An Overview of Recent Work in Media Forensics: Methods and Threats

Kratika Bhagtani, Amit Kumar Singh Yadav, Emily R. Bartusiak, Ziyue Xiang, Ruiting Shao, Sriram Baireddy, Edward J. Delp

Video and Image Processing Laboratory (VIPER)
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, USA

Abstract

In this paper, we review recent work in media forensics for digital images, video, audio (specifically speech), and documents. For each data modality, we discuss synthesis and manipulation techniques that can be used to create and modify digital media. We then review technological advancements for detecting and quantifying such manipulations. Finally, we consider open issues and suggest directions for future research.

1 Introduction

The acquisition and circulation of digital media (i.e., images, video, and audio) has become popular with the proliferation of digital capture devices (e.g., smartphones), free tools for editing and manipulating digital media content (e.g., GIMP [1]), and social networks. With developments in deep learning [2, 3], manipulating digital media and generating synthetic content has become very easy [4, 5]. The manipulations can look extremely realistic and be challenging to detect [6]. The intent behind such manipulations is important to consider. For example, media could be manipulated to create misinformation or to commit financial fraud [7, 8]. Attempts to use manipulated multimedia for influencing social discourse, elections, and the way people interact in a civilized society [9] have driven research efforts in multimedia forensics [10, 11].

Multimedia forensics is the area of research that includes signal processing, computer vision, machine learning, statistics, psychology, sociology, and political science to study the development of manipulated and synthetic media and methods that can be used to detect and mitigate its use. The goals of media forensics are to answer the following questions: Is the media element manipulated (detection)?; Where is it modified (localization)?; What tools and/or who modified it (attribution)?; and Why did they modify it (characterization)? In this paper we focus on detection, localization, and attribution.

Multimedia signals (e.g., image, video, audio) have many characteristics that can be analyzed for detection, attribution, and localization. Some methods analyze pixels to determine if and how images and videos are manipulated. Audio methods can analyze waveform amplitudes and frequencies. Other methods utilize information about the “construction” of a digital asset. All digital assets are created by a system consisting of acquisition and in-device processing. For example, a typical digital camera system consists of acquisition through a system of lenses and in-camera processing such as Color Filter Array (CFA) interpolation, white balance, and gamma correction [12, 13]. Finally, the output is compressed. These processing methods leave traces—or fingerprints—in the information of a media asset that reveals the acquisition device and processing methods [14, 15]. We can attribute the media to its acquisition system by analyzing these fingerprints [16–20]. Additionally, we can use the fingerprints to verify whether the media is pristine, manipulated, or synthesized [16]. When an attacker manipulates the media, the fingerprints can be affected, enabling detection and localization of the manipulations. Deep learning models used to generate media, such as Generative Adversarial Networks (GANs) [21] and diffusion models [22], also have fingerprints. Thus, media forensics methods can also strive to recognize fingerprints of the generation method as part of their analysis [23, 24].

There are several review papers in media forensics, including image forensics [12, 25, 26], video forensics [27–31], audio forensics [32], theoretical analysis [33–35], author attribution [36], and document forensics [37–39]. In this paper we review recent work in media forensics for images, video, audio, and documents and emphasize work in machine learning approaches. Note this is a longer version of a paper presented at the 2022 IEEE International Conference on
2 Image Forensics

In this paper, we discuss images that are considered to be unaltered, manipulated, or synthesized. Unaltered images have been neither manipulated nor synthesized; they are authentic images as captured by a camera system. Typically, synthesized images refer to entire images that have been generated from scratch. Manipulated images are images in which portions of the images have been altered. Note that the manipulated portion of the image could be synthesized. Some common examples of image manipulation techniques include splicing [41] (replacing a section of an image with a section from another image), inpainting [42] (deleting a section of an image and synthesizing pixels to replace the content), copy-move [43] (duplicating a section of an image and moving it to another position within the same image), and photo-montage [44] (composite image from a combination of two or more images). Figure 1 shows examples of splicing and copy-move manipulations.

The level of realism that synthetic images have attained poses challenges in distinguishing them from unaltered images. This is complemented by counter forensic approaches that use strategies to deceive forensic methods. For example, Güera et al. added adversarial noise—specific types of noise that are indiscernible to the human eye and known to fool Convolutional Neural Networks (CNNs)—into images. The authors then showed that CNN-based camera model attribution methods were negatively impacted [45]. Bonettini et al. proposed a method to imperceptibly alter an image by removing camera-specific noise to severely hinder sensor noise-based camera model detectors [46]. Cozzolino et al. injected traces of real cameras into synthetic images to deceive detectors into identifying them as real [47]. Huang et al. proposed a method to evade detectors via notch filtering in the spatial domain [48]. These approaches further motivate multimedia forensics research.

In this section, we first introduce some recent methods for image manipulation and synthetic image generation, and then discuss developments in detection methods.

2.1 Image Manipulation and Synthesis

Common image manipulations such as splicing, copy-move attack, seam-carving, and inpainting can be performed using constraint-based methods [50] and deep learning [2, 3]. Several methods have been proposed for automatic image manipulation. CNNs were described in [44] for image compositing [51]. Yang et al. and Shetty et al. demonstrated automatic image inpainting utilizing a multi-scale neural network [42] and Generative Adversarial Network (GAN) [52], respectively. The methods proposed for image inpainting sometimes produce blurry regions or artifacts. This lack of high-frequency information can make manipulation detection easier. More recent methods aim to generate sharper edges of inpainted regions [53].

There has also been significant progress in research related to synthetic image generation [22, 52]. GANs [52] have enabled the synthesis of high-quality image content that is visually almost indistinguishable from unaltered images. A GAN typically consists of two competing networks: a generator that attempts to learn a data distribution and a discriminator that attempts to distinguish synthetic data created by the generator from the original data it is learning to model [21, 52]. GANs have been used to translate styles of an image collection (e.g., paintings) into an unrelated image collection [49]. Figure 1 shows an example of a style transfer from a painting to a photograph. Karras et al. described a training method for progressive growth of GANs to stabilize training and progressively improve the quality of generated images [54].

A style-based generator (e.g., StyleGAN [56]) can control
image synthesis by adjusting the style of a learned image at each convolution layer. In recent years, attempts have been made to improve the quality of GAN-generated images and address characteristic artifacts through improvements in architecture and training strategies [57,58]. Brock et al. described a real-world class-conditional image synthesis method trained on complex datasets that yielded high-fidelity images [59]. Karras et al. proposed a high-quality synthetic image generation method trained with limited data (i.e., with only a few thousand training images) [60]. In [61], an alias-free GAN known as StyleGAN3 was proposed, and the generating process is equivariant to translation and rotation. Figure 2 shows a synthetic image generated using this alias-free GAN.

Besides GANs, there are other image synthesis methods with the potential to generate high-quality synthetic images. Likelihood-based diffusion models [22] and score-based generative models [62,63] can also create synthetic images with high-frequency information, including sharp edges and fine details. Figure 2 shows a synthetic image generated using a likelihood-based diffusion model [22]. Chan et al. utilized 3D GANs for the synthesis of multi-view images of faces and animals [64]. Transformers, originally developed for natural language processing, can also generate high resolution images [65]. DALL-E [66,67] is a text-to-image method that uses a transformer based on GPT-3 [68]. It receives a text description as an input and produces images that fit that description. The rapid improvement in synthetic image generation methods demands equally robust detection methods.

2.2 Image Manipulation Detection

Many methods have been proposed to detect image splicing, which is a common type of manipulation. Wu et al. used a Deep Matching and Validation Network [69]. The network estimates a probability that a potential donor image has been used to splice a given query image and generates splicing masks for both images. One drawback of this approach is that along with the spliced query image, it also requires the donor image, which may not always be available. Nataraj et al. used a CNN-based method that operates on pixel co-occurrence matrices for image manipulation detection [41], which only requires the spliced query image. First, the co-occurrence matrices of the image under analysis are computed. The co-occurrence matrices describe textures within the image and their locations using a histogram of pixel pair values. Then, a ResNet50 [70] network analyzes the co-occurrence matrices and determines if the image is manipulated. This method is effective for a variety of manipulation techniques, but it does not provide a localization map indicating where the manipulations occur within the image.

Liu et al. proposed a method for detection and localization that utilizes both a matching strategy and an adversarial strategy [71]. In this approach, two images are analyzed to detect whether regions in each of the input images are the same. The method involves three networks: a deep-matching network, a detection network, and a discriminative network. The deep-matching network generates manipulation masks for both input images, highlighting matching regions. This part of the network serves as the generator of the adversarial approach. The detection network predicts whether the two images with their manipulation masks are correlated or not. The role of the discriminator network is to drive the generator to produce manipulation masks that are consistent with their ground truths. Although this method provides extra localization information about manipulation compared to the work proposed by Nataraj et al. [41], one drawback is that it requires two input images. More specifically, it requires the image that was used to splice the region into the host image, which might not always be available.

Other adversarial approaches do not employ a matching technique and thus do not require the image from which the spliced region is taken. For example, a conditional Generative Adversarial Network (cGAN) can be used to detect and localize forgeries in images [72–75]. In these approaches, a generator-discriminator pair is still utilized, but the generator only operates on a single input image. Thus, the generator is forced to detect manipulations based on inherent artifacts in the input image, without relying on a matching network or images that may contain the spliced pixels.

Some media forensics methods detect physical inconsistencies in images to determine if the images are unaltered. For example, Yarlagadda et al. and Kumar et al. investigated shadows [74,76]. Zhu et al. decomposed faces images into their 3D geometry and lighting parameters [77]. Other methods use media information, such as JPEG compression artifacts, for manipulation detection and localization [43,78,79]. Bonettini et al. [80] proposed a manipulation detection method that focuses on the absence of camera sensor noise in images.

![Figure 2: Synthetic images from a GAN (left) and a Diffusion model (right). Both images have mismatched pupil shape and eye color. The image on the left is generated by StyleGAN3 [55]; the image on the right is generated by [22].](image-url)
They showed that CNNs trained with this feature could generalize better in terms of unseen devices. Güera et al. analyzed which regions in images were most suitable for attribution detection [81].

Other techniques use an ensemble of methods to authenticate an image [82, 83]. Charitidis et al. used five detection and localization methods: two are based on detecting JPEG compression using Discrete Cosine Transform (DCT), another two are based on JPEG compression in spatial domain, and one is based on noise fingerprint approach. The decisions from these five methods were fused using deep learning for image manipulation detection and localization [82]. Barni et al. proposed a multi-branch CNN to localize tampered areas in a copy-move attack by identifying the source and target regions [83].

In realistic scenarios, manipulated images contain a combination of manipulations. There are methods that work to detect and localize [85, 86] manipulations in such situations. Noiseprint, a camera model fingerprint using a CNN, was proposed in [84] for image manipulation localization. The method produced heatmaps with brighter regions suggesting a possible manipulation. Figure 3 shows two examples of splicing and the corresponding heatmaps generated using Noiseprint for splicing detection.

The methods described above, which work well for detecting manipulations in “consumer camera images”, sometimes require more effort before they can be applied to satellite images. Satellites usually have different camera sensors and acquisition methods, which can create different images than consumer cameras. For example, Cannas et al. developed methods for satellite attribution based on panchromatic imagery [87]. Horváth et al. proposed splicing detection methods in satellite images using a vision transformer [88], a deep belief network [89], and an autoencoder combined with a one-class classifier [90]. Montserrat et al. used generative auto-regressive models to detect manipulations when their nature is unknown, which is the case in realistic scenarios [91].

2.3 Synthetic Image Detection

Methods for synthetic image detection often involve GAN-specific solutions which assume prior knowledge of the generative process. Giudice et al. proposed a solution by detecting anomalous frequencies while analyzing DCT coefficients of the image [92]. This work also showed that GANs leave architecture-specific fingerprints/signatures, which are not directly perceivable but are present in the spatial frequency domain of the synthetically generated image.

Recently, one of the focus points in image forensics has been detecting whether the image is synthetic without prior knowledge of its source or its manipulation history. Methods capable of accomplishing this level of detection are helpful in real-life scenarios. Several methods focused on identifying the GAN from which the synthetic image is generated. Wang et al. [93], Girish et al. [94] and Cozzolino et al. [95] demonstrated that training a classifier for a specific GAN generator could also be generalized for other GANs by detecting common synthetic image inconsistencies. Guarnera et al. used the detection of convolutional traces in synthetic images as a method of revealing them, which showed good performance for multiple GAN models [96]. Detection of synthetic images generated using GANs has been reviewed in detail [24].

Inconsistencies in GAN-generated images which occur due to the absence of any physiological constraints in the generative process can be used for detection. For example, Hu et al., and Guo et al. utilized corneal specular highlights [97] and irregularity in pupil shapes [98], respectively, to reveal synthetic GAN faces. They exploited the idea that when eyes in real faces look straight into the camera, they will see the same scene. Also, pupil shapes should be symmetrical for humans, which is not always consistent in synthetic images. Figure 2 shows examples of synthetic images with physical inconsistencies. This line of thought can also be extended to detect semantic inconsistencies to expose synthetic images (e.g., asymmetrical eyes and mismatched earrings).

3 Video Forensics

Several review papers [27–31] discussed video manipulation detection techniques. Kaur et al. classified video
manipulation as inter-frame and intra-frame [99]. Inter-frame manipulation involves modifying the order of frames. It can be done either by splicing (inserting frames from other video sequences into an original video sequence), cutting (deleting frames from an original video sequence, or copy-paste (duplicating frames from one temporal location to another within the original video sequence). Figure 4 describes these manipulation techniques further. Intra-frame manipulation involves modifying the contents of individual frames. This modification is sometimes done in multiple consecutive frame sets as shown in Figure 4. Javed et al. provides an in-depth review of detection methods for these manipulations [28].

Video signals are “constructed” in their acquisition and processing steps similar to other media types. This includes the camera sensor system and model-specific in-camera processing and compression. Each step of this system contributes to the video fingerprints. Methods for manipulated video detection often exploit the perturbations added to these video fingerprints by video editing. Bestagini et al. [27], Cozzolino et al. [30], and Shelke et al. [29] reviewed such detection methods in detail, which were based on camera fingerprints such as Photo-Response Non-Uniformity (PRNU) [27,30] and compression artifacts.

This overview covers only the recent works in video sequence manipulation detection. In this paper we shall refer to video that contains synthesized content generated using deep learning methods as “deepfake video” [100,101]. The discussion in this section is limited to methods for video analysis that do not use any audio associated with the video sequence. In Section 6 we discuss metadata and container methods for forensic analysis in video and audio. We will also briefly discuss analysis of both audio sequence and image stream in Section 7.

3.1 Fingerprint-Based Detection

In [20], a mirror network was used to learn camera fingerprints using unaltered video sequences from the camera. The fingerprint was used to classify the camera model that acquired the video sequence and to localize any video manipulation. When an edited video sequence is saved, it goes through another cycle of compression, which leads to double compression artifacts that can be exploited for detecting and localizing video manipulation. For example, Ravi et al. proposed analyzing MPEG compression artifacts using a Huber Markov Random Field (HMRF) to determine if the video was multiply compressed [102]. Subsequently, transition probability matrices of the noise were used to detect doubly compressed video sequences. Mahfoudi et al. demonstrated that DCT coefficients had a Laplacian distribution dependent on the quantization parameter used in the encoder [103], which could be used for double compression detection. Ud din et al. investigated HEVC/H.265 [104] encoded videos and utilized both statistical and deep convolutional neural network features for multiple compression detection [105]. In [106], the authors used Scale-Invariant Feature Transform (SIFT) [107] features to reveal intra-frame copy-paste attacks by detecting edge lines. In [108], a difference frame was generated for each frame and the edges detected using Canny edge detector [109] were thresholded to detect foreground manipulation. Ouadrhiri et al. demonstrated near-duplicate video detection using content-based features such as luminance (visual features), color motion features based on motion vectors, and high-level features [110].

3.2 Deepfake Video Detection

Deepfakes are realistic-looking synthetic video sequences generated using deep learning. Although they may look indistinguishable from real video sequences, the deepfake
video sequences featuring human faces are often inconsistent with how real human faces talk or move (e.g., abnormal blinking or breathing). To detect whether a video sequence is a deepfake, we can exploit these inconsistencies if we have access to the characteristics of unaltered video sequences. Several deepfake detection methods have been proposed based on this idea.

One of the earliest deepfake detection methods was proposed by Güera et al. [100,101]. The authors developed a temporal-aware pipeline to automatically detect deepfake videos. Their system uses a CNN to extract frame-level features. The features are used to train a Recurrent Neural Network (RNN) [111] that learns to classify if a video has been altered or not.

In [112], the authors used temporal facial features of a person (e.g., face movement) while talking. Facial features such as shape, expression, and pose [113] were used with a modified ResNet architecture [70,112]. The network, coupled with an adversarial training strategy, provided a representation of temporal facial behaviour. There were inconsistencies even in realistic-looking deepfakes with respect to the temporal face expression, which were used for deepfake detection. This method worked well on facial reenactment detection even in the presence of strong video compression.

In [114], high-level semantic facial information from an ensemble of CNNs was used to detect deepfakes. An attention mechanism was used to infer and focus only on the relevant parts of the face. This mechanism along with a modified EfficientNetB4 [115] were used to infer which part of face were relevant, and the method only focused on those parts for manipulation detection. In [116,117], face bounding boxes and facial landmark features from a Multi-Task Cascaded CNN (MTCNN) were used for deepfake video detection [118]. These feature were used with EfficientNetB4 [115] to classify real and synthetic faces. For each face region in the video sequence, a classification and a weight were assigned using the attention mechanism [119], which together provided the probability of the video sequence being synthetic. For temporal features, a Recurrent Neural Network (RNN) [111] was used to merge all features and generate the final decision.

Detecting deepfakes in highly compressed or low quality video sequences is challenging. Guhagarkar et al. used super resolution to improve the quality of the video sequence, which made it easier for the CNN to detect features for each frame [120]. These features were then used by a Long Short-Term Memory (LSTM) network [121] to capture temporal features for deepfake video detection.

Social media platforms are often the targets of manipulated video sequences. In [122], the authors observed a performance drop of CNN-based video manipulation detection methods when they were tested on manipulated videos shared on social media. They created a dataset of manipulated videos (non-shared), shared them on a social plat-
4.1 Methods for Manipulated Audio

One type of audio manipulation involves removing parts of an audio signal or copying them to another location within the same signal. Another type of manipulation involves pasting parts of an audio signal into another signal to create a spliced audio signal [129]. From acquisition to compression, each block in a digital audio system leaves a fingerprint, which can be analyzed for manipulation detection. For example, the microphone used for audio acquisition leaves characteristic fingerprints which are detectable in the frequency domain [130] and background noise [131]. These fingerprints can be used to attribute the audio signal to the microphone device. The existence of signals from more than one microphone can be an indication of audio signal tampering [132]. The environment in which the audio signal is recorded is referred to as the acoustic environment. The acoustic environment also contributes to the audio as signature smearing and ambient noise [32,133]. The presence of multiple dissimilar acoustic environments within an audio signal can also be used for audio splicing detection [32,133]. In [134], magnitudes of channel impulse response were captured using the audio spectrum and were used to classify the acoustic environment for small temporal segments of the audio signal.

After manipulation, the audio signal is typically re-compressed [32]. In [135–137], the authors explored the detection of double compression within an audio signal for manipulation detection. Such detection techniques often exploit features from Modified Discrete Cosine Transform (MDCT) coefficients used during compression [138, 139]. In [140], authors used MDCT coefficients with other MP3 codec data such as scalefactors, quantization step sizes, Huffman table indices, and sub-band window selection information to train a transformer to identify temporal location of multiply compressed audio signal.

To detect a copy-move attack in a speech signal, the signal can be divided into several voice segments and features such as pitch [141], Mel Frequency Cepstral Coefficients (MFCC) [142], and Delta-MFCC [142] can be used. Higher similarity (e.g., similarity evaluated using Pearson correlation coefficient) [141] between features of two segments can indicate a copy-move attack.

4.2 Methods for Synthesized Audio

Recent deep-learning based speech synthesis and voice conversion systems [143–146] can generate realistic human speech, which can fool Automatic Speaker Verification (ASV) systems. This is referred to as Logical Access (LA) attack to speech [147]. Synthetic audio can also be uploaded to online platforms to spread misinformation. Such attacks are referred to as deepfake attacks in [148]. Challenges such as ASVspoof2021 [147] encourage research in detecting such deepfake audio signals and LA attacks. When a recorded audio signal is replayed to fool ASV systems, the attacks are referred to as physical access attacks [147]. We have categorized synthetic audio detection methods into three categories: feature-based, image-based, and waveform-based.

Feature-Based Approaches: Audio features either from short-term window transforms such as Mel Frequency Cepstral Coefficients (MFCC) [142] or from long-term window transforms such as Constant-Q Cepstral Coefficients (CQCC) [149] are used to detect audio attacks. Hassan et al. used short-term spectral features consisting of MFCC, Gammatone Cepstral Coefficient (GTCC), spectral flux, and spectral centroid [150] to detect LA attacks. These features were used with a Recurrent Neural Network (RNN) [111] for

Figure 5: Example of an audio signal and its mel-spectrogram.
In [515], a normalized log-spectrogram of temporal segments was used for classifying the signal as synthetic or real. For classifying each temporal segment, an Efficient Convolutional Neural Network (EfficientCNN) [115] and its residual variant RES-EfficientCNN were used. In [152], the authors investigated long-term features based on Constant-Q Transform (CQT). Das et al. and Li et al. proposed long-term and high-frequency features such as Inverted Constant-Q Coefficient (ICQCC), Inverted Constant-Q Cepstral Coefficient (ICQCC), and Long-term Variable Q Transform (L-VQT) [153, 154] for detection. These features were then passed through a deep neural network to detect synthetic audio signals. In [155], the authors trained a Res2Net network with features such as log power magnitude spectrogram, Linear Frequency Cepstral Coefficient (LFCC), and CQT. Among all these features, the CQT-trained Res2Net performed better in detecting LA attacks. Various features used for synthetic speech detection have been reviewed in detail in [156].

**Image-Based Approaches:** In image-based synthetic audio detection, the spectrogram or the melspectrogram of the audio data is treated as an image and then analyzed using computer vision methods. A melspectrogram refers to a spectrogram with frequencies in mel scale [157, 158]. It graphically represents variation of frequency and intensity with respect to time. Figure 5 shows a melspectrogram. Bartusiak et al. used normalized gray-scale spectrograms of audio signal for synthetic speech detection using a CNN and a convolution transformer [117, 159, 160]. While in [161], the authors trained a temporal CNN and a spatial transform network using melspectrograms. For synthetic audio detection, these image-based methods outperformed feature-based methods including the ones using features related to energy, bandwidth, frequency, and short-term transform features such as MFCCs.

**Waveform-Based Approaches:** In waveform-based methods, the audio signal or waveform is used as an input to a deep neural network. The authors of [162] hypothesized that deep networks learned high-level features while spoofing generated subtle artifacts that could be better captured by shallower networks. They proposed the Time-domain Synthetic Speech Detection Network (TSSDNet) having multiple blocks similar to ResNet [70] and Inception [163]. In [164], the authors presented a convolutional RNN to detect synthetic speech. In this method, features from CNN were passed through a bidirectional LSTM [121] for synthetic speech detection.

In summary, existing methods focus on detecting synthetic audio to counter LA and deepfake attacks. More challenges (e.g., Audio Deep Synthesis Detection Challenge 2022 [148]) will promote research further in this area.

## 5 Text and Document Forensics

Documents such as news articles, bank receipts, scientific publications, social media posts, and business forms can be manipulated with malicious intent. Manipulated news articles can spread fake news or misinformation to negatively influence public opinion. Often, the documents circulated in companies do not contain marks to check their integrity visually. In such cases, document manipulation can lead to financial losses. Document manipulation detection methods have been proposed to prevent such scenarios [39]. Most common document manipulation detection methods can be of two types: (1) attributing the document to its original printer/scanner to check if it is authentic and not synthesized [17, 165]; and (2) checking for irregularities in a document by detecting manipulation [166].

We categorize document manipulation detection analysis as text forensics analysis and document forensics analysis. We also briefly discuss multimodal analysis involved in documents.

**Text Forensics Analysis:** Text forensics methods detect inconsistencies by analyzing the text of the document. Several methods have been proposed to detect text manipulation using machine learning. Aldwairi et al. proposed a method to identify potential fake news in websites by detecting misleading words and informal phrases in the website links, analyzing the length of titles and the use of punctuation marks [167]. The method used text to detect malicious websites after they were retrieved by the search engine and notified the user that they may contain misinformation. Ahmad et al. used linguistic features with an ensemble of different text classifiers to distinguish fake news articles from real ones [168]. These linguistic features included punctuation, emotions associated with words (positive or negative), and grammar. Recent Text Generative Models (TGMs) [68, 169–171] can produce convincing human language-like text which can spread fake news, generate false online reviews on products, and can be used for spamming emails. Zellers et al. proposed Grover, which is both a text generator and detector for fake news [172]. Given an article heading, Grover generates text for the article and can produce persuasive misinformation. It can also discriminate fake content using artifacts introduced by the Grover generator. Jawahar et al. discussed recent methods for detecting text generated using TGMs in detail [166].

Author attribution and author verification are also important and challenging problems in text forensics analysis. Author attribution relates to identifying the author of a given document to prevent deceit, while author verification is concerned with finding out whether multiple given documents were written by the same author. Traditionally, methods for authorship analysis are based on the extraction of stylometric features. Stylometric features, *i.e.*, statistical features of a
Document Forensics Analysis: Document image forensics methods treat a document as an image and then use variations of image forensics methods to analyze the document. Images of handwritten signatures can be used for verification of printed documents. Gideon et al. proposed a method using Convolutional Neural Networks (CNNs) to detect the probability of a handwritten signature being genuine or manipulated [180]. Beusekom et al. analyzed text-line alignment in document images for their authentication [181]. They detected text-lines in the document image and measured skew angles between lines to determine the probability of the document being manipulated. Cruz et al. used uniform Local Binary Patterns (LBP) for texture features that are characteristic of manipulated regions [182]. They used a Support Vector Machine (SVM) [183] to identify whether patches of document images were manipulated. Document images can also be checked for manipulation by attributing them to an original scanner or printer, which leave their fingerprints on the images [38,184].

The prevalence of scientific publications with problematic images has risen significantly over the past decade [185]. Retouched or reused images are important reasons for retractions of scientific publications. These edits are examples of scientific misconduct that undermine the integrity of the presented research. Traditionally, progress in scientific publishing has relied on a relatively slow cycle of peer review, where human experts need to inspect the authenticity of images [186,187]. Technologies that can automatically examine figures in published scientific papers will benefit experts who need to accomplish such an important and difficult task. Zhuang et al. analyzed graphical integrity issues in open access publications by verifying if the size of shaded areas in scientific figures were consistent with their corresponding quantities [188]. Moreira et al. demonstrated a human-in-the-loop end-to-end scientific publication analysis process. It starts by extracting content from uploaded PDFs, performs image manipulation detection on the automatically extracted figures, and ends with image provenance graphs expressing the relationships between the images in question, allowing experts to examine potential problems [187].

Multimodal Analysis: Document forensics often overlaps with multimodal analysis (analysis of a media which contains some combination of image, video, audio, and text). For example, news articles frequently contain images, captions, and text. Several methods have been proposed for fact-checking news articles. Vo et al. proposed a method for fact-checking articles using both text and images through a Multimodal Attention Network (MAN) [189]. Fung et al. used a graph-based neural network for fake news detection [190].

In articles with multimedia assets, authors may manipulate images to match text or manipulate the text in article to convince readers of what they are trying to convey, usually to present a false narrative. Some authors may even misuse some images in their articles in hopes of grabbing attention. Often, such manipulations lead to image-text mismatch, which can be used for cross-model manipulation detection. Image-text matching [191,192] refers to checking semantic similarity of an image in an article with its caption. Zhang et al. and Li et al. proposed a projection matching loss [193] and a Visual Semantic Reasoning Network [194], respectively, for image-text matching. Li et al. proposed a deep learning method known as Object-Semantics Aligned Pre-training (Oscar), which used object tags detected in images as anchor points to significantly ease the learning of alignments [195].

Tan et al. used visual-semantic representations for detecting inconsistencies in news articles [196]. These visual-semantic representations which included representations of text, objects in the images, and captions were used for classifying the article as either machine-generated or human-generated. McCrae et al. proposed a fusion method to detect manipulations in social media posts by identifying inconsistencies between videos and their captions [197]. These methods used machine learning and deep learning methods to classify media as genuine or manipulated.

6 Metadata Forensics

Almost all media files contain metadata. For example, the MP4 video container stores information about the modification date, video bitstream format, and video length [198]. MP3 compressed audio signals contain encoding parameters about the details of windowing, quantization, and Huffman encoding [199]. Metadata may not be used by end users directly, but media files cannot be decoded correctly without them. Many recent publications have shown that metadata in video and audio files can be used for forensic analysis [101,135,140,198,200,201]. There are mainly two advantages for forensic analysis using metadata. First, since the metadata format and structure are strictly specified in the standards of video/audio bitstream, it is more challenging for one to conceal or alter the forensic traces left in metadata without damaging the media file. Second, many existing video/audio manipulation or synthesis methods do not at-
The encoding parameters in video bitstream can be used for forensic analysis. In [205], the motion vectors from motion compensation were used for video device attribution. The encoding parameters are also used extensively for double compression detection. In [103, 105, 201, 206], the authors investigated double compression detection using encoding parameters in H.264 bitstreams. In [105, 207–209], the authors used encoding parameters in H.265 video bitstreams for double compression detection. Altinisik et al. used both encoding parameters and video container metadata for video device attribution, which resulted in improved performance compared to methods using container metadata only [210].

The encoding parameters in audio bitstream can be used for forensic analysis. Most related work focuses on two popular audio compression methods, namely MP3 and AAC [139]. In [135, 136, 211], the authors used the Modified Discrete Cosine Transform (MDCT) coefficients for double compression detection in MP3. Ma et al. used the MP3 scalefactors for double compression detection [212]. In [213], the authors achieved MP3 double compression detection using the Huffman table indices. Yan et al. used scalefactors and Huffman table indices information in MP3 files for double and triple compression detection [214]. In [215], the authors used the MDCT coefficients in MP3 compressed signals for multiple compression localization. Xiang et al. used a transformer neural network to process many types of MP3 encoding parameters, and their method achieved multiple compression localization with high accuracy [140]. In [216], the authors used the Huffman table indices in AAC files for double compression detection. Huang et al. achieved AAC double compression detection using the Quantized MDCT (QMDCT) coefficients [217]. In [218], the authors used the scalefactors from AAC compressed signals for double compression detection.

7 Discussion and Future Work

Research in media forensics is advancing rapidly, driven by competition with advancements in techniques that manipulate and generate media. In this paper we discuss some of the recent research for detecting manipulated images, videos, audio signals, and documents. Developments in deep learning have made it easy to individually generate deepfake audio, image, video, and text. Multimedia assets typically have more than one media type present (e.g., a video has image sequences and audio sequences, news articles have text along with images). Current counter-forensic approaches struggle to simultaneously manipulate these multiple data modalities consistently. Future work should focus on leveraging multi-modal analysis to detect these inconsistencies in order to identify manipulated media.

For example, in [219], the authors detected deepfake videos of humans by analyzing both audio and image sequences individually by looking for emotional inconsistencies across them. Audio and image sequences were divided into
temporal segments. For each audio segment using speech features, an LSTM was used to predict the subject’s emotion. For each image sequence segment, a similar LSTM using facial features was also used to predict emotions. The Lin’s Concordance Correlation Coefficient [219] estimated the correlation between the video and audio and predicted inconsistencies of the emotions. We aim to apply a similar concept to image-text cross-modal forensic analysis, with an approach that exploits the alignment between object labels and text, as well as attention regions in an image (such as faces) and name entities in text to examine the consistency between image and text.

Another direction to explore is combining statistical features and neural network features together to improve attribution and verification of a document. Lastly, we can look at intent prediction. Multiple media assets may be manipulated for a common purpose. Identifying common manipulation motivations of different media types would provide forensics analysts with more information about the nature of the manipulations.

Acknowledgments

This paper is based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement numbers FA8750-20-2-1004 and FA8750-16-2-0173. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL, or the U.S. Government.

Address all correspondence to Edward J. Delp at ace@ecn.purdue.edu.

References

[1] “GIMP - GNU Image Manipulation Program,” https://www.gimp.org/, 2021.
[2] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.
[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York Springer, 2001.
[4] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenhaim, L. RP, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang, “DeepFaceLab: Integrated, Flexible and Extensible Face-Swapping Framework,” arXiv preprint arXiv:2005.05535, May 2020.
[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaityl, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, “Natural TTS Synthesis by Conditioning WaveNet on MEL Spectrogram Predictions,” Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4779–4783, April 2018, Calgary, Canada.
[6] H. Farid, “Seeing is not Believing,” IEEE Spectrum, vol. 46, no. 8, pp. 44–51, August 2009.
[7] S. J. Nightingale and H. Farid, “AI-Synthesized Faces are Indistinguishable from Real Faces and More Trustworthy,” Proceedings of the National Academy of Sciences, vol. 119, no. 8, February 2022.
[8] J. Wakefield, “Deepfake Presidents Used in Russia-Ukraine War,” https://www.bbc.com/news/technology-60780142, March 2022.
[9] C. Vaccari and A. Chadwick, “Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News,” Social Media + Society, vol. 6, no. 1, pp. 1–12, February 2020.
[10] S. Lyu, “Fighting AI-Synthesized Fake Media,” Proceedings of the Workshop on Synthetic Multimedia - Audiovisual Deepfake Generation and Detection, pp. 1–2, October 2021.
[11] S. Battito, O. Giudice, and A. Paratore, “Multimedia Forensics: Discovering the History of Multimedia Contents,” Proceedings of the International Conference on Computer Systems and Technologies, pp. 5–16, June 2016, Palermo, Italy.
[12] A. Šiva, “An Overview on Image Forensics,” International Scholarly Research Notices, vol. 2013, pp. 1–22, January 2013.
[13] R. Boehme and M. Kirchner, “Counter-forensics: Attacking Image Forensics,” in Digital Image Forensics: There is More to a Picture than Meets the Eye. Springer, July 2013, pp. 327–366.
[14] H. Farid, “Image Forgery Detection,” IEEE Signal Processing Magazine, vol. 26, no. 2, pp. 16–25, March 2009.
[15] E. J. Delp, “Multimedia Security: The 22nd Century Approach,” Multimedia Systems, vol. 11, no. 2, pp. 95–97, December 2005.
[16] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, “Determining Image Origin and Integrity Using Sensor Noise,” IEEE Transactions on Information Forensics and Security, vol. 3, no. 1, pp. 74–90, February 2008.
[17] N. Khanna, A. K. Mikkilineni, G. T. C. Chiu, J. P. Allebach, and E. J. Delp, “Scanner Identification Using Sensor Pattern Noise,” Proceedings of the Society of Photo-Optical Instrumentation Engineers, vol. 6505, pp. 563–573, March 2007, Florence, Italy.
[18] R. Shao and E. J. Delp, “Forensic Scanner Identification Using Machine Learning,” Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 1–4, March 2020, Virtual.
[19] N. Khanna, A. K. Mikkilineni, P.-J. Chiang, M. V. Ortiz, S. Suh, G. T.-C. Chiu, J. P. Allebach, and E. J. Delp, “Sensor Forensics: Printers, Cameras and Scanners, They Never Lie,” Proceedings of the IEEE International Conference on Multimedia and Expo, pp. 20–23 July 2007, Beijing, China.
[20] D. Cozzolino, G. Poggi, and L. Verdoliva, “Extracting Camera-Based Fingerprints for Video Forensics,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop, pp. 130–137, June 2019, Long Beach, CA.
[21] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, “A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications,” IEEE Transactions on Knowledge and Data Engineering, pp. 1–1, November 2021.
[22] P. Dhariwal and A. Q. Nichol, “Diffusion Models Beat GANs on Image Synthesis,” Advances in Neural Information Processing Systems, vol. 34, pp. 8780–8794, December 2021, Virtual.
[23] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, “Do GANs Leave Artificial Fingerprints?” Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval, pp. 506–511, March 2019, San Jose, CA.
[24] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and
Systems, vol. 30, pp. 6000–6010, December 2017, Long Beach, CA.

N. Gugaharkar, S. Desai, S. Vaishampayan, and A. Save, “A Novel Approach to Detect Low Quality Deepfake Videos,” Journal of Scientific and Engineering Research, vol. 14, no. 1, pp. 1–7, 2021.

S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, p. 1735–1780, November 1997.

D. Marcon, C. Pasquini, and G. Boato, “Detection of Manipulated Face Videos over Social Networks: A Large-Scale Study,” Journal of Imaging, vol. 7, p. 193, September 2021.

L. Bondi, E. Daniele Cannas, P. Bestagini, and S. Tubaro, “Training Strategies and Data Augmentations in CNN-Based DeepFake Video Detection,” Proceedings of the IEEE International Workshop on Information Forensics and Security, pp. 1–6, December 2020, Shanghai, China.

G. E. Hinton, A. Krizhevsky, and S. D. Wang, “Transforming Auto-Encoders,” in Artificial Neural Networks and Machine Learning. Springer Berlin Heidelberg, 2011, pp. 44–51.

H. H. Nguyen, J. Yamagishi, and I. Eichen, “Capsule-Forensics Networks for Deepfake Detection,” in Handbook of Digital Forensics: Detection and Prevention. John Wiley & Sons, Inc., 2019, pp. 275–301.

V. Mazzia, F. Salvetti, and M. Chiaberge, “Efficient-CapsNet: Capsule Network with Self-Attention Routing,” Scientific Reports, Springer Nature, vol. 11, p. 14634, July 2021.

W. Huang and F. Zhou, “DA-CapsNet: Dual Attention Mechanism Capsule Network,” Scientific Reports, Springer Nature, vol. 10, p. 11538, July 2020.

O. Alamaryeh and M. Barni, “Detection of GAN-Synthesized Street Videos,” Proceedings of the European Signal Processing Conference, pp. 811–815, September 2021, Virtual.

P. R. Bevinamarad and M. Shirldonkar, “Audio Forgery Detection Techniques: Present and Past Review,” Proceedings of the International Conference on Trends in Electronics and Informatics, pp. 613–618, June 2020, Tirunelveli, India.

R. Buchholz, C. Kraetzer, and J. Dittmann, “Microphone Classification Using Fourier Coefficients,” International Workshop on Information Hiding, pp. 235–246, June 2009.

S. Ikram and H. Malik, “Digital Audio Forensics Using Background Noise,” Proceedings of the IEEE International Conference on Multimedia and Expo, pp. 106–110, July 2010, Singapore.

L. Bondi, E. Daniele Cannas, P. Bestagini, and S. Tubaro, “Training Strategies and Data Augmentations in CNN-Based DeepFake Video Detection,” Proceedings of the IEEE International Workshop on Information Forensics and Security, pp. 1–6, December 2020, Shanghai, China.

G. E. Hinton, A. Krizhevsky, and S. D. Wang, “Transforming Auto-Encoders,” in Artificial Neural Networks and Machine Learning. Springer Berlin Heidelberg, 2011, pp. 44–51.

H. H. Nguyen, J. Yamagishi, and I. Eichen, “Capsule-Forensics Networks for Deepfake Detection,” in Handbook of Digital Forensics: Detection and Prevention. John Wiley & Sons, Inc., 2019, pp. 275–301.

V. Mazzia, F. Salvetti, and M. Chiaberge, “Efficient-CapsNet: Capsule Network with Self-Attention Routing,” Scientific Reports, Springer Nature, vol. 11, p. 14634, July 2021.

W. Huang and F. Zhou, “DA-CapsNet: Dual Attention Mechanism Capsule Network,” Scientific Reports, Springer Nature, vol. 10, p. 11538, July 2020.

O. Alamaryeh and M. Barni, “Detection of GAN-Synthesized Street Videos,” Proceedings of the European Signal Processing Conference, pp. 811–815, September 2021, Virtual.

P. R. Bevinamarad and M. Shirldonkar, “Audio Forgery Detection Techniques: Present and Past Review,” Proceedings of the International Conference on Trends in Electronics and Informatics, pp. 613–618, June 2020, Tirunelveli, India.

R. Buchholz, C. Kraetzer, and J. Dittmann, “Microphone Classification Using Fourier Coefficients,” International Workshop on Information Hiding, pp. 235–246, June 2009.

S. Ikram and H. Malik, “Digital Audio Forensics Using Background Noise,” Proceedings of the IEEE International Conference on Multimedia and Expo, pp. 106–110, July 2010, Singapore.

L. Bondi, E. Daniele Cannas, P. Bestagini, and S. Tubaro, “Training Strategies and Data Augmentations in CNN-Based DeepFake Video Detection,” Proceedings of the IEEE International Workshop on Information Forensics and Security, pp. 1–6, December 2020, Shanghai, China.

G. E. Hinton, A. Krizhevsky, and S. D. Wang, “Transforming Auto-Encoders,” in Artificial Neural Networks and Machine Learning. Springer Berlin Heidelberg, 2011, pp. 44–51.

H. H. Nguyen, J. Yamagishi, and I. Eichen, “Capsule-Forensics Networks for Deepfake Detection,” in Handbook of Digital Forensics: Detection and Prevention. John Wiley & Sons, Inc., 2019, pp. 275–301.

V. Mazzia, F. Salvetti, and M. Chiaberge, “Efficient-CapsNet: Capsule Network with Self-Attention Routing,” Scientific Reports, Springer Nature, vol. 11, p. 14634, July 2021.

W. Huang and F. Zhou, “DA-CapsNet: Dual Attention Mechanism Capsule Network,” Scientific Reports, Springer Nature, vol. 10, p. 11538, July 2020.

O. Alamaryeh and M. Barni, “Detection of GAN-Synthesized Street Videos,” Proceedings of the European Signal Processing Conference, pp. 811–815, September 2021, Virtual.

P. R. Bevinamarad and M. Shirldonkar, “Audio Forgery Detection Techniques: Present and Past Review,” Proceedings of the International Conference on Trends in Electronics and Informatics, pp. 613–618, June 2020, Tirunelveli, India.

R. Buchholz, C. Kraetzer, and J. Dittmann, “Microphone Classification Using Fourier Coefficients,” International Workshop on Information Hiding, pp. 235–246, June 2009.

S. Ikram and H. Malik, “Digital Audio Forensics Using Background Noise,” Proceedings of the IEEE International Conference on Multimedia and Expo, pp. 106–110, July 2010, Singapore.

L. Bondi, E. Daniele Cannas, P. Bestagini, and S. Tubaro, “Training Strategies and Data Augmentations in CNN-Based DeepFake Video Detection,” Proceedings of the IEEE International Workshop on Information Forensics and Security, pp. 1–6, December 2020, Shanghai, China.

G. E. Hinton, A. Krizhevsky, and S. D. Wang, “Transforming Auto-Encoders,” in Artificial Neural Networks and Machine Learning. Springer Berlin Heidelberg, 2011, pp. 44–51.

H. H. Nguyen, J. Yamagishi, and I. Eichen, “Capsule-Forensics Networks for Deepfake Detection,” in Handbook of Digital Forensics: Detection and Prevention. John Wiley & Sons, Inc., 2019, pp. 275–301.

V. Mazzia, F. Salvetti, and M. Chiaberge, “Efficient-CapsNet: Capsule Network with Self-Attention Routing,” Scientific Reports, Springer Nature, vol. 11, p. 14634, July 2021.

W. Huang and F. Zhou, “DA-CapsNet: Dual Attention Mechanism Capsule Network,” Scientific Reports, Springer Nature, vol. 10, p. 11538, July 2020.
