




## Review

# Detection of Manipulations in Digital Images: A Review of Passive and Active Methods Utilizing Deep Learning

Paweł Duszejko , Tomasz Walczyna  and Zbigniew Piotrowski 

Faculty of Electronics, Military University of Technology, 00-908 Warszawa, Poland;  
tomasz.walczyna@wat.edu.pl (T.W.); zbigniew.piotrowski@wat.edu.pl (Z.P.)

\* Correspondence: pawel.duszejko@wat.edu.pl

**Abstract:** The modern society generates vast amounts of digital content, whose credibility plays a pivotal role in shaping public opinion and decision-making processes. The rapid development of social networks and generative technologies, such as deepfakes, significantly increases the risk of disinformation through image manipulation. This article aims to review methods for verifying images' integrity, particularly through deep learning techniques, addressing both passive and active approaches. Their effectiveness in various scenarios has been analyzed, highlighting their advantages and limitations. This study reviews the scientific literature and research findings, focusing on techniques that detect image manipulations and localize areas of tampering, utilizing both statistical properties of images and embedded hidden watermarks. Passive methods, based on analyzing the image itself, are versatile and can be applied across a broad range of cases; however, their effectiveness depends on the complexity of the modifications and the characteristics of the image. Active methods, which involve embedding additional information into the image, offer precise detection and localization of changes but require complete control over creating and distributing visual materials. Both approaches have their applications depending on the context and available resources. In the future, a key challenge remains the development of methods resistant to advanced manipulations generated by diffusion models and further leveraging innovations in deep learning to protect the integrity of visual content.



Academic Editors: Abdussalam Elhanashi and Pierpaolo Dini

Received: 1 December 2024

Revised: 10 January 2025

Accepted: 14 January 2025

Published: 17 January 2025

**Citation:** Duszejko, P.; Walczyna, T.; Piotrowski, Z. Detection of Manipulations in Digital Images: A Review of Passive and Active Methods Utilizing Deep Learning. *Appl. Sci.* **2025**, *15*, 881. <https://doi.org/10.3390/app15020881>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** image manipulation; deep learning; active protection; passive protection; image forensics; deepfakes; authenticity verification; multimodal analysis; watermarking

## 1. Introduction

Image manipulation is not a new phenomenon. From the dawn of photography and cinematography, intentional attempts to modify images were undertaken, sometimes in very rudimentary ways, to add special effects [1], introduce simple elements such as end credits in films, or alter landscapes or their components [2]. History demonstrates how dangerous this practice can be, often carrying political implications. A notable example is the infamous removal of Nikolai Yezhov from a photograph with Stalin in 1937, aimed at erasing evidence of his association with the ruling regime [3], as shown in Figure 1.

Throughout most of the 20th century, image and video manipulation was a labor-intensive task performed primarily by skilled artists [5]. Using templates or their own creativity, they modified parts of an image or added new elements. This process was time-consuming, requiring specialized skills and often artistic training to accurately recreate, for example, the perspective or the lighting distribution of a scene [6]. However, the

late 20th and early 21st centuries brought a breakthrough in the field of graphics and multimedia with the advent of high-performance industrial computers, followed shortly by personal computers. This development necessitated the digitization of multimedia, significantly simplifying and accelerating the processes of manipulation and editing while also improving the quality of the resulting materials [7].



**Figure 1.** An example of reality falsification by a totalitarian regime: the removal of Nikolai Yezhov from a photograph with Stalin [4].

With the popularization of graphic editing software, often available under open licenses, such as GIMP, high-resolution multimedia content editing became accessible to a wide range of users. However, it still required appropriate skills and specialized equipment. Subsequent technological breakthroughs [8], particularly the development of social media and artificial intelligence offering advanced processing algorithms, have made it possible for virtually anyone to edit multimedia content almost professionally, thereby increasing the risks to its integrity. As a result, the phenomenon of the “deepfake” is increasingly discussed across various types of media [9].

In this article, we discuss recent changes in the concept of image manipulation and the detection of such alterations, with a particular focus on methods based on deep learning. We will present typical manipulation scenarios and explore detection techniques that leverage deep learning methodologies.

## 2. Image Manipulating Methods

Based on existing research [10–16] in the field of image manipulation, several types of modifications to the original image at the pixel level can be identified. One such technique is steganography [15,16], which involves covert communication by embedding a hidden message within an overt medium. In this context, it refers to embedding data in the spatial domain of the image (pixel values) or the frequency domain (e.g., using Discrete Cosine Transform, DCT). This manipulation is typically minimal, often altering the least significant bits of the image. High-entropy regions are often chosen for such modifications, as they enable changes that remain imperceptible to human perception [17]. Steganography stands out as one of two image processing techniques (alongside digital watermarking) that aim to minimize alterations to the structure of the original image.

When we speak of image manipulation, we often refer to actions that alter the interpretation of the photograph. Based on research [18,19], several primary types of such operations can be identified:

- Retouching involves subtle modifications aimed at enhancing aesthetics or removing minor imperfections without significantly altering the semantics of the image or its overall message. Retouching may include adjustments to color, contrast, lighting, depth of field, and the elimination of unwanted artifacts such as noise, discoloration, or vignetting. Such techniques are commonly employed in photography to

improve the visual appeal of an image from the viewer's perspective, thereby influencing its emotional impact. Retouching is frequently used in portrait, fashion, and advertising photography.

- The copy-and-move operation involves duplicating a single object or a group of objects within the same image. This process is often supplemented by geometric transformations, such as rotation or perspective changes, as well as pixel value modifications, including blurring, contrast adjustment, and brightness regulation. Since the manipulation occurs within a single image, the objects share similar characteristics (having been captured at the same time and scene using the same sensor). This makes the transformation relatively straightforward to execute, making it one of the most commonly encountered types of image manipulation.
- The object removal operation affects the semantics of the image and is typically executed using the "copy-and-move" technique, where the pixels of the target object are overwritten with pixels sourced from another part of the image. Alternatively, it may involve completely removing and replacing the object's pixels with pixels from surrounding areas, such as the background, to achieve a visually coherent result.
- The image compositing operation combines elements from different sources to create a cohesive image. This process results in a new scene that does not exist in reality but is composed of actual objects that were previously captured. This technique is commonly used to manipulate a given shot, altering the context of the individuals or objects depicted within it.
- Painting is the process of creating graphic elements to replace portions of the original image. Similar to "copy-and-move" or "object removal" operations, the goal is to alter the semantics of the image. However, the critical difference is that the new elements are entirely generated from scratch and do not originate from any other pre-existing photograph.

The aforementioned manipulations rely on modifications to the original image. Detecting potential fraud can therefore involve comparing the transformed version with the original. As noted in the introduction, recent breakthroughs in deep learning have introduced revolutionary changes across various fields, opening new possibilities for data processing and generation. In the domain of image analysis, segmentation, detection, and generation algorithms have significantly advanced. Generative artificial intelligence, mainly through modern diffusion networks, can produce highly realistic images often based solely on textual descriptions (known as prompts). This presents a challenge for manipulation detection, as there may be no original image to compare against the generated image or its elements. Generated images are created based on patterns learned from millions of examples in generative network training datasets. As a result, they are not direct copies but entirely new compositions inspired by existing examples [9,14,20].

The purpose of this article is to review the most significant methods for ensuring image integrity, including both passive solutions (analyzing the properties of the image itself) and active solutions (based on, among other things, watermarks). The research methodology is based on an analysis of the available literature, considering both theoretical concepts and practical implementations, enabling an assessment of the effectiveness and limitations of various approaches. The authors' most significant contribution is identifying current trends in research on automated image manipulation detection using deep neural networks and discussing datasets used for training and testing modern algorithms.

The structure of this article is as follows. Section 2 defines key concepts and discusses the types of image manipulations most commonly employed. Section 3 describes the datasets used to validate the discussed solutions and includes a summary table outlining the key parameters of these datasets. Section 4 focuses on passive and active strategies,

presenting examples of deep learning applications. This section also includes a tabular comparison of methods to facilitate their evaluation. Section 5 provides a summary of this study, the main conclusions, and suggestions for future research directions in the field of digital image integrity protection.

### 3. Image Manipulation Datasets

The effectiveness of image manipulation detection methods largely depends on the quality and diversity of datasets used for training and testing models [21]. In recent years, several specialized datasets containing both original and manipulated images have been developed, enabling researchers to compare various approaches and detection techniques [22]. In this section, we provide an overview of the most important datasets used in the field of image manipulation detection, discussing their characteristics, applications, and availability.

CoMoFoD (Copy–Move Forgery Detection) [23] is a dataset developed by Tralić et al., in 2013, designed for studying methods of detecting copy–move manipulations in digital images. The dataset consists of two groups of images divided by size: small images with dimensions of  $512 \times 512$  pixels and large images with dimensions of  $3000 \times 2000$  pixels. In the small image category, there are 10,000 examples, of which 5000 are original images and 5000 are images containing forgeries created using the copy–move method. In the large image category, there are 3000 examples, split into 1500 originals and 1500 manipulated images.

The forgery creation process included various geometric transformations, such as translation, rotation, scaling, distortions, and their combinations, which enhances the diversity and realism of manipulations. Additionally, all images—both original and manipulated—underwent a series of post-processing operations. These include JPEG compression with different quality factors (ranging from 20 to 90), Gaussian blur with various parameters, noise addition using median filters of different sizes, brightness adjustments, color reduction, and contrast regulation. This extensive range of post-processing operations allows testing the robustness of detection algorithms against various distortions and transformations that may occur in practice.

The advantage of the CoMoFoD dataset is its diversity and complexity. The availability of ground truth masks enables the precise evaluation of manipulation detection methods at the pixel level. However, the disadvantage is the lack of information about the exact parameters of the applied geometric transformations.

CASIA [24] is one of the most significant public datasets used in research on detecting manipulations in digital images, particularly for operations such as splicing or copying parts of an image. This dataset was developed by the Chinese Academy of Sciences and is available in two versions: CASIA V1.0 and CASIA V2.0.

- CASIA V1.0 is the original version of the dataset, containing images of relatively low resolution. It includes 800 original images and 921 manipulated images. The uniform sizes of the images in this version facilitate processing and analysis.
- CASIA V2.0 is an extended version of the dataset, significantly larger and more diverse in terms of image resolution. It contains 7491 original images and 5123 manipulated images. The resolutions of the images in this version are varied—see Table 1.

Manipulations introduced in the CASIA dataset include various techniques such as splicing and copying and pasting elements within the same image. To increase the realism of the forgeries and make detection more challenging, additional post-processing operations were applied to the manipulated areas, such as blurring, contrast adjustment, and noise addition.

A disadvantage of the CASIA dataset is the lack of ground truth masks for manipulated images, which makes it difficult to accurately evaluate the effectiveness of detection

methods at the level of change localization. Researchers must primarily rely on classification metrics that assess whether an image is original or manipulated, without the capability to identify the areas of intervention precisely.

Despite this limitation, CASIA serves as a benchmark in numerous scientific studies.

MICC (Media Integration and Communication Center) [25] is one of the earliest and most renowned datasets used in research on detecting manipulations in digital images, particularly for methods involving copy–paste operations. Developed by a team from the University of Florence, this dataset consists of several subsets: MICC-F220, MICC-F2000, MICC-F600, and MICC-F8multi.

- MICC-F220 contains 220 high-resolution images, half of which are original and half manipulated using copy–paste techniques. These manipulations often include additional geometric transformations, such as rotation or scaling, increasing the difficulty of detection.
- MICC-F2000 is an extended dataset consisting of 2000 images, including 1300 original and 700 manipulated.
- MICC-F600 includes 600 high-resolution images featuring diverse scenes and objects. As in the previous subsets, the manipulations involve copying and pasting parts of the image, with 440 original images and 160 manipulated ones.
- MICC-F8multi is a small subset comprising eight images with multiple manipulations. This subset allows for testing algorithms in scenarios where a single image contains several independent forgeries.

A drawback of this dataset (similar to the CASIA dataset) is the lack of ground truth masks that would indicate the precise manipulated regions at the pixel level.

NIST16 [26] is a dataset created in 2016 by the National Institute of Standards and Technology (NIST) to support research on detecting manipulations in digital images. This dataset contains 3032 files, including 1422 test images (probe) in .jpg format, with the remainder consisting of reference images, ground truth masks, and metadata files. The dataset is characterized by high-resolution images, ranging from  $360 \times 480$  to  $4032 \times 3024$  pixels. The dataset is divided into three main manipulation categories: manipulation, removal, and splice. Each category represents a different type of forgery:

- Manipulation includes images with subtle changes, such as modifications to specific parts of the photo or alterations aimed at enhancing its appearance. This category includes 143 ground truth masks that indicate the edited areas at the pixel level.
- Removal contains images where certain elements have been removed. This group includes 101 ground truth masks.
- Splice involves images created by merging fragments from different sources, resulting in realistic but fabricated compositions. This category includes 146 ground truth masks.

In addition to manipulated images, the dataset includes 1244 original images, serving as a baseline for comparisons with the modified images. Metadata files are also available, specifying the transformations applied to the images, although they lack detailed transformation parameters.

COVERAGE [27] is a dataset designed for research on detecting copy–paste manipulations in digital images. Created by a team of researchers from Jiao Tong University in Shanghai, it includes high-quality images that enable testing the effectiveness of various forgery detection algorithms.

The dataset consists of 100 images, of which 50 are original, and the remaining 50 are manipulated using the copy–paste method. The manipulations involve a variety of geometric transformations, such as rotation, scaling, and perspective changes. Additionally,



the images include similar but authentic objects, adding complexity for algorithms that must distinguish genuine similarities from those resulting from manipulations.

The images in the dataset have a relatively high resolution of  $400 \times 486$  pixels. For each manipulated image, ground truth masks are available.

DEFACTO is a comprehensive dataset created to support research on detecting image and face manipulations, developed by Mahfoudi et al [28]. The DEFACTO dataset includes over 200,000 images, which were automatically generated to represent four main categories of image manipulation:

- **Splicing:** This involves inserting an external element from one image into another, creating a composite image. The dataset contains 105,000 images with this type of manipulation.
- **Copy-move (within the same image):** Elements within a single image are duplicated and placed elsewhere within the same image. To maintain realism, the position of copied objects is controlled along specific axes (vertical or horizontal) depending on the object's dimensions. Alpha matting techniques are used to refine the edges of the objects. The dataset contains 19,000 images with this type of manipulation.
- **Removing objects (inpainting):** This involves removing objects from images using inpainting techniques. This method fills empty areas by synthesizing the background based on surrounding pixels, allowing for smooth and natural image completion. Objects selected for removal are typically located against relatively uniform backgrounds. The dataset contains 19,000 images with this type of manipulation.
- **Morphing (face morphing):** Two images are deformed and merged to create a single image combining features of both sources. For facial images, this includes face blending and swapping. Facial landmarks are detected using the Dlib library [29], enabling the precise alignment and merging of faces. Additionally, color-matching techniques are applied to ensure consistent skin tone and lighting between the combined images. The dataset contains 80,000 images with this type of manipulation.

To generate these forgeries, the DEFACTO dataset uses the MSCOCO dataset [30] as a source of images and initial object annotations. However, the raw segmentation masks from MSCOCO are not precise enough for high-quality manipulations. Therefore, the authors applied alpha matting techniques to refine these masks, resulting in improved object edges and eliminating obvious manipulation artifacts, such as sharp edges or inconsistent lighting.

For each image in this dataset, a binary ground truth mask and metadata in the form of a JSON file describing the image transformation process are available.

The Columbia Image Splicing Detection Evaluation Dataset [31] is one of the first and most well-known datasets used in research on detecting splicing manipulations in digital images. Developed by the DVMM Laboratory at Columbia University, this dataset was created to support the development of passive forgery detection techniques that do not require additional information about the image or metadata. The dataset consists of 1845 images, all of a fixed size of  $128 \times 128$  pixels. These images were extracted from larger photographs available in the CalPhotos collection [32], as well as from a small number of photos taken with digital cameras. The dataset contains a comparable number of original and manipulated images. The manipulations primarily involve splicing, and the data are categorized based on specific characteristics:

- **Smooth vs. textured:** Images with uniform surfaces compared to those with complex textures.
- **Arbitrary object boundaries vs. simple boundaries:** Manipulations with irregular, complex edges compared to straight dividing lines.

One limitation of this dataset is the lack of pixel-level ground truth masks.

IMD2020 [33,34] is a comprehensive dataset developed by Novozamsky and collaborators. It consists of several subsets encompassing diverse types of manipulations and a wide range of authentic images, enabling thorough testing of detection methods under realistic conditions.

The first component of the dataset is IMD2020 Real-Life Manipulated Images, containing 2010 forged images sourced from the Internet. Each manipulated image has a corresponding original version, allowing precise comparison and analysis of the manipulations. Additionally, manually created binary masks are provided for each forged image, localizing the manipulated areas at the pixel level. Another component is the IMD2020 Large-Scale Set of Real Images, comprising 35,000 authentic (non-manipulated) images collected from 2322 camera models. All images were manually reviewed, and those showing visible traces of digital manipulation were excluded.

The IMD2020 Large-Scale Set of Inpainting Images includes 35,000 manipulated images. The manipulations in this subset involve the use of inpainting techniques, where regions are randomly selected and filled using the method described by Jiahui Yu and collaborators [34]. The IMD2020 Guaranteed Set of Images contains 2759 authentic images from 32 unique cameras (19 models). This subset is specifically designed to analyze sensor noise (PRNU) and other features related to image authenticity. The diversity of cameras used allows for the study of the impact of different sensors on manipulation detection.

Below, we present Table 1, aggregating data on all the mentioned datasets.

**Table 1.** Summary of popular image manipulation datasets.

Dataset Name	Nb of Samples (Authentic; Manipulated)	Resolutions [px]	Types of Manipulations	GT Mask Available	Metadata Available
CoMoFoD	small images: 5000; 5000 large images: 1500; 1500	from $512 \times 512$ to $3000 \times 2000$	copy-move; geometric transformations; post-processing;	yes	no
CASIA V1.0	800; 921	$384 \times 256$	splicing; copy-move; post-processing;	no	no
CASIA V2.0	7491; 5123	from $240 \times 160$ to $900 \times 600$	splicing; copy-move; post-processing;	no	no
MICC-F220	110; 110	from $722 \times 480$ to $800 \times 600$	copy-move; geometric transformations;	no	no
MICC-F2000	1300; 700	$2048 \times 1536$	copy-move; geometric transformations;	no	no
MICC-F8multi	N/A; 8	from $800 \times 532$ to $2048 \times 1536$	multiple copy-move manipulations;	no	no
NIST16	1244; 390	from $360 \times 480$ to $4032 \times 3024$	removal; splicing;	yes	yes (without transformation parameters)
COVERAGE	50; 50	$400 \times 486$	copy-move; geometric transformations;	yes	not specified
DEFACTO	over 223,000; over 223,000	not specified (but based on MSCOCO)	splicing; copy-move; inpainting; face morphing;	yes	yes
Columbia	933; 912	$128 \times 128$	splicing;	no	no
IMD2020	39,769; 37,010	from $291 \times 1024$ to $1024 \times 930$	splicing; inpainting	yes	not specified

## 4. Methods for Manipulation Protection and Detection

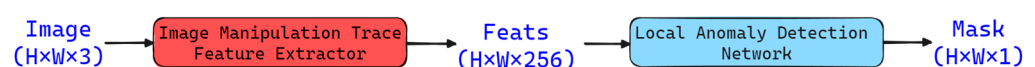
Manipulation detection methods not only enable the identification of tampering in an image but often also pinpoint the specific areas subjected to modification [35,36]. Some advanced techniques can even reconstruct an approximate version of the original image prior to the manipulation [37]. In the years preceding the deep learning revolution, classical image processing algorithms were used to detect image manipulation [18,38,39]. However, these methods faced significant limitations in terms of effectiveness and generalizability. Today, the advancements in deep learning dominate both current research and practical applications in manipulation detection [19,40]. Methods based on neural networks allow for more efficient and accurate detection of modifications, even in complex scenarios.

Active methods require embedding additional information, such as digital watermarks or signatures, into the image. Digital watermarks involve hiding imperceptible data within the image, which can serve as redundant information for identifying manipulated areas or even partially reconstructing the original. Additionally, watermarks may include metadata such as author details, source information, rights data, or unique identifiers. In cases of suspected manipulation, this embedded information can be extracted and used to verify the authenticity of the image. Digital signatures, on the other hand, use cryptographic encryption techniques to generate unique identifiers for the image, similar to solutions used in telecommunication systems [41]. Any alteration to the image changes the signature, making unauthorized modifications easier to detect. Active methods are particularly effective when the entire process of image creation and distribution is under control. Passive methods, also known as blind methods, rely solely on analyzing the image itself without requiring any additional data or prior embedded information. These methods leverage the statistical and structural characteristics of the image, such as color histograms, textures, noise patterns, or compression artifacts. By analyzing these features, passive methods can detect inconsistencies and anomalies indicative of manipulation. For instance, they can identify discrepancies in lighting, shadows, or perspective, which are challenging to replicate accurately during editing. Thanks to deep learning techniques, passive methods have gained the ability to detect manipulations by extracting highly abstract and complex image features [42]. Neural networks with a large number of parameters can learn data representations at multiple levels, identifying subtle anomalies and patterns that are invisible to traditional methods.

In the following subsections of this section, we will focus on the analysis of both active and passive image integrity assurance methods based on deep learning.

### 4.1. ManTra-Net

One of the significant approaches within passive methods is the concept presented in the publication ManTra-Net [43], which introduced one of the first end-to-end solutions based on fully convolutional networks. The authors designed the architecture by dividing it into two main components: the first, called the “image manipulation trace feature extractor”, which generates a unified feature representation, and the second, the “local anomaly detection network” (LADN), which focuses on anomaly localization. This structure directly detects tampered areas without requiring further post-processing—see Figure 2.



**Figure 2.** ManTra-Net high level architecture.

According to the information provided by the authors, the developed method can distinguish up to 385 types of known manipulations, including operations generated by other neural networks, such as image inpainting. For the backbone of the architecture, three



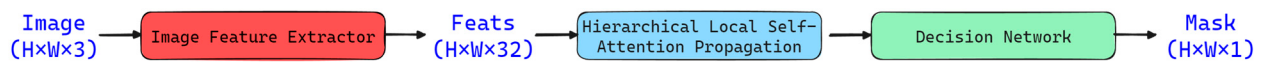
types of networks were tested as follows: VGG [44], ResNet [42], and DnCNN [45]. The backbones were evaluated on the “IMC-7” (Image Manipulation Classification) task, which includes attacks such as compression, blurring, morphological manipulations, contrast adjustments, noise addition, resampling, and quantization. The VGG network achieved the best results, with 92.1% accuracy on the validation set, while ResNet reached 90.8% and DnCNN 91.2%. Additionally, studies were conducted to assess the impact of the type of first network layer, comparing SRMConv2D, BayarConv2D, and a classic 2D convolutional layer. The results for the different layer types were similar, with differences of approximately 1%. The authors incorporated a Conv-LSTM module into the LADN component to effectively model spatial dependencies between feature vectors. This module processes data sequentially, line by line, across the image, allowing it to account for local context during anomaly detection. As a result, the network can more effectively identify subtle irregularities in the image structure, such as barely noticeable changes in texture or color, which may indicate manipulations. Another key element of LADN is the Z-score function, used to normalize features and highlight deviations from the norm. Z-score enables the identification of pixels or regions in the image that significantly differ from their surroundings, a hallmark of manipulations. After being processed by Conv-LSTM, the data are passed through additional convolutional layers, which further refine and integrate spatial information, enhancing the network’s ability to distinguish complex manipulation patterns. During the training process of LADN, a loss function based on the difference between the predicted and actual localization masks was used, enabling the network to learn precise anomaly detection. The results presented by the authors demonstrate that ManTra-Net, leveraging LADN with the Conv-LSTM module, achieves high effectiveness in detecting and localizing forgeries across various datasets. The network not only handles individual types of manipulations but also copes with complex combinations of manipulations and previously unseen types of forgeries.

#### 4.2. SPAN

SPAN [46] (Spatial Pyramid Attention Network) is an architecture designed to enhance manipulation localization by effectively capturing dependencies between points and objects at various scales, utilizing self-attention mechanisms [47–49].

SPAN offers an improved interpretation of spatial relationships within an image, resulting in more precise manipulation detection compared to ManTra-Net. While ManTra-Net primarily focuses on analyzing the same image points across different scales of the feature map, it does not account for spatial dependencies between different regions of the image. SPAN builds upon this approach by modeling both inter-scale relationships through multi-scale information propagation and spatial dependencies between image regions using a self-attention module. This capability allows SPAN to capture complex patterns and relationships within the image more effectively, leading to more accurate manipulation localization. The SPAN architecture consists of three main components: a feature extractor, a spatial pyramid attention block, and a decision module—see Figure 3. The self-attention module aims to establish relationships between different parts of the image, enabling the network to capture both local and global context. The first stage of processing in the described architecture is the feature extractor. The authors utilized a pre-trained network from ManTra-Net, trained on synthetic data based on images from the Dresden Image Database [50]. This training employed supervised classification, covering 385 types of image manipulations. To streamline subsequent processing, the output from the feature extractor is passed through a  $1 \times 1$  convolutional layer, reducing the feature depth from 256 (as in ManTra-Net) to 32. This enables faster data propagation through the self-attention modules, which are organized into five cascaded layers analyzing features

progressively from the most local to the most global. Each step employs a self-attention module with an increasing dilation parameter (1, 3, 9, 27, 81) and a convolutional layer to normalize the output channels back to 32.



**Figure 3.** SPAN high level architecture.

This approach efficiently captures dependencies at various scales, enhancing the accuracy of manipulation localization within the image. In machine translation models, positional embeddings—learned weights added to inputs at each position—are commonly used and are effective for linguistic tasks. However, this approach is suboptimal for image processing due to differences in data characteristics. Instead, the authors of this model introduced learned matrix projections to represent all possible relative spatial relationships in a pixel’s neighborhood. In the local self-attention block, each pixel considers information from its neighbors through these projections.

The final component of the processing pipeline, the decision module, is a relatively simple convolutional network consisting of several layers. It analyzes the output of the preceding self-attention module and concludes with a sigmoid activation function [51]. This approach achieved results that were 11.2% better than ManTra-Net under comparable testing conditions.

#### 4.3. Asnani et al. [35]

One of the newer approaches in the field of image manipulation detection is the method proposed by Asnani and collaborators [35]. The authors introduce an innovative solution that involves embedding pre-learned templates into images to facilitate subsequent manipulation detection. This method stands out for its active nature, as it does not passively analyze images for manipulations but instead embeds visually imperceptible templates into the original images. As a result, any manipulation performed by generative models alters the embedded templates in detectable ways, enabling the effective identification of changes—see Figure 4.



**Figure 4.** Asnani et al. [35] high level architecture.

The process of embedding templates into an image can be defined as follows: for each image from the set of real images  $X_a$  a template  $S_i$  is randomly selected from a pre-prepared set of templates  $S = \{S_1, S_2, \dots, S_n\}$ . This template is then added to the image in a controlled manner, regulated by a parameter  $m$ , which determines the strength of the template addition while ensuring that the changes remain imperceptible to the human eye.

This process can be formalized as

$$T(X_a^j) = X_a^j + m \times S_i \quad (1)$$

where:

$X_a^j$ —original image;

$T(X_a^j)$ —modified image;

$m$ —parameter controlling the intensity of the template embedding.

The template recovery module is used for manipulation detection by comparing the recovered template with the original one. This process employs an encoder network  $E$ , which is trained to recover the embedded templates from images that may have been manipulated by a generative model  $G$ .

Thus, for each image, the process can be described as follows:

$$S_R = E\left(T\left(X_a^j\right)\right) \quad (2)$$

and

$$S_F = E\left(G\left(T\left(X_a^j\right)\right)\right) \quad (3)$$

If the image has not been manipulated, the recovered template  $S_R$  will be very similar to the original template  $S_i$ . In the case of manipulation, the recovered template  $S_F$  will differ significantly from the original  $S_i$ . The recovered template is compared with the original template using a classification network, which performs a binary evaluation to determine whether the image has been manipulated or not.

The encoder network begins with two convolutional layers that perform initial feature extraction from the image. Each of these layers uses  $3 \times 3$  convolutions with a stride of one, processing the three-channel input image (RGB) and generating 64 feature maps as the output, utilizing the ReLU activation function. Following this, the network proceeds through ten convolutional blocks, where each convolutional layer is supported by batch normalization and the ReLU activation function. In the end, an additional  $3 \times 3$  convolutional layer reduces the number of output channels to one, allowing for the generation of the final template recovered from the image.

The architecture of the classifier network is very similar to that of the encoder, with the primary differences being in the final portion and the number of convolutional blocks, reduced to eight instead of ten. Unlike the encoder, the classifier concludes with a sequence of fully connected layers that transform the extracted features into a classification decision. The fully connected layers contain 512, 256, and 1 neurons, respectively, enabling the assignment of the image to either the “original” or “manipulated” class.

#### 4.4. ObjectFormer

An intriguing method for detecting image manipulation within the realm of passive approaches is ObjectFormer [52], proposed by a team of researchers from Fudan University, the University of Maryland, Huya Inc., and Meta AI. ObjectFormer stands out for its multimodal approach to manipulation detection, which enables analysis in both the RGB and frequency domains. This dual-domain capability allows it to capture invisible traces of editing in the standard color spectrum. The architecture of this method is based on transformers [32,33], which model spatial dependencies within the image (Figure 5). By leveraging the strengths of both RGB and frequency representations, ObjectFormer enhances its ability to detect subtle and sophisticated manipulation artifacts.

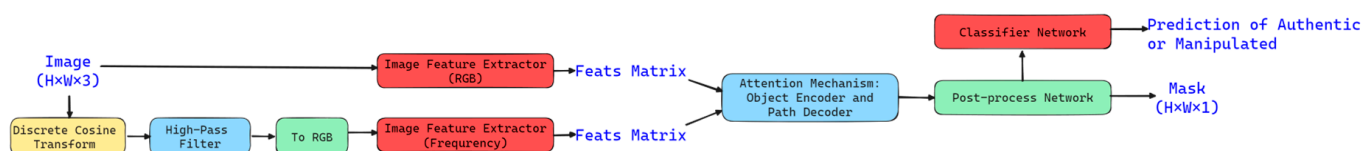


Figure 5. ObjectFormer high level architecture.

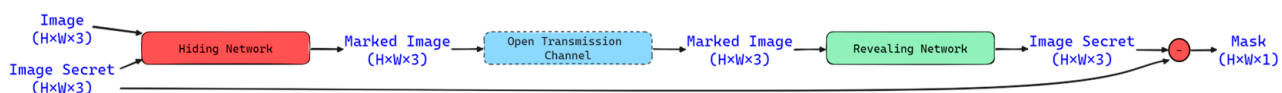
Many image manipulation detection methods classify an image as either authentic or manipulated, often without considering the structure of objects within the image. However, the effective detection of visual modifications requires not only identifying pixel-level

anomalies but also evaluating whether entire objects in the image are consistent with each other. The input RGB image is transformed into the frequency domain using the Discrete Cosine Transform (DCT). A high-pass filter is then applied, retaining only the components that may potentially contain information about image tampering (low frequencies are removed). After this process, the image is converted back to the RGB domain. The architecture begins with two parallel processing paths: the spatial (color) and frequency domains. For data representation extraction, two feature extractors based on EfficientNet-b4 [53] (pre-trained on ImageNet [54]) are employed.

Both representations—derived from the RGB and frequency domains—are then divided into spatial embeddings, which are vectors combined to form multimodal embeddings. In the next stage, the object encoder utilizes an attention mechanism and learnable object prototypes to analyze the consistency between elements in the image. These prototypes enable the model to identify various objects in the image and examine their interrelationships. Subsequently, the patch decoder leverages this information to refine the patch embeddings, enriching them with knowledge about the objects. The applied BCIM module (Boundary-sensitive Contextual Incoherence Modeling) is designed to detect contextual inconsistencies at the pixel level. This module improves the precision of boundary detection in manipulated regions, enhancing their sharpness by analyzing local pixel similarities and incorporating this information into the generated feature representations. The final stage includes a post-processing network consisting of interpolation and convolution operations aimed at upsampling the matrix to match the input image dimensions and producing a single-channel output to create a binary manipulation detection mask. The extracted image features are also passed through a simple fully connected layer [55], which classifies the image as either authentic or manipulated.

#### 4.5. Zhao et al. [36]

According to the authors, one of the most intriguing recently proposed approaches to active image integrity assurance is the study “Proactive Image Manipulation Detection via Deep Semi-Fragile Watermark” by Zhao et al [36]. This method is based on the concept of a transparent watermark embedded in the image in the form of a mask, which can later be extracted to detect potential modifications—see Figure 6. A key element of the solution involves two neural network models responsible for generating and recovering the watermark. The watermark itself is another image, transformed into a feature matrix and added to the original image. The algorithm’s concept resembles data-hiding techniques such as “image-in-image” [56,57]. However, the decoding component of the hidden image has been modified so that any changes in the watermark indicate the regions of the image that have been manipulated.



**Figure 6.** Zhao et al. [36] high level architecture.

The authors proposed using the UNet architecture [58] for both models in their solution. These networks are trained in a supervised process utilizing two loss functions: hiding loss, which focuses on maintaining the high transparency of the watermark, and revealing loss, which ensures the accuracy of extracting the hidden image while accounting for manipulations introduced during training. The manipulation process is simulated by an attack module consisting of two groups of operations. The first group includes natural processing operations commonly occurring in transmission pipelines, such as JPEG compression [59] and Gaussian blur. Artifacts introduced by these operations should

not be interpreted by the decoding network as distortions and must be ignored during the detection of manipulated regions. The second group of the attack module simulates intentional manipulations of selected image regions, such as inpainting. For each such manipulation, a binary mask is generated to indicate the attack region, which assists in validating the detection results by the decoding component.

This algorithm represents an advancement over previous methods because it avoids the inclusion of specific manipulations in the training pipeline, which could potentially limit the ability to generalize across various sabotage operations. Instead, it randomly selects and modifies a region within the container image to simulate the effect of manipulation, using this modification as the ground truth label during training. The output of the attack module is therefore a manipulated image and a binary mask indicating the altered regions. The task of the decoding network is to recover the hidden image from the carrier image (manipulated or unmanipulated). The original hidden image is then subtracted from the recovered image to extract the manipulated regions—assuming correct decoder performance, the manipulated areas should have values significantly different from zero.

#### 4.6. MSCL-Net

The MSCL-Net method [19] is a passive approach for detecting and localizing image manipulations, based on common structures in such algorithms, including an encoder–decoder architecture (for multi-level analysis), feature fusion, and attention mechanisms. The standard approach in these algorithms relies on using color features (RGB) along with additional representations, such as noise features or compression artifacts, to identify manipulations. In this context, the creators of MSCL-Net proposed the Feature Cross Fusion Module (FCFM), which facilitates deeper integration of RGB and Spatial Rich Model (SRM) features through a cross-fusion mechanism, rather than the simple concatenation or addition used in other methods. The Spatial Rich Model (SRM) is characterized by its ability to capture local noise features at the pixel neighborhood level, enabling the detection of noise inconsistencies between altered and untouched regions while minimizing the influence of the image’s content. Cross-fusion, on the other hand, involves dynamic interaction between RGB and SRM features, allowing for a more detailed and balanced representation of manipulation information. This approach neutralizes semantic content and amplifies forgery traces, resulting in improved detection and localization of manipulations.

The authors also introduced the Adaptive Self-Attention Module (ASAM) [60], which analyses dependencies in both spatial and channel dimensions, unlike traditional attention mechanisms that operate in a single dimension. This enables the model to identify better global and local differences between altered and untouched regions, enhancing the precision of manipulation detection. Additionally, the method incorporates a Supervised Contrastive Learning Module (SCLM), which employs multi-scale contrastive learning. This approach maximizes the differences between the features of altered and untouched regions at various representation levels, improving accuracy and reducing false positives. The loss function in MSCL-Net goes beyond standard approaches, which typically focus on classification and segmentation losses. The authors introduced an additional component in the form of contrastive losses calculated at four different scales. This allows for a more comprehensive analysis of differences, further improving the model’s ability to detect and localize manipulations.

Using ConvNeXt-T [61] as the encoder, instead of the popular ResNets, allows for a larger receptive field and greater sensitivity to local traces of manipulation. This is achieved by using larger convolutional kernels, which cover more extensive areas of the image in a single step, and replacing traditional max-pooling operations with downsampling convolu-



tions. As a result, the network is more sensitive to local traces of manipulation, as it better preserves critical details related to differences between altered and untouched regions.

#### 4.7. Summary

The summary of the methods discussed in this section is presented below—Tables 2 and 3.

**Table 2.** Summary of key features of the discussed methods.

Method Name	Publication Year	Method Type	Output Type	Key Features	Datasets Used in Evaluation
ManTra-Net [43]	2019	Passive	Localization	End-to-end CNN; detects 385 types of manipulations	NIST16, CASIA, CoMoFoD
SPAN [46]	2020	Passive	Localization	Self-attention mechanisms; better spatial relation interpretation	NIST16, CASIA
Asnani et al. [35]	2022	Active	Detection	Adding pre-trained templates to images	Own dataset
ObjectFormer [52]	2022	Passive	Detection and Localization	Multimodal approach; architecture based on transformers	CASIA, Coverage, NIST16
Zhao et al. [36]	2023	Active	Detection	Deep semi-fragile watermark; use of UNet architecture	Own dataset
MSCL-Net [19]	2024	Passive	Detection and Localization	Multi-scale contrastive learning; attention mechanisms	CASIA, Coverage, NIST16

We are witnessing the rapid development of image manipulation detection methods based on deep learning, driven by increasing threats associated with the ease of editing and generating multimedia content using advanced generative models [9,14,20]. Current deep learning methods [62,63], such as ManTra-Net, SPAN, and MSCL-Net, demonstrate significant improvements in manipulation detection accuracy, achieving progressively higher metrics such as F1 Score and AUC on standard datasets. The importance of statistical analysis and frequency-domain features in detecting inconsistencies in images is emphasized [64,65]. Trends indicate a growing interest in passive methods, which, by leveraging the latest advancements in deep learning such as self-attention mechanisms, contrastive learning, and hybrid architectures [66], effectively identify subtle inconsistencies in image structures.

Active methods, such as those proposed by Asnani et al. [35] and Zhao et al. [36], are gaining importance, despite requiring additional effort during the image generation process. Techniques based on watermarking (or other invisible markers) significantly enhance effectiveness in both detecting modified areas and overall manipulation detection.

Furthermore, watermarks allow for the transmission of a small data payload, which can include information about the source, authenticity, or context of the image, providing an additional layer of protection. Solutions in this category offer potential benefits for applications demanding a high level of data integrity.

When comparing both strategies, passive methods are more versatile as they do not require prior embedding of data into the image. However, they may be less effective against advanced manipulations generated by diffusion models or GANs. On the other hand, active methods offer higher security, but their effectiveness depends on controlling the image creation and distribution process. Moving forward, the focus will be on developing models capable of generalizing to new types of manipulation algorithms and resistant to attacks by

generative models, where no original image is available for comparison. Researchers will emphasize creating more advanced neural network architectures that leverage emerging innovations in deep learning and multimodal analysis to counter increasingly sophisticated manipulation techniques effectively.

**Table 3.** Performance metrics of methods on various datasets.

Method Name	Dataset	Metric	Value	Comments
ManTra-Net [43]	CASIA	F1 Score	-	used CASIA v1.0
		AUC	81.7%	
	NIST16	F1 Score	-	
		AUC	79.5%	
	COVERAGE	F1 Score	-	
		AUC	81.9%	
SPAN [46]	CASIA	F1 Score	38.2%	some data from the NIST16 test set were also present in the training set
		AUC	79.72%	
	NIST16	F1 Score	58.2%	
		AUC	83.95%	
	COVERAGE	F1 Score	55.8%	
		AUC	92.22%	
ObjectFormer [52]	CASIA	F1 Score	57.9%	-
		AUC	88.2%	
	NIST16	F1 Score	82.4%	
		AUC	99.6%	
	COVERAGE	F1 Score	75.8%	
		AUC	95.7%	
MSCL-Net [19]	CASIA	F1 Score	85.9%	used both CASIA v1 and CASIA v2
		AUC	86.2%	
	NIST16	F1 Score	82.1%	
		AUC	84.6%	
	COVERAGE	F1 Score	71.9%	
		AUC	80.4%	
Zhao et al. [36]	MS-COCO	PSNR	38.05 dB	to obtain an unambiguous F1 metric and precision, the results were averaged for three types of attacks: copy-move, inpainting, and splice.
		SSIM	0.94	
	CelebaHQ	F1 Score	78.8%	
		Precision	80.1%	

## 5. Conclusions

This paper presents the most significant methods published in recent years for verifying the integrity of digital images in the context of increasingly advanced manipulation techniques. Both passive and active approaches were reviewed to achieve this goal, highlighting their main advantages, limitations, and practical applications. This paper discusses the development trends in neural network structures and architecture, ranging from traditional convolutional models through attention modules to innovative solutions based on transparent watermarking. This provides a better estimation of trends in forgery detection, aligning with this study's primary objectives. A notable contribution is the detailed presentation of the most popular datasets, aiding researchers in quickly selecting the optimal set for their needs. Additionally, the limitations of individual image collections and the types of data they contain are outlined.

One of the current challenges is creating a reliable testing environment that enables accurate and meaningful performance measurement of anomaly detection algorithms under conditions resembling real-world scenarios. Although datasets such as DEFACTO are available, their scale, data quality, and annotations do not always keep pace with contemporary, sophisticated image manipulation techniques. Therefore, collaboration

between institutions responsible for creating multimedia content is crucial to developing further high-quality reference datasets for evaluating forgery detection systems.

Another somewhat overlooked aspect is the model's efficiency and the conditions under which they are deployed. For example, many publications lack references to processing time and hardware requirements, which can hinder the practical implementation of a method. Additionally, some models operate at low resolutions (e.g., 200 px), and the downsampling process can result in significant information loss. Equally important is the analysis of video materials, which must account for spatiotemporal aspects and annotations' high costs and complexity.

In the future, special emphasis should be placed on designing solutions resilient to automated manipulations generated by advanced AI models and integrating various approaches into cohesive detection systems—tailored to specific needs and use cases. Ensuring the availability of these technologies on mobile platforms and embedded systems is crucial for widespread adoption and protecting the integrity of information. Another key issue is the efficient processing of high-resolution images and video materials, which are increasingly subject to manipulation.

Ultimately, the effective fight against visual forgeries requires advanced technical solutions, raising public awareness of potential threats, and promoting education to foster critical multimedia content analysis.

**Author Contributions:** Conceptualization, P.D., T.W. and Z.P.; funding acquisition, Z.P.; methodology, P.D., T.W. and Z.P.; project administration, Z.P.; software, P.D. and T.W.; supervision, Z.P.; visualization, P.D. and T.W.; writing—original draft, P.D., T.W. and Z.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Military University of Technology, Faculty of Electronics, grant number UGB/22-747, on the application of artificial intelligence methods to cognitive spectral analysis, satellite communications, and watermarking and technology deepfakes.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Das, S. The evolution of visual effects in cinema: A journey from practical effects to CGI. *J. Emerg. Technol. Innov. Res.* **2023**, *10*, e303–e309.
2. Pierson, M. CGI Effects in Hollywood Science-Fiction Cinema 1989–95: The Wonder Years. *Screen* **1999**, *40*, 158–176. [CrossRef]
3. King, D. *The Commissar Vanishes/Anglais: The Falsification of Photographs and Art in Stalin's Russia*; Tate: London, UK, 2014; ISBN 978-1-84976-251-9.
4. Kraków, I.P.N. Operacja polska NKWD 1937–1938. Available online: <https://krakow.ipn.gov.pl/pl4/aktualnosci/56290,Operacja-polska-NKWD-1937-1938.html> (accessed on 28 December 2024).
5. Cubitt, S. *The Cinema Effect*; MIT Press: Cambridge, MA, USA, 2005; ISBN 978-0-262-53277-8.
6. Keating, P. Out of the Shadows. In *A Companion to Film Noir*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2013; pp. 265–283, ISBN 978-1-118-52372-8.
7. Braham, R. The Digital Backlot [Cinema Special Effects]. *IEEE Spectr.* **1995**, *32*, 51–63. [CrossRef]
8. Agbaji, D.; Lund, B.; Mannuru, N.R. Perceptions of the Fourth Industrial Revolution and Artificial Intelligence Impact on Society. *arXiv* **2023**, arXiv:2308.02030.
9. Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education | Cyberpsychology, Behavior, and Social Networking. Available online: <https://www.liebertpub.com/doi/abs/10.1089/cyber.2020.0174> (accessed on 3 November 2024).
10. Digital Image Forensics: A Booklet for Beginners | Multimedia Tools and Applications. Available online: <https://link.springer.com/article/10.1007/s11042-010-0620-1> (accessed on 5 November 2024).
11. Zheng, L.; Zhang, Y.; Thing, V.L.L. A Survey on Image Tampering and Its Detection in Real-World Photos. *J. Vis. Commun. Image Represent.* **2019**, *58*, 380–399. [CrossRef]
12. Capasso, P.; Cattaneo, G.; De Marsico, M. A Comprehensive Survey on Methods for Image Integrity. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 347:1–347:34. [CrossRef]

13. Zanardelli, M.; Guerrini, F.; Leonardi, R.; Adami, N. Image Forgery Detection: A Survey of Recent Deep-Learning Approaches. *Multimed. Tools Appl.* **2023**, *82*, 17521–17566. [\[CrossRef\]](#)
14. Langguth, J.; Pogorelov, K.; Brenner, S.; Filkuková, P.; Schroeder, D.T. Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes. *Front. Commun.* **2021**, *6*, 632317. [\[CrossRef\]](#)
15. Fridrich, J.; Kodovsky, J. Rich Models for Steganalysis of Digital Images. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 868–882. [\[CrossRef\]](#)
16. Zhu, J.; Kaplan, R.; Johnson, J.; Fei-Fei, L. HiDDeN: Hiding Data with Deep Networks. *arXiv* **2018**, arXiv:1807.09937.
17. Swain, G.; Lenka, S.K. LSB Array Based Image Steganography Technique by Exploring the Four Least Significant Bits. In Proceedings of the Global Trends in Information Systems and Software Applications, Vellore, India, 9–11 December 2011; Krishna, P.V., Babu, M.R., Ariwa, E., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 479–488.
18. Piva, A. An Overview on Image Forensics. *Int. Sch. Res. Not.* **2013**, *2013*, 496701. [\[CrossRef\]](#)
19. Bai, R. Image Manipulation Detection and Localization Using Multi-Scale Contrastive Learning. *Appl. Soft Comput.* **2024**, *163*, 111914. [\[CrossRef\]](#)
20. Dang, M.; Nguyen, T.N. Digital Face Manipulation Creation and Detection: A Systematic Review. *Electronics* **2023**, *12*, 3407. [\[CrossRef\]](#)
21. Ali Qureshi, M.; Deriche, M. A Review on Copy Move Image Forgery Detection Techniques. In Proceedings of the 2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14), Castelldefels, Spain, 11–14 February 2014; pp. 1–5.
22. Zhou, P.; Han, X.; Morariu, V.I.; Davis, L.S. Learning Rich Features for Image Manipulation Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
23. Tralic, D.; Zupancic, I.; Grgic, S.; Grgic, M. CoMoFoD—New Database for Copy-Move Forgery Detection. In Proceedings of the Proceedings ELMAR-2013, Zadar, Croatia, 25–27 September 2013; pp. 49–54.
24. Dong, J.; Wang, W.; Tan, T. Casia Image Tampering Detection Evaluation Database. In Proceedings of the 2013 IEEE China Summit and International Conference on Signal and Information Processing, Beijing, China, 6–10 July 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 422–426.
25. Copy-Move Forgery Detection and Localization. Available online: <http://lci.micc.unifi.it/labd/2015/01/copy-move-forgery-detection-and-localization/> (accessed on 19 November 2024).
26. Guan, H.; Kozak, M.; Robertson, E.; Lee, Y.; Yates, A.N.; Delgado, A.; Zhou, D.; Kheyrkhah, T.; Smith, J.; Fiscus, J. MFC Datasets: Large-Scale Benchmark Datasets for Media Forensic Challenge Evaluation. In Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 63–72.
27. Wen, B.; Zhu, Y.; Subramanian, R.; Ng, T.T.; Shen, X.; Winkler, S. COVERAGE—A Novel Database for Copy-Move Forgery Detection. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016.
28. Mahfoudi, G.; Tajini, B.; Retraint, F.; Morain-Nicolier, F.; Dugelay, J.L.; Pic, M. DEFACTO: Image and Face Manipulation Dataset. In Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO), A Coruña, Spain, 2–6 September 2019; pp. 1–5.
29. King, D.E. Dlib-Ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
30. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2014; Volume 8693, pp. 740–755, ISBN 978-3-319-10601-4.
31. Columbia Image Splicing Detection Evaluation Dataset. Available online: <https://www.ee.columbia.edu/in/dvmm/downloads/AuthSplicedDataSet/AuthSplicedDataSet.htm> (accessed on 19 November 2024).
32. CalPhotos. Available online: <https://calphotos.berkeley.edu/> (accessed on 19 November 2024).
33. Novozamsky, A.; Mahdian, B.; Saic, S. IMD2020: A Large-Scale Annotated Dataset Tailored for Detecting Manipulated Images. In Proceedings of the 2020 IEEE Winter Applications of Computer Vision Workshops (WACVW), Snowmass Village, CO, USA, 1–5 March 2020; IEEE: Snowmass Village, CO, USA, 2020; pp. 71–80.
34. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative Image Inpainting with Contextual Attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
35. Asnani, V.; Yin, X.; Hassner, T.; Liu, S.; Liu, X. Proactive Image Manipulation Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
36. Zhao, Y.; Liu, B.; Zhu, T.; Ding, M.; Yu, X.; Zhou, W. Proactive Image Manipulation Detection via Deep Semi-Fragile Watermark. *Neurocomputing* **2024**, *585*, 127593. [\[CrossRef\]](#)
37. A Novel Color Image Tampering Detection and Self-Recovery Based on Fragile Watermarking. *J. Inf. Secur. Appl.* **2023**, *78*, 103619. [\[CrossRef\]](#)

38. Triaridis, K.; Tsigos, K.; Mezaris, V. MMFusion: Combining Image Forensic Filters for Visual Manipulation Detection and Localization. *arXiv* **2024**, arXiv:2312.01790.
39. Khan, E.S.; Kulkarni, E.A. An Efficient Method for Detection of Copy-Move Forgery Using Discrete Wavelet Transform. *Int. J. Comput. Sci. Eng.* **2010**, *2*, 2010.
40. VidalMata, R.G.; Saboia, P.; Moreira, D.; Jensen, G.; Schlessman, J.; Scheirer, W.J. On the Effectiveness of Image Manipulation Detection in the Age of Social Media. *arXiv* **2023**, arXiv:2304.09414.
41. Nawaal, B.; Haider, U.; Khan, I.; Fayaz, M. Signature-Based Intrusion Detection System for IoT. In *Cyber Security for Next-Generation Computing Technologies*; CRC Press: Boca Raton, FL, USA, 2023; pp. 141–158, ISBN 978-1-00-340436-1.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
43. Wu, Y.; AbdAlmageed, W.; Natarajan, P. ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries with Anomalous Features. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Long Beach, CA, USA, 2019; pp. 9535–9544.
44. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
45. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Trans. Image Process.* **2016**, *26*, 3142–3155. [[CrossRef](#)]
46. Hu, X.; Zhang, Z.; Jiang, Z.; Chaudhuri, S.; Yang, Z.; Nevatia, R. SPAN: Spatial Pyramid Attention Network for Image Manipulation Localization. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12366, pp. 312–328.
47. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information processing Systems*; MIT Press: Cambridge, MA, USA, 2017; pp. 5998–6008.
48. Alexey, D.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
49. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:1409.0473.
50. Gloe, T.; Böhme, R. The Dresden Image Database for Benchmarking Digital Image Forensics. *J. Digit. Forensic Pract.* **2010**, *3*, 150–159. [[CrossRef](#)]
51. Narayan, S. The Generalized Sigmoid Activation Function: Competitive Supervised Learning. *Inf. Sci.* **1997**, *99*, 69–82. [[CrossRef](#)]
52. Wang, J.; Wu, Z.; Chen, J.; Han, X.; Shrivastava, A.; Lim, S.-N.; Jiang, Y.-G. ObjectFormer for Image Manipulation Detection and Localization. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: New Orleans, LA, USA, 2022; pp. 2354–2363.
53. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
54. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
55. Basha, S.H.S.; Dubey, S.R.; Pulabaigari, V.; Mukherjee, S. Impact of Fully Connected Layers on Performance of Convolutional Neural Networks for Image Classification. *Neurocomputing* **2020**, *378*, 112–119. [[CrossRef](#)]
56. Das, A.; Wahi, J.S.; Anand, M.; Rana, Y. Multi-Image Steganography Using Deep Neural Networks. *arXiv* **2021**, arXiv:2101.00350.
57. Duan, X.; Liu, N.; Gou, M.; Wang, W.; Qin, C. SteganoCNN: Image Steganography with Generalization Ability Based on Convolutional Neural Network. *Entropy* **2020**, *22*, 1140. [[CrossRef](#)] [[PubMed](#)]
58. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer: Berlin/Heidelberg, Germany, 2015.
59. Zhang, C.; Karjauv, A.; Benz, P.; Kweon, I.S. Towards Robust Data Hiding Against (JPEG) Compression: A Pseudo-Differentiable Deep Learning Approach. *arXiv* **2020**, arXiv:2101.00973.
60. Liu, J.; Wei, Z.; Li, Z.; Mao, X.; Wang, J.; Wei, Z.; Zhang, Q. SAM: A Self-Adaptive Attention Module for Context-Aware Recommendation System. In Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024.
61. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
62. Cozzolino, D.; Verdoliva, L. Single-Image Splicing Localization through Autoencoder-Based Anomaly Detection. In Proceedings of the 2016 IEEE International Workshop on Information Forensics and Security (WIFS), Abu Dhabi, United Arab Emirates, 4–7 December 2016; pp. 1–6.
63. Shvetsova, N.; Bakker, B.; Fedulova, I.; Schulz, H.; Dylov, D.V. Anomaly Detection in Medical Imaging with Deep Perceptual Autoencoders. *IEEE Access* **2021**, *9*, 118571–118583. [[CrossRef](#)]



64. Rao, Y.; Ni, J. A Deep Learning Approach to Detection of Splicing and Copy-Move Forgeries in Images. In Proceedings of the 2016 IEEE International Workshop on Information Forensics and Security (WIFS), Abu Dhabi, United Arab Emirates, 4–7 December 2016; pp. 1–6.
65. Huh, M.; Liu, A.; Owens, A.; Efros, A.A. Fighting Fake News: Image Splice Detection via Learned Self-Consistency. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 101–117.
66. Bappy, J.H.; Simons, C.; Nataraj, L.; Manjunath, B.S.; Roy-Chowdhury, A.K. Hybrid LSTM and Encoder-Decoder Architecture for Detection of Image Forgeries. *IEEE Trans. Image Process.* **2019**, *28*, 3286–3300. [[CrossRef](#)] [[PubMed](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.