates forgery features shared by all manipulation methods from method-specific ones; most of their reported results are on data or manipulation types not seen during training. A plain XceptionNet, in contrast, drops from 99.26% to 81.00% when moving from raw to heavily compressed FaceForensics++ videos [34], and the detector evaluated on DeeperForensics-1.0 falls from 100% on the standard test set to 77% on the hidden test set [46]. In steganalysis, the corresponding problem is cover-source mismatch, which [16] addresses through domain adaptation from labeled source data to unlabeled target data rather than through a deeper network. In document forgery detection, methods that combine RGB features with noise or frequency features [75], [79], [15] exploit traces that are less visible in the image content, while FLiD [31] restricts the classifier to the fields that are usually forged instead of the whole document. The common element is that generalization improves when the training signal is less tied to the source of the training data.

3) Effect of dataset bias. All three domains depend on benchmarks that are narrower than real-world conditions. BOSSBase was used by 16 of the 17 steganalysis studies, although its images come from a small number of cameras and were processed with a single pipeline; ALASKA II was created to reduce this gap [39]. FaceForensics++ or its earlier version FaceForensics was used by 11 of the 18 deepfake studies, and performance drops when detectors trained on it are tested on Celeb-DF or DFDC (e.g., AUCs of 72.42% for SBI [61] and 73.5% for LipForensics [48] on DFDC). In document forgery detection, results are reported on many small or private datasets, and the only zero-shot result in Table 3(c), DocVerify trained on FantasyID and tested on SIDTD, drops from a PR-AUC of 0.9921 to 0.555 [82]. High scores in all three tables therefore partly reflect the benchmark and not only the method.

4) Gaps common to the three domains. Four gaps appear in all three domains. (i) Source shift: the same problem appears as cover-source mismatch in steganalysis [16], as unseen generators or datasets in deepfake detection [18], [61], and as unseen document templates in document forgery detection [82], yet each domain evaluates it with its own protocol. (ii) Post-processing: JPEG recompression, resizing, and additive noise weaken the low-level traces on which all three domains rely [17], [34], [46]. (iii) Detection versus localization: most steganalysis and deepfake methods produce only an image-level decision, whereas document forgery methods increasingly localize the forged region [14], [75], [82], which is useful for forensic reporting in every domain. (iv) Evaluation: there is no common set of metrics (P_e in steganalysis, AUC in deepfake detection, and F1-score or PR-AUC in document forgery detection) and no common reporting of computational cost (see item 5), and robustness to adversarial attacks [83] is not reported for the methods in Tables 3(a)–3(c). A method designed for one domain should therefore also be tested under the source-shift and post-processing conditions used in the other two.

5) Metric comparability and computational cost. The metrics in Tables 3(a)–3(c) are not interchangeable, so no cross-metric ranking is made and no metric is converted into another. A single normalized value cannot be derived out of results reported in the literature since conversion among P_e, accuracy, AUC, AP, PR-AUC, F1-score, and EER requires sample-level values and decision thresholds for each detector, which is not provided in the original papers. For balanced steganalysis, the ($100-P_e$) value is included in Table 4 just as descriptive measure but not as substitute for the accuracy reported in other papers. AUC, AP, PR-AUC, F1-score and EER are compared only when the metric, dataset, and protocol match. The Evaluation setting column includes the values of payload, JPEG quality or compression ratio and conditions of cross-domain and hidden tests when these values are provided by the source; otherwise, the column reports only task and dataset names. The Cost* column differentiates full-models from the backbones and individual modules; results obtained in different tasks and hardware configurations cannot be compared directly. The claims that transformer or hybrid methods have higher cost are not justified without benchmark comparison. Future works should report the number of parameters, FLOPs, and the inference time of each detector as well as hardware; they should also test the detectors for both pristine and compressed images.

Table 4. Selected Results Reported Under Similar Evaluation Settings

Domain and setting	Method (Year)	Result
Steganalysis: BOSSBase 1.01, 0.2 bpp, P_e (%) for WOW / S-UNIWARD (error-rate complement $100 - P_e$, balanced test set)	Xu-Net (2016) [8]	32.4 / 39.1 ($100 - P_e$: 67.6 / 60.9)
	Transfer CNN (2016) [9]	30.78 / 34.38 ($100 - P_e$: 69.22 / 65.62)
	Ye-Net (2017) [27]	33.1 / 40.0 ($100 - P_e$: 66.9 / 60.0)
	Yedroudj-Net (2018) [62]	27.8 / 36.7 ($100 - P_e$: 72.2 / 63.3)
	SRNet (2019) [10]	16.76 / 20.90 ($100 - P_e$: 83.24 / 79.10)
Deepfake detection: cross-dataset AUC (%) on Celeb-DF, training on FaceForensics++	LipForensics‡ (2021) [48]	82.4
	STIL‡ (2021) [49]	75.58
	SBI (2022) [61]	93.18
	RealForensics‡ (2022) [50]	86.9
	UCF (2023) [18]	82.4

Note: ‡ video-based method. Even within one setting, training details such as compression level, frame sampling, and video- or frame-level evaluation differ between studies, so the values show trends and are not a ranking.

7. CHALLENGES AND OPEN PROBLEMS

Even with the advancements of deep learning-based image manipulation detection methods, there are still many unsolved problems. Firstly, the generalization over different datasets is weak: the models with good performance on benchmark datasets demonstrate poor performance on unseen data and new manipulation techniques, which limits practical applicability [18], [6], [61]. Secondly, the detectors are vulnerable to the

adversarial examples and the anti-forensic interventions as carefully constructed perturbations deceive detectors but are visually imperceptible [83]. Thirdly, some detectors suffer from instability in case of typical image transformations such as JPEG compression, resizing, noise and blurring [17]. Since the images in the real world are subjected to the transformations, the models trained on pristine images underperform on them. Fourth, computational cost may limit deployment: transformer and hybrid architectures are reported to be more demanding than conventional CNNs, which can hinder their use in real-time or resource-constrained settings [11], [14], although the available studies do not provide standardized cost reports that would allow a controlled comparison (Section 6.2). Finally, the lack of unified benchmarking frameworks and standardized evaluation protocols makes it difficult to compare existing models. Differences in datasets, manipulation types, evaluation metrics, and experimental protocols can lead to inconsistent or even contradictory conclusions.

Addressing these issues is essential for improving the reliability, generalization, and practical applicability of future image manipulation detection systems.

8. COMPARISON AND DISCUSSION

The benchmark datasets listed in Table 2 are central to evaluating deep learning-based image manipulation detectors. Well-known deepfake benchmarks such as FaceForensics++ [34], Celeb-DF [35], and the DeepFake Detection Challenge (DFDC) dataset [36] are widely used because they provide large collections of facial images and videos manipulated with several techniques, which allows detectors to be compared under diverse manipulation conditions. However, these benchmarks have limitations related to compression and to generalization to unseen manipulation techniques [34]–[36].

In image steganalysis, BOSSBase and BOWS2 are commonly used because their controlled embedding process allows consistent comparison between methods [37], [38]. Although these datasets have supported major progress in spatial-domain steganalysis, their limited content diversity may restrict the generalization of some models [10], and ALASKA II [39] was introduced to provide JPEG images closer to real acquisition conditions. For general image forgery, CASIA [40], Columbia [41], IMD2020 [42], and DEFACTO [43] are used by the general image-forgery detectors TruFor [75] and the lightweight ViT of [78], which are included in Table 3(c) as reference detectors (see the note under the table). For identity-document and document forgery detection, representative benchmarks include MIDV-2020 [33], FakeIDet2 [7], FantasyID [32], SIDTD [45], and DocTamper [44], which contain document images captured or generated under different conditions. However, large and diverse benchmark datasets specifically designed for document forgery detection remain limited [7], [32].

9. FUTURE RESEARCH DIRECTIONS

Future research must address increasingly complex manipulation types. First, multimodal learning that combines visual data with complementary information is a promising way to improve the robustness and scope of forensic analysis [50]. Second, self-supervised learning can reduce the dependence on large annotated datasets and improve the generalization of detection models [19]. Lightweight architectures with low computational cost are also important, particularly for real-time and resource-constrained applications [14], [78]; to support this direction, future studies should report parameters, FLOPs, and inference time in a standard way. Another important direction is robustness to adversarial attacks and common image degradations such as compression, noise, and blur, which is essential for practical deployment [83], [17], [46]. Finally, the rapid development of generative AI, including diffusion models, introduces new forensic challenges [5] that call for adaptive forensic frameworks able to detect previously unseen manipulation techniques and to generalize across emerging manipulation scenarios [18], [19].

10. CONCLUSION

This survey has reviewed deep learning-based image manipulation detection, with particular emphasis on three domains: deepfake detection, image steganalysis, and digital document forgery detection. A unified taxonomy was proposed to organize existing approaches, and benchmark datasets, architectures, representative methods, evaluation metrics and settings, reported performance, computational cost, and limitations were analyzed systematically. The comparative analysis indicates that CNN-based approaches are effective at modeling local forensic features, whereas transformer-based and hybrid CNN–Transformer architectures add complementary local–global representations and better modeling of long-range dependencies. The open challenges include limitations in generalization, high computation costs, vulnerability to adversarial examples and anti-forensic methods, instability under realistic image transformations, lack of standardized metrics and cost calculation in research. Further research should focus on the robustness, computation cost reduction and generalization. Prospective directions include multi-modal learning, lightweight architectures, and adaptable forensic approaches which can take into account new manipulation techniques.

DECLARATIONS

Funding: Not applicable.

Conflicts of interest: The authors declare that they have no conflict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Data availability: No new data were created or analyzed in this study.

REFERENCES

[1] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “DeepFakes and beyond: A survey of face manipulation and fake detection,” Information Fusion, vol. 64, pp. 131–148, 2020, doi: 10.1016/j.inffus.2020.06.014.

[2] J. De La Croix, T. Ahmad, and F. Han, “Comprehensive survey on image steganalysis using deep learning,” Array, vol. 22, Art. no. 100353, 2024, doi: 10.1016/j.array.2024.100353.

[3] A. B. Hassan and Y. A. Fadlalla, “A survey on techniques of detecting identity documents forgery,” in Proc. Sudan Conference on Computer Science and Information Technology (SCCSIT), 2017, doi: 10.1109/SCCSIT.2017.8293052.

[4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8110–8119, doi: 10.1109/CVPR42600.2020.00813.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695, doi: 10.1109/CVPR52688.2022.01042.

[6] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, “Face X-ray for More General Face Forgery Detection,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[7] J. Muñoz-Haro, R. Tolosana, J. Fierrez, R. Vera-Rodriguez, and A. Morales, “Privacy-aware detection of fake identity documents: Methodology, benchmark, and improved algorithms (FakeIDet2),” Information Fusion, vol. 128, Art. no. 103969, 2026, doi: 10.1016/j.inffus.2025.103969.

 © 2020 FJIECE, All Rights Reserved

[8] G. Xu, H.-Z. Wu, and Y.-Q. Shi, "Structural Design of Convolutional Neural Networks for Steganalysis," IEEE Signal Processing Letters, vol. 23, no. 5, pp. 708–712, 2016, doi: 10.1109/LSP.2016.2548421.

[9] Y. Qian, J. Dong, W. Wang, and T. Tan, "Learning and Transferring Representations for Image Steganalysis Using Convolutional Neural Network," in Proc. IEEE International Conference on Image Processing (ICIP), 2016, pp. 2752–2756, doi: 10.1109/ICIP.2016.7532860.

[10] M. Boroumand, M. Chen, and J. Fridrich, "Deep Residual Network for Steganalysis of Digital Images," IEEE Transactions on Information Forensics and Security, vol. 14, no. 5, pp. 1181–1193, 2019, doi: 10.1109/TIFS.2018.2871749.

[11] K.-K. Wei, W.-Q. Luo, S.-Q. Tan, and J.-W. Huang, "CTNet: A Convolutional Transformer Network for Color Image Steganalysis," Journal of Computer Science and Technology, vol. 40, no. 2, pp. 413–427, 2025, doi: 10.1007/s11390-023-3006-3.

[12] J. Wang, H. Yan, and J. Gu, "Image Steganalysis Based on CNN-Transformer," Journal of Shenzhen University Science and Engineering, vol. 42, no. 2, pp. 233–241, 2025, doi: 10.3724/SP.J.1249.2025.02233.

[13] H. Zhao, Y. Zhou, X. Liu, Z. Cao, and K. Zhang, "Multi-attentional Deepfake Detection," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[14] A. George and S. Marcel, "EdgeDoc: Hybrid CNN-Transformer model for accurate forgery detection and localization in ID documents," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW), 2025. [Online]. Available: arXiv:2508.16284.

[15] L. Li, Y. Bai, S. Zhang, and M. Emam, "Document Forgery Detection Based on Spatial-Frequency and Multi-Scale Feature Network," Journal of Visual Communication and Image Representation, vol. 107, Art. no. 104393, 2025, doi: 10.1016/j.jvcir.2025.104393.

[16] L. Yu, Z. Zhang, S. Weng, P. Cao, and G. Cao, "A Deep Steganalysis Network Combining Source-Supervised and Target-Unsupervised Information for Cover-Source Mismatch," Expert Systems with Applications, vol. 255, Art. no. 124790, 2024, doi: 10.1016/j.eswa.2024.124790.

[17] O. A. Alrusaini, "Deep Learning for Steganalysis: Evaluating Model Robustness Against Image Transformations," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1532895, 2025, doi: 10.3389/frai.2025.1532895.

[18] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "UCF: Uncovering Common Features for Generalizable Deepfake Detection," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

[19] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-Supervised Learning of Adversarial Examples: Towards Good Generalizations for Deepfake Detection," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[20] L. Verdoliva, "Media forensics and deepfakes: An overview," IEEE J. Sel. Topics Signal Process., vol. 14, no. 5, pp. 910–932, 2020, doi: 10.1109/JSTSP.2020.3002101.

[21] M. Zanardelli, F. Guerrini, R. Leonardi, and N. Adami, "Image forgery detection: A survey of recent deep-learning approaches," Multimedia Tools and Applications, vol. 82, no. 12, pp. 17521–17566, 2023, doi: 10.1007/s11042-022-13797-w.

[22] B. Kitchenham and S. Charters, "Guidelines for Performing Systematic Literature Reviews in Software Engineering," EBSE Technical Report EBSE-2007-01, Keele University, 2007.

[23] H. Snyder, "Literature Review as a Research Methodology: An Overview and Guidelines," Journal of Business Research, vol. 104, pp. 333–339, 2019, doi: 10.1016/j.jbusres.2019.07.039.

[24] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, et al., "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews," BMJ, vol. 372, Art. no. n71, 2021, doi: 10.1136/bmj.n71.

[25] S. Tan and B. Li, "Stacked Convolutional Auto-Encoders for Steganalysis of Digital Images," in Proc. APSIPA Annual Summit and Conference, 2014.

[26] Y. Qian, J. Dong, and W. Wang, "Deep Learning for Steganalysis via Convolutional Neural Networks," in Proc. SPIE Media Watermarking, Security, and Forensics, vol. 9409, 2015, Art. no. 94090J.

[27] J. Ye, J. Ni, and Y. Yi, "Deep Learning Hierarchical Representations for Image Steganalysis," IEEE Transactions on Information Forensics and Security, vol. 12, no. 11, pp. 2545–2557, 2017, doi: 10.1109/TIFS.2017.2710946.

[28] S. Wu, S.-H. Zhong, and Y. Liu, "Residual Convolution Network Based Steganalysis with Adaptive Content Suppression," in Proc. IEEE International Conference on Multimedia and Expo (ICME), 2017, pp. 241–246, doi: 10.1109/ICME.2017.8019304.

[29] G. Xu, "Deep Convolutional Neural Network to Detect J-UNIWARD," in Proc. ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec), 2017, pp. 67–73.

[30] R. Zhang, F. Zhu, J. Liu, and G. Liu, "Efficient Feature Learning and Multi-Size Image Steganalysis Based on CNN," arXiv:1807.11428, 2018.

[31] A. Kumar, R. Tapwal, C. Maple, and M. Hooper, "Field-localized forgery detection for digital identity documents," arXiv:2605.09089v2, 2026 (preprint).

[32] P. Korshunov, A. Mohammadi, V. Vidit, C. Ecabert, and S. Marcel, "FantasyID: A dataset for detecting digital manipulations of ID-documents," in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), 2025.

[33] A. V. Arlazarov, K. Bulatov, T. Chernov, and V. V. Arlazarov, "MIDV-2020: A benchmark dataset for identity document analysis and recognition," in Proc. Int. Conf. Pattern Recognit. Workshops, 2020.

[34] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1–11.

[35] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020.

[36] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton-Ferrer, "The DeepFake Detection Challenge (DFDC) Dataset," arXiv:2006.07397, 2020.

[37] P. Bas, T. Filler, and T. Pevný, "Break Our Steganographic System: The Ins and Outs of Organizing BOSS," in Information Hiding, Lecture Notes in Computer Science, vol. 6958, pp. 59–70, 2011.

[38] P. Bas and T. Furon, "BOWS-2: Break Our Watermarking System," in Proceedings of the International Workshop on Information Hiding, 2007.

[39] R. Cogranne, E. Giboulot, and P. Bas, "ALASKA#2: Challenging academic research on steganalysis with realistic images," in Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS), 2020, pp. 1–5, doi: 10.1109/WIFS49906.2020.9360896.

[40] J. Dong, W. Wang, and T. Tan, "CASIA image tampering detection evaluation database," in Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP), 2013, pp. 422–426.

[41] Y.-F. Hsu and S.-F. Chang, "Detecting image splicing using geometry invariants and camera characteristics consistency," in Proc. IEEE Int. Conf. Multimedia Expo (ICME), 2006, pp. 549–552, doi: 10.1109/ICME.2006.262447.

[42] A. Novozámský, B. Mahdian, and S. Saic, "IMD2020: A large-scale annotated dataset tailored for detecting manipulated images," in Proc. IEEE Winter Conf. Appl. Comput. Vis. Workshops (WACVW), 2020, pp. 71–80, doi: 10.1109/WACVW50321.2020.9096940.

[43] G. Mahfoudi, B. Tajini, F. Retraint, F. Morain-Nicolier, J.-L. Dugelay, and M. Pic, "DEFACTO: Image and face manipulation dataset," in Proc. 27th Eur. Signal Process. Conf. (EUSIPCO), 2019, doi: 10.23919/EUSIPCO.2019.8903181.

[44] C. Qu, C. Liu, Y. Liu, X. Chen, D. Peng, F. Guo, and L. Jin, "Towards robust tampered text detection in document image: New dataset and new solution," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 5937–5946, doi: 10.1109/CVPR52729.2023.00575.

[45] C. Boned, M. Talarmain, N. Ghanmi, G. Chiron, S. Biswas, A. M. Awal, and O. Ramos Terrades, "Synthetic dataset of ID and travel documents," Scientific Data, vol. 11, Art. no. 1356, 2024, doi: 10.1038/s41597-024-04160-9.

[46] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2889–2898.

[47] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018.

[48] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 5039–5049, doi: 10.1109/CVPR46437.2021.00500.

[49] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, F. Huang, and L. Ma, "Spatiotemporal inconsistency learning for DeepFake video detection," in Proc. 29th ACM Int. Conf. Multimedia (MM), 2021, doi: 10.1145/3474085.3475508.

[50] A. Haliassos, R. Mira, S. Petridis, and M. Pantic, "Leveraging real talking faces via self-supervision for robust forgery detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 14950–14962, doi: 10.1109/CVPR52688.2022.01453.

[51] M. Bibi, A. Hamid, M. Moetesum, and I. Siddiqi, "Document forgery detection using source printer identification: A comparative study of text-dependent versus text-independent analysis," Expert Systems, vol. 39, no. 8, Art. no. e13020, 2022, doi: 10.1111/exsy.13020.

[52] S.-Y. Lim and J. Cho, "Reinforced CNN forensic discriminator to detect document forgery by DCGAN," Computers, Materials & Continua, vol. 71, no. 3, pp. 6039–6051, 2022, doi: 10.32604/cmc.2022.024862.

[53] G. Jaiswal, A. Sharma, and S. K. Yadav, "Deep feature extraction for document forgery detection with convolutional autoencoders," Computers & Electrical Engineering, vol. 99, Art. no. 107770, 2022, doi: 10.1016/j.compeleceng.2022.107770.

[54] J.-F. Chen, Z.-J. Fu, W.-M. Zhang, X. Cheng, and X.-M. Sun, "Review of Image Steganalysis Based on Deep Learning," Journal of Software, vol. 32, no. 2, pp. 551–578, 2021 (in Chinese), doi: 10.13328/j.cnki.jos.006135.

[55] G. Xu, Y. Xu, S. Zhang, and X. Xie, "SFRNet: Feature Extraction-Fusion Steganalysis Network Based on Squeeze-and-Excitation Block and RepVgg Block," Security and Communication Networks, 2021, Art. no. 3676720, doi: 10.1155/2021/3676720.

[56] J. Tabares-Soto et al., "Sensitivity of deep learning applied to spatial image steganalysis," PeerJ Computer Science, vol. 7, e616, 2021, doi: 10.7717/peerj-cs.616.

[57] M. Maaz et al., "EdgeNeXt: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications," in Proc. European Conference on Computer Vision (ECCV), 2022, pp. 3–20.

[58] Keras Team, “Xception,” Keras Applications API documentation, https://keras.io/api/applications/xception/ (accessed Oct. 12, 2026).

[59] Keras Team, “EfficientNet B0 to B7,” Keras Applications API documentation, https://keras.io/api/applications/efficientnet/ (accessed Oct. 12, 2026).

[60] TorchVision maintainers, “Models and pretrained weights,” PyTorch/TorchVision documentation, https://pytorch.org/vision/stable/models.html (accessed Oct. 12, 2026).

[61] K. Shiohara and T. Yamasaki, “Detecting Deepfakes with Self-Blended Images,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18720–18729.

[62] M. Yedroudj, F. Comby, and M. Chaumont, “Yedroudj-Net: An Efficient CNN for Spatial Steganalysis,” in Proc. IEEE ICASSP, 2018, pp. 2092–2096.

[63] 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, vol. 25, no. 5, pp. 650–654, 2018.

[64] J. Zeng, S. Tan, B. Li, and J. Huang, “Large-Scale JPEG Image Steganalysis Using Hybrid Deep-Learning Framework,” IEEE Transactions on Information Forensics and Security, vol. 13, no. 5, pp. 1200–1214, 2018.

[65] D. Hu et al., “Digital image steganalysis based on visual attention and deep reinforcement learning,” IEEE Access, vol. 7, 2019.

[66] A. Singhal and P. Bedi, “USteg-DSE: Universal quantitative steganalysis framework using DenseNet merged with squeeze & excitation net,” Signal Processing: Image Communication, vol. 128, Art. no. 117171, 2024, doi: 10.1016/j.image.2024.117171.

[67] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: A Compact Facial Video Forgery Detection Network,” in Proc. IEEE International Workshop on Information Forensics and Security (WIFS), 2018.

[68] L. Nataraj, T. M. Mohammed, B. S. Manjunath, S. Chandrasekaran, A. Flenner, J. Bappy, and A. K. Roy-Chowdhury, “Detecting GAN Generated Fake Images Using Co-Occurrence Matrices,” Electronic Imaging, 2019.

[69] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos,” in Proc. IEEE ICASSP, 2019, pp. 2307–2311.

[70] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-Forensics: Official implementation,” GitHub repository, https://github.com/nii-yamagishilab/Capsule-Forensics (accessed Oct. 12, 2026).

[71] X. Xuan, B. Peng, W. Wang, and J. Dong, “On the Generalization of GAN Image Forensics,” in Proc. Chinese Conference on Biometric Recognition (CCBR), 2019, pp. 134–141.

[72] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, “On the Detection of Digital Face Manipulation,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[73] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, “Leveraging Frequency Analysis for Deep Fake Image Recognition,” in Proc. International Conference on Machine Learning (ICML), 2020.

[74] R. Durall, M. Keuper, and J. Keuper, “Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[75] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, “TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 20606–20615, doi: 10.1109/CVPR52729.2023.01974.

[76] K. Rouis, P. Gomez-Krämer, M. Coustaty, S. Kébairi, and V. Poulain d'Andecy, "Deep discriminative feature learning for document image manipulation detection," in Proc. 19th Int. Conf. Comput. Vis. Theory Appl. (VISAPP), vol. 4, 2024, pp. 567–574, doi: 10.5220/0012410800003660.

[77] A. Diarra, T. F. Bissyandé, and P. Poda, "Doc-Patch: An unsupervised approach for documents forgery detection," in Proc. 7th Int. Conf. Algorithms, Computing and Artificial Intelligence (ACAI), 2024, doi: 10.1109/ACAI63924.2024.10899704.

[78] P. Aryan, R. M. R. Yanamala, A. Pallakonda, R. D. Amar Raj, and K. K. Prakasha, "Lightweight end-to-end patch-based self-attention network for robust image forgery detection," IEEE Access, vol. 13, pp. 157674–157686, 2025, doi: 10.1109/ACCESS.2025.3607206.

[79] L. Wang, Z. Li, and W. Zhao, "Research on identity document image tampering detection based on texture understanding and multistream networks," Journal of Electronic Imaging, vol. 34, no. 4, Art. no. 043018, 2025, doi: 10.1117/1.JEI.34.4.043018.

[80] Y.-Y. Bae, D.-J. Cho, and K.-H. Jung, "Enhancing document forgery detection with edge-focused deep learning," Symmetry, vol. 17, no. 8, Art. no. 1208, 2025, doi: 10.3390/sym17081208.

[81] W. Li, B. Li, K. Zheng, S. Li, and H. Li, "Document image forgery detection and localization in desensitization scenarios," Signal Processing, vol. 238, Art. no. 110123, 2026, doi: 10.1016/j.sigpro.2025.110123.

[82] H. Díaz Rodríguez et al., "Multi-objective Pareto hyperparameter optimization for joint detection and localization of forged identity documents," Pattern Recognition Letters, vol. 209, pp. 22–28, 2026, doi: 10.1016/j.patrec.2026.08.021.

[83] P. Neekhara, B. Dolhansky, J. Bitton, and C. C. Ferrer, "Adversarial threats to DeepFake detection: A practical perspective," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2021, pp. 923–932.

ATU-FJIECE, Volume: 5, Issue: 2 , 11,23 , 2026, © 2020 FJIECE, All Rights Reserved