Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 1 (2024)

A Comprehensive Survey on the Advancements in Deep Learning Techniques for Detecting Deep Fake Images

Sheela Ramachandra, Smitha Rajagopal*

Dept of CSE, Alliance University, Bengaluru, India.

Abstract: -Understanding and detecting deepfake images is crucial to safeguarding the authenticity and reliability of visual content in digital media, preserving trust and credibility in online information dissemination. Effective detection using advanced deep learning techniques is vital in mitigating the potential risks posed by the proliferation of deceptive content.This survey explores recent advances in deep learning methods used to identify deepfake photos, emphasising their increasing importance in the modern digital environment. The paper provides an overview of the problems caused by deepfakes, examines the methods used to create false content, and studies the use of different deep learning models in the field of deepfake detection, including convolutional neural networks (CNNs), XceptionNet, and capsule networks. It draws attention to the ongoing struggle between the advancement of deepfakes and the development of detection techniques, highlighting the necessity of highly developed and adaptable neural network architectures in order to effectively prevent the spread of misleading data.

Keywords: *deep Fake Images, Deep Learning,XceptionNet,Convolution Neural Networks (CNN).*

1. Introduction

Deepfakes, primarily generated through deep learning models like autoencoders and generative adversarial networks (GANs), have become a significant concern due to their potential for misuse. Leveraging vast image and video datasets, these AI-driven methods fabricate realistic facial expressions and movements, often targeting public figures like celebrities and politicians. [1] The threat posed by deep fakes extends beyond creating fabricated adult content; it encompasses manipulating political scenarios, influencing elections, spreading false information, and even misleading military intelligence. This technology, if misused, can generate deceptive content such as forged speeches by world leaders, fake satellite images, and even alter historical narratives. However, amidst these concerns, deepfakes also hold positive applications, aiding in visual effects, digital avatars, voice generation for individuals without speech capabilities, and enhancing entertainment experiences. [2] Despite their potential benefits, the widespread use of deepfakes demands stringent measures to mitigate the risks associated with their misuse. Nevertheless, the concerning aspect remains that the negative applications of deepfakes far outweigh the positive ones. The evolution of sophisticated deep neural networks combined with the abundance of accessible data has rendered manipulated images and videos nearly indiscernible from authentic ones, challenging both human perception and advanced computer algorithms. This ease of manipulation has significantly lowered the barrier to creating deceptively authentic content, often requiring only an individual's photo or brief video snippet. Remarkably, recent advancements have even enabled the generation of convincing deep fakes using merely a static image. Consequently, the threat posed by deep fakes extends beyond public figures to impact everyday individuals as well. [3] The emergence of software like DeepNude and apps like Zao illustrates even more disturbing threats posed by deepfake technology. DeepNude's ability to transform individuals into non-consensual pornography and Zao's viral trend of inserting less-skilled users into movie scenes using face-swapping features raise serious concerns regarding privacy violations and identity manipulation, significantly impacting various facets of human life. [4] Verifying authenticity in the digital realm has become increasingly crucial, especially in light of deepfakes' rampant use for malicious intents. The accessibility of deepfake tools allows almost anyone to create deceptive content, intensifying the challenge of distinguishing between truth and falsified information.

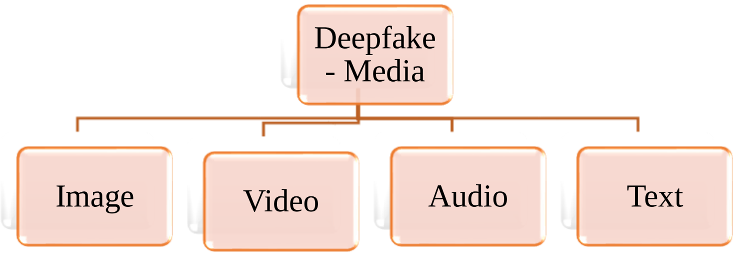

Figure 1. Types of Media – DeepFake

Deepfake technology is primarily used to manipulate and create deceptive content in various types of media. The common types of media where deepfake technology is applied include:

Images: Deepfake techniques can manipulate still images by altering facial features, expressions, backgrounds, or inserting elements that were not present in the original image. These manipulated images can be highly convincing and difficult to differentiate from authentic ones.

Videos: Deepfake videos involve the alteration or synthesis of video content. This can include replacing faces, changing facial expressions, modifying speech or gestures, or creating entirely synthetic videos of individuals saying or doing things they never did. Deepfake videos have gained attention due to their potential to spread misinformation or fabricate events.

Audio: Deepfake technology is also used to create synthetic audio content, known as "audio deepfakes" or "voice cloning." This involves imitating someone's voice or generating speech that mimics a specific individual, allowing the creation of false audio recordings.

Text: While not as prevalent as other forms, deepfake text generation can produce written content that mimics the style and tone of a particular writer. This technology can generate fake articles, reviews, or messages that appear authentic but are actually generated by AI algorithms.

Consequently, various methods have been proposed to detect deepfakes, many leveraging deep learning techniques. This evolution in detection methods reflects an ongoing conflict between the malicious exploitation and positive applications of deep learning methodologies. We classify deepfake detection methods according to the type of data they target, distinguishing between images and videos. We classify the features used in false image detection techniques as either handcrafted features or deep features. Our thorough analysis includes a thorough look at the difficulties faced in this area, identifies popular research topics, and suggests future paths for issues with multimedia forensics and deepfake detection.

The following research questions guide this exploration, aiming to uncover insights that contribute to a deeper understanding of the nuances surrounding deepfake detection:

- How effective are current deep learning models, such as convolutional neural networks (CNNs), XceptionNet, and capsule networks, in detecting a wide range of deepfake manipulation techniques?
- To what extent do these models generalize across different types of deepfake content, including images created using various algorithms and tools?
- What strategies can be employed to enhance the robustness of deep learning models against adversarial attempts to evade detection?

- What interdisciplinary approaches are promising in advancing the state-of-the-art in identifying manipulated content, considering the evolving nature of deepfake creation techniques?

2. Deep Learning for Deepfake Detection:

Deep Learning is a subset of machine learning that involves learning representations from data using artificial neural networks with multiple layers. The foundational unit of a neural network is the neuron, which receives inputs, processes them using weighted connections, and applies an activation function to produce an output. Deep Learning encompasses a wide range of architectures, such as feedforward neural networks, recurrent neural networks (RNNs), convolutional neural networks (CNNs), etc.

A neural network can be represented mathematically as follows:

Given an input vector **x** and parameters **W** (weights) and **b** (biases), a simple fully connected feedforward neural network can be represented by the following equations:

$$\mathbf{z}=\mathbf{W}\mathbf{x}+\mathbf{b}$$

$$\mathbf{a}=f(\mathbf{z})$$

Where:

- **x** is the input vector.
- **W**- weight matrix.
- **b**- bias vector.
- **z** is the intermediate output before activation.
- $f(\cdot)$ is the activation function.
- **a** is the final output after activation.

Deep learning techniques are commonly utilized in the detection and identification of deepfake content due to their ability to analyze and recognize patterns in complex data such as images and videos. [5] Some popular deep learning methods which used for deepfake detection include:

Convolutional Neural Networks (CNNs): CNNs are effective in extracting hierarchical features from images and videos, making them suitable for detecting inconsistencies or anomalies in visual data, which could indicate the presence of deepfake alterations.

Recurrent Neural Networks (RNNs): RNNs are utilized for processing sequential data and can analyze temporal information across frames in videos, aiding in detecting irregularities or inconsistencies in the flow of frames within a video, which may indicate manipulation.

Generative Adversarial Networks (GANs): Although GANs are often used to create deepfakes, they are also employed in detecting them. GAN-based detection models are trained to differentiate between authentic and manipulated content by identifying discrepancies in generated images or videos.

Capsule Networks: These networks are designed to handle hierarchical relationships between features, potentially enabling better understanding of spatial hierarchies and aiding in identifying unnatural distortions in deepfake images or videos.

Siamese Networks: Siamese networks are utilized for one-shot learning and can be used in identifying discrepancies between original and manipulated content by learning feature representations that help in recognizing differences.

Deep Feature Extraction: Pre-trained models like VGG, ResNet, or EfficientNet are used to extract deep features from images or videos. These features are then employed in traditional machine learning classifiers for detecting anomalies in the data.

The detection of deepfake content remains a challenging task due to the continuous advancements in deepfake generation techniques. Researchers are constantly developing and refining deep learning-based methods to effectively detect and mitigate the proliferation of misleading content.

2.1 Process of Fake Content Creation (Deepfakes):

Data Collection and Preprocessing:

The process of developing a deepfake model commences with the gathering of a dataset comprising target faces, voices, or other pertinent content essential for training purposes. This dataset acquisition phase involves collecting a diverse range of data samples, which serve as the foundational material for the subsequent model training. Subsequently, preprocessing techniques are employed to refine and prepare the gathered data before training the deepfake model. This preprocessing stage encompasses various tasks, including data cleaning, alignment, and meticulous preparation to appropriately formatted dataset and optimized for effective utilization in the training process. The goal of this preparatory phase is to enhance the dataset's quality, standardize its structure, and address any inconsistencies or imperfections, thereby facilitating robust and accurate training of the deepfake model.

Model Training:

The development of deepfake models involves the utilization of diverse deep learning architectures, which includes Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and encoder-decoder networks, among others, as the primary frameworks for model training. This training process revolves around optimizing the model's parameters, encompassing neural network layers, weights, and biases, through extensive utilization of large-scale datasets. GANs, a prominent example among these architectures, operate by orchestrating a competition between adiscriminator and a generator. The generator component continually generates content, aiming to produce increasingly realistic outputs to deceive the discriminator, while the discriminator endeavors to distinguish between genuine and artificially generated content.

Synthesis and Manipulation:

After training, the model is formed with the ability to create or modify content by adjusting certain data pieces. This manipulation includes a variety of transformative behaviours, such as creating fake photos depicting non-existent people or modifying speech or faces in recordings. By utilising deepfake techniques, these models are able to alter voices, change facial expressions, or change the context of the content. These advanced manipulation skills make it possible to create and modify multimedia information with amazing fidelity, frequently making it difficult to tell what is modified compared to what is original.

Refinement and Improvement:

In the field of deepfake technology, iterative refining techniques are essential for improving the calibre and reality of generated content. By means of these iterative cycles, the generated content is subjected to incremental upgrades with the goal of improving its overall quality and authenticity. Simultaneously, researchers and professionals in the domain continuously strive to improve the complexity of the underlying algorithms used in deepfake creation. The goal of this continuous algorithmic evolution is to produce increasingly complex and subtle modifications, making the generated information harder and harder to recognise as fake.

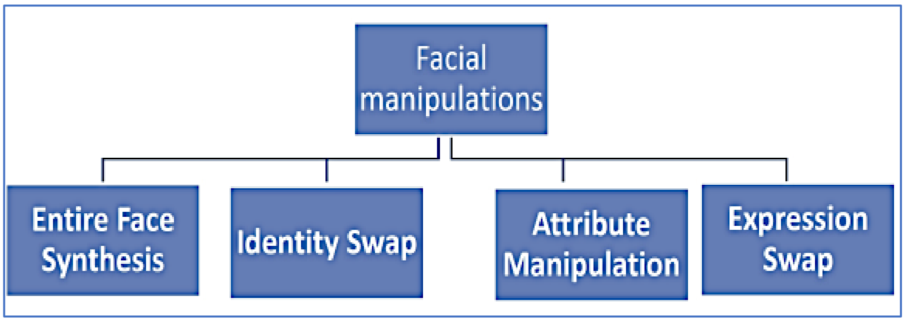

Figure 2. Face Image Manipulation

3. Deep Fake Detection of images using Deep Learning

In recent years, advancements in AI-driven image manipulation have led to an increased prevalence of falsified visual content. To combat this challenge, various CNN architectures such as XceptionNet, Meso Inception-Net, and ResNet have been instrumental in detecting signs of image manipulations. The listed are some of the prominent techniques employed by these models to identify different types of forgeries.

Visual Artifact Detection: CNN architectures are trained extensively to discern specific visual anomalies indicative of image manipulations. These models excel in detecting irregularities in textures, pixelation, blending borders, and inconsistencies arising from the alteration of original images. Leveraging learned features, CNNs efficiently recognize aberrations caused by editing tools and processes.

Inconsistency in Color and Blending Operations*:* During the creation of manipulated images, blending techniques often induce color irregularities or inconsistencies. CNN-based models are adept at scrutinizing color distribution, boundaries, and artifacts introduced during blending operations. This scrutiny enables the detection of subtle visual cues left by tampering or manipulation, facilitating the identification of potentially forged content.

Detection of Face X-ray Forgery*:* Face X-ray forgeries typically involve the blending of altered boundaries with authentic images. Employing CNN models, researchers can meticulously analyze these images, focusing on identifying distinct indicators of blending or inconsistencies inherent in the X-ray imagery. This targeted analysis assists in revealing discrepancies that signal potential tampering or manipulation in face X-ray images.

Classification of Loss in Detection*:* Following image analysis, CNN models perform classification tasks to ascertain the extent or nature of the detected forgery. This classification step aids in quantifying the level of manipulation present in the image, providing insights into the type and severity of the identified forgery.

3.1 CNN

Convolutional Neural Networks (CNNs or ConvNets) are a class of deep neural networks primarily designed for processing and analyzing visual data, such as images. They have revolutionized the field of computer vision and are widely used in various image-related tasks due to their effectiveness in learning hierarchical representations directly from raw pixel data.

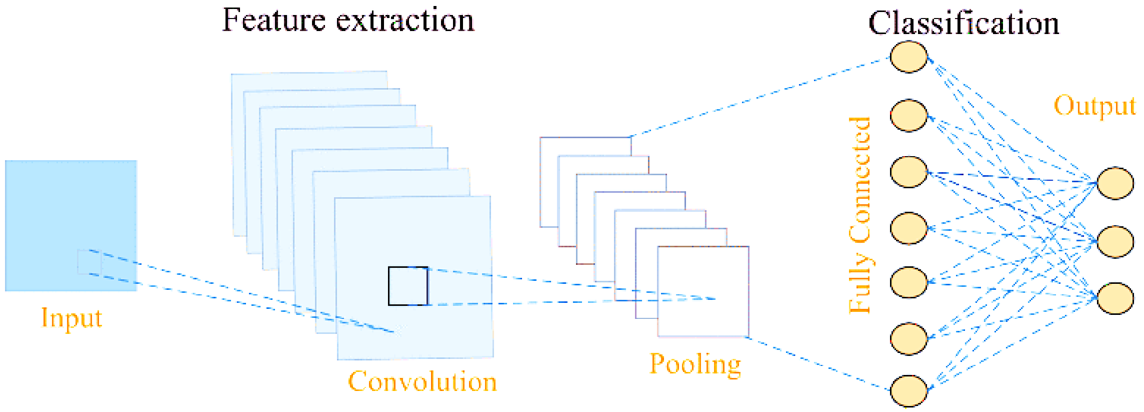

Figure 3. CNN Architecture

In the paper,***Ali. S. et. al (2022)*** proposed a novel image forgery detection system that leveraged neural networks, with an emphasis on Convolutional Neural Network (CNN) architecture, was introduced. The method exploited variations in image compression by utilizing differences between original and recompressed images as a means to train the model. This approach demonstrated effectiveness in detecting prevalent types of image forgeries, including splicing and copy-move manipulations.The authors highlight a notable gap in existing CNN-based forgery detection methods, which tend to be specialized in detecting a particular type of forgery. The proposed system aims to overcome this limitation by efficiently identifying a broader range of unseen forgeries within images.The core innovation lies in using the differences between an image's original and recompressed versions as cues for forgery detection. This suggests that the model is designed to capture artifacts and anomalies introduced during the recompression process, which can be indicative of manipulations.The authors emphasize the lightweight nature of the proposed model, suggesting a focus on computational efficiency and real-time processing capabilities. This is crucial for practical applications, especially in scenarios where quick and efficient forgery detection is essential. Experimentation produced highly promising results, showcasing an impressive overall validation accuracy of 92.23% within the iteration constraints.

The study by ***Suganthi et.al. 2022***implemented the FF-LBPH DBN (fisherface Linear binary pattern histogram using the DBN classifier) technique as a method to detect deepfake images, showcasing notable speed in execution while demonstrating highly effective differentiation between real and fake images. The proposed approach involves a hybrid methodology. Initially, Fisherface with Local Binary Pattern Histogram (FF-LBPH) is employed for face recognition, focusing on dimension reduction in the face space. Subsequently, Deep Belief Network (DBN) with Restricted Boltzmann Machines (RBM) is applied for deep fake detection. This combination aims to enhance the accuracy and efficiency of the detection process.This advancement holds promise in preventing unwarranted defamation resulting from manipulated images.The authors identify inaccuracies and high time consumption as major challenges in existing deep fake detection techniques. These challenges likely impact the efficiency and practicality of such methods, prompting the need for more effective solutions. The FF-LBPH model achieved impressive accuracy rates, reaching 98.82% in the CASIA-WebFace dataset and 97.82% in the DFFD dataset. The research's conclusion highlighted the superior performance of FF-LBPH in detecting and analyzing deepfake face images.By addressing existing issues and leveraging diverse datasets, the research contributes to the ongoing efforts in developing accurate and efficient techniques for detecting manipulated facial images created using deep learning methods.

In the paper, by ***Lakshmanan Nataraj et al. 2019***, a novel method was introduced to identify fake images by employing a blend of co-occurrence matrices and deep learning techniques. Co-occurrence matrices were extracted from three color channels in the pixel domain, and a deep CNN model was trained using this information.Co-occurrence matrices are commonly used in texture analysis, pattern recognition, and feature extraction in image processing tasks. They provide valuable information about the texture and spatial structure of images, which can be utilized in various image analysis and classification algorithms.Co-occurrence matrices are a statistical method used in image processing and analysis to capture the relationships between pixel values within an image. They provide information about the spatial distribution of pixel intensities by measuring how often pairs of pixel values occur in relation to each other at specific distances and angles within an image.The authors propose a unique detection approach that combines co-occurrence matrices, extracted on three color channels in the pixel domain, with deep learning. This hybrid methodology signifies an integration of traditional image processing techniques with advanced deep learning frameworks.The deep learning aspect involves the use of a convolutional neural network (CNN) framework. CNNs are well-suited for image-related tasks due to their ability to automatically learn hierarchical features, making them relevant for the complex task of GAN-generated fake image detection.The experimental evaluation encompassed two distinct and challenging GAN datasets, comprising over 56,000 images generated through unpaired image-to-image translations (cycleGAN) and facial attributes/expressions (StarGAN).The proposed approach is evaluated on two challenging GAN datasets: cycleGAN, focused on unpaired image-to-image translations, and StarGAN, centered around facial attributes/expressions. The inclusion of diverse datasets emphasizes the robustness and versatility of the proposed method. The findings illustrated the promise of this approach, achieving classification accuracies exceeding 99% in both datasets.The high classification accuracy and generalization across diverse datasets underscore the potential effectiveness of the proposed method in identifying manipulated images created using GAN-based techniques.

Using a multi-task learning approach, a CNN was meticulously developed in their study by ***Nguyen et.al (2019)***to simultaneously recognise modified images and videos and to accurately identify the edited regions within them. The strategy used by the network allowed information obtained from one activity to benefit from the other, creating a symbiotic relationship that improved the performance of both the detection and localization tasks. The implementation of a semi-supervised learning strategy greatly enhanced the network's capacity to generalise over a wide range of datasets. The architecture included a Y-shaped decoder and an encoder that were intended to maximise performance for the assigned tasks. The research's novelty lies in concentrating on detecting image forgeries linked to double image compression. This signaled a departure from conventional forgery types and introduced a specific contextual framework for the study.The fundamental methodology involved utilizing the disparity between an image's original and recompressed versions to train the deep learning model. This strategic choice of training data aligned precisely with the particular forgery type under scrutiny, providing a targeted investigative approach.

The intended objective of the framework proposed by ***Ambica Ghai et. al. 2021*** was to identify forged images manipulated through copy-move and splicing techniques. Employing an image transformation technique assisted in isolating pertinent features crucial for effective network training. Subsequently, a pre-trained customized CNN was utilized for training on publicly available benchmark datasets.The choice of a CNN indicates a reliance on a type of deep learning architecture well-suited for image-related tasks. Performance evaluation was conducted on the test dataset, employing diverse parameters to assess the model's effectiveness.The research addresses the pervasive issue of misinformation online, emphasizing the influence of manipulated content on decision-making. Focusing on the critical role of images in supplementing textual information and the rise of image manipulation, the study sets out to develop a deep-learning-based framework for detecting image forgeries.

Luca Guarnera et.al., 2020 proposedresearch involved utilizing an EM (Expectation Maximization) algorithm to extract a set of localized features tailored to model the inherent generative process.The research addresses the pervasive issue of misinformation online, emphasizing the influence of manipulated content on decision-making. The research addresses the prevalent and increasingly realistic phenomenon of Deepfakes, particularly focusing on human faces. The primary goal is to introduce a novel detection method capable of revealing forensics traces embedded in images during the deepfake generation process. The core methodology involves employing an EM algorithm to extract local features, specifically designed to characterize the convolutional generative process used in deepfake creation. The research validates the proposed technique through experimental tests, utilizing naive classifiers across five different architectures. The CELEBA dataset serves as the ground truth for non-fake images, enabling a comprehensive assessment of the technique's effectiveness.The approach, utilizing an EM algorithm and tested across various deepfake architectures, proves effective in distinguishing between different generative processes. This research contributes to the ongoing efforts to detect and mitigate the impact of deepfake images through advanced deep learning methodologies.

A multi-attentional network is a neural network architecture that incorporates multiple attention mechanisms to capture diverse and detailed information from input data. In the context of deep learning, attention mechanisms allow networks to focus on specific parts or aspects of the input, enabling more effective processing of complex data. One such network was proposed by ***Zhao et.al 2020*** whichadopts a deep learning architecture that incorporates multiple spatial attention heads, allowing the model to focus on different local parts of the images. The primary focus is on enhancing the detection of deepfake contents, considering the subtleties and local differences between real and fake images. The paper critiques the prevalent approach of treating deepfake detection as a standard binary classification problem and proposed a novel multi-attentional deepfake detection network.

The proposed deepfake detection framework adopts a fine-grained classification perspective, aiming to capture subtle and local differences between real and fake images. The key components of the methodology include:

Multiple Spatial Attention Heads: Introducing multiple attention heads to make the network focus on different local parts of the images.

Textural Feature Enhancement Block: Enhancing textural features to zoom in on subtle artifacts present in shallow features.

Aggregation of Features: Aggregating low-level textural features and high-level semantic features guided by attention maps to improve overall detection accuracy.

Additionally, the paper introduced specific strategies to address learning difficulties within the network, including a new regional independence loss and an attention-guided data augmentation approach.Extensive experiments conducted on various datasets demonstrate the effectiveness and superiority of the proposed method over conventional binary classifier counterparts. The research achieves state-of-the-art performance in deepfake detection, emphasizing the importance of a fine-grained classification approach and the incorporation of multiple attentional mechanisms.

Densely Connected Convolutional Networks, is a type of deep neural network architecture specifically designed to address challenges related to feature propagation and parameter efficiency in deep learning models.In DenseNet, each layer receives direct input from all preceding layers and passes its own feature maps to all subsequent layers. This dense connectivity pattern facilitates stronger feature reusability, encourages feature propagation, and combats the vanishing-gradient problem commonly encountered in very deep networks. DenseNet's design fosters feature reuse by concatenating feature maps from previous layers, enabling the network to exploit information flow efficiently throughout the architecture.a novel deep learning-based approach was introduced by ***Hsu et.al 2020*** for detecting fake images, employing contrastive loss as a pivotal component. The methodology commenced by utilizing several cutting-edge Generative Adversarial Networks (GANs) to generate pairs of fake and real images. Subsequently, a modified DenseNet architecture was developed, structured into a two-streamed network to process pairwise information as input.The core of the proposed approach revolved around training a common fake feature network using pairwise learning techniques, aimed at effectively distinguishing features characteristic of fake and real images. Following this, a classification layer was appended to the fake feature network to discern if the input image was genuine or fake.

3.2. Res-Net

ResNet, short for Residual Network, is a type of CNN architecture that introduced a novel building block called the "residual block." Residual networks were proposed by Kaiming He et al. in their paper "Deep Residual Learning for Image Recognition" in 2015.ResNet was designed to meet the challenge which is training very deep neural networks, where deeper networks often suffered from the problem of vanishing gradients or degradation.The key innovation of ResNet lies in the residual blocks. These blocks learn residual mappings, i.e., the difference between the learned representation and the input. Instead of learning the actual mapping from input to output directly, ResNet learns the residual mapping and adds this to the input, allowing the network to learn the residual functions.The architecture's fundamental principle is to use skip connections or shortcuts that directly connect earlier layers to later layers (bypassing one or more layers) to mitigate the vanishing gradient problem. ResNet architectures come in different depths, such as ResNet-18, ResNet-34, ResNet-50 etc., with varying numbers of layers. Deeper variants have shown impressive performance in various computer vision tasks, particularly in image classification, object detection, and segmentation, winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2015.

The approach utilized in study by ***Nawaz et.al. 2023***involved employing a deep learning methodology known as **ResNet-Swish-Dense54** for the precise and dependable detection of deepfake content. Initially, the method involves the extraction of human faces from input video frames. Subsequently, these extracted facial components undergo content classification using the ResNet-Swish-Dense54 model to discern whether they are genuine or manipulated representations.The chosen deep learning model, ResNet-Swish-Dense54, is employed for content classification, distinguishing between real and manipulated human faces.The proposed approach is evaluated on challenging datasets, including DFDC, FaceForensic++datasets, to assess its robustness. The evaluation involves testing against adversarial attacks, and the explainability power of the ResNet-Swish-Dense54 model is demonstrated through heatmap generation and cross-dataset validation.In this study, a novel ResNet-Swish-Dense54 framework was introduced for the effective capture of complex patterns in videos, specifically targeting deepfake detection. The architecture incorporates residual blocks with the Swish activation method to allow the flow of small negative values, optimizing the model's learning behavior. Additional dense layers were introduced at the end of the framework to enhance the extraction of reliable features for classification.The ResNet50 model served as the basis, known for its identity shortcut links and residual mapping to improve performance. Customizations included the incorporation of the swish activation method and the addition of dense layers to enhance keypoint nomination. The motivation for using swish activation was to improve learning behavior and empower the model to recognize intricate video patterns effectively.The structural details of the ResNet-Swish-Dense54 model include 33 convolution layers organized into five phases, each containing multiple residual blocks. Global average pooling and four added dense layers were also integrated. The dense layers contribute to emphasizing manipulated regions, filtering unnecessary background information, and improving deepfake detection under varying conditions such as light, color, and face orientations. The additional dense layers enhance the model's capability to learn a reliable set of image features with minimal architectural overhead.In summary, the study aimed to enhance deepfake detection using a tailored deep learning framework, leveraging the ResNet-Swish-Dense54 model with specific modifications. The architecture proved effective in capturing intricate video patterns, showcasing its potential for addressing the challenges associated with deepfake manipulation.

Image saliency refers to the visual distinctiveness or importance of certain regions within an image. Image saliency plays a role in feature extraction within computer vision and image processing. It determines the parts of an image that draw human attention most prominently. Saliency maps are used to highlight areas that are most relevant or striking to human perception. These maps are generated using computational algorithms that analyze various features of an image, such as color, contrast, texture, and spatial information, to identify regions with high saliency.The study by ***Yang et.al. 2021*** aimed to uncover subtle texture distinctions between authentic and manipulated facial images by leveraging the image saliency method. Utilizing an improved guided filter for image preprocessing, termed guided features, enhanced texture artifacts within both real and fake face images.

By employing a **ResNet18 network** capable of downsampling and resampling, the enlarged texture disparities were learned, facilitating precise detection of genuine and manipulated facial images.The research abstract focuses on spare-parts classification in heavy-asset industries using transfer learning via deep convolutional neural networks (CNNs). The authors propose a three-phase model for multi-criteria spare-parts classification, employing image-processing methods. The model involves reconstructing the neural network architecture based on the pre-trained VGG-19, building a hierarchical inference model for spare parts, and training/testing the model using real-world data.In the first phase, the authors adapt the VGG-19 architecture for spare-parts classification. In the second phase, a hierarchical inference model is established, aiming at data visualization and image conversion for spare parts.To validate the model, a comparative scheme is introduced, involving other CNN-based learning methods such as AlexNet and ResNet-50.The proposed model demonstrates outstanding performance with an overall accuracy of 95.87%.

Gandhi, A et.al 2020 proposed a method where adversarial perturbations were employed to deceive deepfake detectors, resulting in a significant reduction in detection accuracy to below 27% on perturbed deepfakes. To fortify detection robustness, two strategies were explored: Lipschitz regularization, demonstrating a 10% accuracy improvement in blackbox scenarios, and Deep Image Prior (DIP), achieving 95% accuracy in identifying perturbed deepfakes while maintaining 98% accuracy in other cases on a subset of 100 images. These methodologies aimed to strengthen deepfake detection against adversarial attacks, presenting promising advancements in defense techniques.Moreover, the research incorporated VGG and **ResNet**architectures, popular convolutional neural network models, to enhance the identification of distinguishing features within images, contributing to the advancement of deepfake detection against adversarial attacks.

3.3. Capsule Networks

Capsule networks (CapsNets) are a one of neural network architecture introduced by Geoffrey Hinton and his team. They are designed to overcome certain limitations of traditional convolutional neural networks (CNNs) in handling hierarchical spatial relationships within data.In contrast to CNNs, CapsNets focus on preserving the spatial relationships between different parts of an object by utilizing "capsules," which are neurons representing various properties of a specific part of an object. Capsules store details about the pose (orientation, position) and instantiation parameters of the entity they detect.In the context of deepfake detection, CapsNets have shown promise in efficiently identifying various types of attacks, including manipulated images and videos, while using fewer parameters compared to traditional CNNs. Their ability to preserve spatial hierarchies and handle complex relationships within data makes them a potential candidate for forensic applications, such as detecting digital forgeries and ensuring content authenticity.

Nguyen, H et.al. 2019introduced, a novel approach: a capsule network capable of detecting an array of attacks, including attacks involving printed images, replayed video, and deep learning-generated fake videos. Notably, this network demonstrated comparable performance to traditional convolutional neural networks while utilizing significantly fewer parameters.The authors have proposed a Capsule-Forensics method for detecting different types of deepfake images. The pipeline of their approach involves pre-processing, feature extraction using a part of the VGG-19 network, capsule network architecture, and post-processing.For video input, frames are separated, and for image input, facial areas are extracted using a face detection algorithm.Input images are pre-processed with common sizes like 100 × 100, 128 × 128, 256 × 256, and 299 × 299, with the authors choosing 300 × 300 for their model.The pre-processed images pass through a part of the VGG-19 network, specifically from the first layer to the third max pooling layer.This pre-trained CNN acts as a feature extractor, providing benefits of regularization, guidance for training, reduction of overfitting, and transfer learning.

The capsule network consists of primary capsules and two output capsules (for "real" and "fake").Primary capsules are designed with a 2D convolutional part, statistical pooling layer, and a 1D convolutional part. The statistical pooling layer ensures independence of input image size.Dynamic routing is used to calculate agreement between features extracted by primary capsules.The output capsules' activations determine the final result, and dynamic routing differs from classical fusion in its dynamic run-time calculation of agreement.The capsule network includes several primary capsules, with three typically used for light networks and ten for full ones.Each primary capsule is divided into three parts: a 2D convolutional part, a statistical pooling layer, and a 1D convolutional part.The dynamic routing algorithm calculates agreement between features for binary classification (real or fake).

Another research conducted by ***Nguyen, H et.al. 2019***gavecomprehensive investigations into several critical aspects including Replay Attack Detection, Face Swapping Detection, Reenactment Detection, and Computer-Generated Image Detection. This extensive exploration showcased the versatility of capsule networks beyond the realm of computer vision domains. Additionally, the study underscored the advantageous utilization of random noise during the training phase, demonstrating its efficacy across various scenarios.

3.4. XceptionNet

XceptionNet is a type of convolutional neural network architecture. It can achieve better performance in image classification tasks by exploring deeper and more complex network structures while maintaining computational efficiency. The separation of channel-wise and spatial convolutions helps in reducing the parameters while preserving the network's expressive power, leading to improved efficiency and performance compared to traditional CNN architectures.***Marra et.al 2018*** proposed a study on utilizing XceptionNet to detect fake images over social media. The study's main objective was to identify photos that had been modified using GAN-based image-to-image translation methods. Investigation indicated significant deterioration in performance, especially on Twitter-like compressed photos, even though some detectors showed high performance on undamaged images.

The given consolidated table provides an overview of the key studies, including the year of the study, the type of network used, success metrics, and key aspects of each approach for image forgery detection using deep learning techniques.

Table 1. Overview of Deep Learning Approaches for Deepfake Detection

Study	Year	Network Type	Success Metrics	Key Aspects
Ali. S. et.al (2022)	2022	CNN	Effective in detecting image forgeries	Lightweight model, leverages differences between original and recompressed images for training
Suganthi et.al. (2022)	2022	FF-LBPH with DBN	Impressive accuracy rates	Hybrid methodology, utilizes FF-LBPH for face recognition and DBN for deepfake detection
Lakshmanan Nataraj et.al. (2019)	2019	CNN with Co-occurrence Matrices	Effective integration of matrices	Hybrid approach, combines co-occurrence matrices with CNN
Luca Guarnera et.al. (2020)	2020	EM Algorithm	Distinguishing deepfake architectures	Leveraging EM algorithm for feature extraction in deepfake creation
Ambica Ghai et.al. (2021)	2021	CNN	Addressing image forgery detection	Image transformation technique for relevant feature isolation
Zhao et.al. (2020)	2020	Multi-Attentional CNN	State-of-the-art deepfake detection	Fine-grained classification, multiple spatial attention heads, and textural feature enhancement
Hsu et.al. (2020)	2020	Modified DenseNet	Effective discrimination in fake/real	Two-streamed network, pairwise learning, and classification layer
Nawaz et.al. (2023)	2023	ResNet-Swish-Dense54	Robust deepfake detection	Tailored ResNet-Swish-Dense54 framework, evaluation on

				challenging datasets, adversarial attack testing
Yang et.al. (2021)	2021	ResNet18 with Image Saliency	Precise detection of genuine and manipulated facial images	Utilizes image saliency to uncover texture distinctions, improved guided filter for preprocessing
Gandhi et.al. (2020)	2020	VGG, ResNet	Fortifying detection against adversarial attacks	Adversarial perturbations, Lipschitz regularization, Deep Image Prior (DIP), VGG and ResNet architectures
Nguyen et.al. (2019)	2019	Capsule Network	Comparable performance to CNNs	Capsule-Forensics method, detection of various attacks, dynamic routing for agreement
Nguyen et.al. (2019) (Comprehensive)	2019	Capsule Network	Versatility in various domains	Comprehensive investigation, random noise utilization during training
Marra et.al. (2018)	2018	XceptionNet	Detecting modified images on social media	Identifying photos modified using GAN-based image-to-image translation methods

4. Data Set

The foundation for the development and assessment of deepfake detection and generation algorithms is a wide range of datasets. Precisely selected and with a wide range of material, these datasets are essential tools for domain researchers and developers. From face-swapped deepfake videos of famous people in a variety of settings to high-resolution facial images produced by Generative Adversarial Networks (GANs), these datasets provide extensive sets that support evaluating and enhancing the effectiveness of deepfake detection and generation methods. A summary of these datasets, together with their individual information and attributes, is presented in the table below, highlighting their importance in promoting advancement and creativity in this constantly developing field.

Table 2. Deepfake Image Datasets Overview

Dataset Name	Description
Flickr-Faces-HQ, FFHQ	Contains 70,000 high-resolution face images generated by Generative Adversarial Networks (GANs).
100K-Faces	Comprises 100,000 unique human facial images generated using StyleGAN.
Fake face dataset (DFFD)	Consists of 100,000 to 200,000 fake images created by ProGAN and StyleGAN. The dataset contains approximately 47.7% male and 52.3% female images, predominantly aged between 21 to 50 years old.
CASIA-WebFace	Database with 10,000 subjects and 500,000 images collected from the IMDB website, including pictures of 10,575 renowned actors and actresses.
Celeb-DF (Ver. 2) dataset (CDF)	Largest publicly available dataset with 5639 face-swapped deepfake videos extracted from 590 YouTube videos featuring 61 celebrities in various settings.
Deep Fake Detection Challenge Preview dataset (DFDC-P)	Contains 1131 genuine and 4113 face-swap deepfake videos portraying 66 individuals of diverse demographics.
DeepFake Detection dataset (DFD)	Consists of 363 real and 3068 face-swap deepfake videos featuring 28 consenting actors performing various activities and expressions.

5. Conclusion

In conclusion, this survey has investigated the application of sophisticated deep learning methods for identifying deepfake images, which provide serious difficulties for content verification on web pages. The introduction underlined the necessity for efficient detection techniques and the growing issues regarding deepfakes.We have looked at a number of deep learning techniques in this work that are intended to identify these deceptive images. We discussed about the importance of deepfake detection and prevention tools like XceptionNet, capsule networks, and convolutional neural networks (CNNs).Our investigation also uncovered the complex processes involved in producing false images, particularly deepfakes, which shed light on the processes involved in these manipulations.The evaluation of previous studies highlighted the advancements made in of several neural network architectures for deepfake identification. While evaluating previous studies, the progress made in developing diverse neural network architectures for deepfake identification became evident. However, it is crucial to emphasize the persisting concerns about the inability to consistently and accurately identify deepfake images. As deepfake techniques evolve rapidly, there remains a pressing need for more advanced and flexible deep learning models to effectively combat the challenges posed by increasingly sophisticated manipulations. Future research endeavors will be directed towards developing cutting-edge solutions that can adapt to the dynamic nature of deepfake technologies, addressing the ongoing concerns surrounding the identification of deceptive content.

Competing Interests

We hereby declare that the authors do not have any competing interests

Data availability

We will deliver the data in accordance with the request.

Research Data policy

No datasets were generated or analysed during the current study

Compliance with Ethical Standards

We hereby declare that accepted principles of ethical and professional conduct have been followed during the course pf this work

References

[1] Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., ... & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, *223*, 103525.

[2] Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR), 54(1), 1-41.

[3] Chai, L., Bau, D., Lim, S. N., & Isola, P. (2020). What makes fake images detectable? understanding properties that generalize. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16 (pp. 103-120). Springer International Publishing.

[4] James Vincent,"New AI deepfake app creates nude images of women in seconds".June 27,2019.[Online].Avaiable: https://www.theverge.com/2019/6/27/18760896/deepfake- nude-ai-appwomen-deepnude-non-consensual- pornography.[Accessed:25-march2020]

[5] Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2023). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1520.

[6] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face x-ray for more general face forgery detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5001-5010).

[7] Suganthi, S. T., Ayoobkhan, M. U. A., Bacanin, N., Venkatachalam, K., Štěpán, H., & Pavel, T. (2022). Deep learning model for deep fake face recognition and detection. *PeerJ Computer Science*, *8*, e881.

[8] Shiohara, K., & Yamasaki, T. (2022). Detecting deepfakes with self-blended images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18720-18729).

[9] Ali, S. S., Ganapathi, I. I., Vu, N. S., Ali, S. D., Saxena, N., &Werghi, N. (2022). Image forgery detection using deep learning by recompressing images. Electronics, 11(3), 403.

[10] Ghai, A., Kumar, P., & Gupta, S. (2021). A deep-learning-based image forgery detection framework for controlling the spread of misinformation. Information Technology & People.

[11] Nawaz, M., Javed, A., & Irtaza, A. (2023). ResNet-Swish-Dense54: a deep learning approach for deepfakes detection. *The Visual Computer*, *39*(12), 6323-6344.

[12] Nataraj, L., Mohammed, T. M., Chandrasekaran, S., Flenner, A., Bappy, J. H., Roy-Chowdhury, A. K., & Manjunath, B. S. (2019). Detecting GAN generated fake images using co-occurrence matrices. arXiv preprint arXiv:1903.06836.

[13] Gandhi, A., Jain, S. Adversarial Perturbations Fooldeepfake Detectors. arXiv preprint arXiv2003.10596, 2020. https://doi.org/10.1109/IJCNN48605.2020.9207034

[14] Guarnera, L., Giudice, O., Battiato, S. Deepfake Detection by Analyzing Convolutional Traces. Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, 666-667. ttps://doi. org/10.1109/CVPRW50498.2020.00341

[15] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467.

[16] Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2185-2194).

[17] Marra, F., Gragnaniello, D., Cozzolino, D., &Verdoliva, L. (2018, April). Detection of gan-generated fake images over social networks. In 2018 IEEE conference on multimedia information processing and retrieval (MIPR) (pp. 384-389). IEEE.

[18] Nguyen, H. H., Fang, F., Yamagishi, J., & Echizen, I. (2019, September). Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th international conference on biometrics theory, applications and systems (BTAS)* (pp. 1-8). IEEE.

[19] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019, May). Capsule-forensics: Using capsule networks to detect forged images and videos. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2307-2311). IEEE.

[20]Hsu, C. C., Zhuang, Y. X., & Lee, C. Y. (2020). Deep fake image detection based on pairwise learning. Applied Sciences, 10(1), 370.