Media Forensics and DeepFakes: an overview

Luisa Verdoliva

arXiv:2001.06564v1 [cs.CV] 18 Jan 2020

***Abstract*—With the rapid progress of recent years, techniques that generate and manipulate multimedia content can now guarantee a very advanced level of realism. The boundary between real and synthetic media has become very thin. On the one hand, this opens the door to a series of exciting applications in different fields such as creative arts, advertising, film production, video games. On the other hand, it poses enormous security threats. Software packages freely available on the web allow any individual, without special skills, to create very realistic fake images and videos. So-called deepfakes can be used to manipulate public opinion during elections, commit fraud, discredit or blackmail people. Potential abuses are limited only by human imagination. Therefore, there is an urgent need for automated tools capable of detecting false multimedia content and avoiding the spread of dangerous false information. This review paper aims to present an analysis of the methods for visual media integrity verification, that is, the detection of manipulated images and videos. Special emphasis will be placed on the emerging phenomenon of deepfakes and, from the point of view of the forensic analyst, on modern data-driven forensic methods. The analysis will help to highlight the limits of current forensic tools, the most relevant issues, the upcoming challenges, and suggest future directions for research.**

***Index Terms*—Digital image forensics, video forensics, deep learning, deepfakes.**

I. INTRODUCTION

Fake multimedia has become a central problem in the last few years, especially after the advent of the so called *Deepfakes*, i.e., images and videos manipulated using advanced deep learning tools, like autoencoders (AE) or generative adversarial networks (GAN). With this technology, creating realistic manipulated media assets may be very easy, provided one can access large amounts of data. Applications include movie productions, photography, video-games and virtual reality. The very same technology, however, can also be used for malicious purposes, like creating fake porn videos to blackmail people, or building fake-news campaigns to manipulate the public opinion. In the long run, it may also reduce trust in journalism, including serious and reliable sources. Figure 1 shows some popular deepfakes circulating on the internet. These fakes are easy to spot since they were generated for fun and involve well-known actors and politicians in unlikely situations. In addition, on the web it is usually possible to retrieve both the original and the manipulated version, removing any doubt about authenticity. However, verifying digital integrity becomes much more difficult if the video portrays a less known person and only the manipulated version is publicly available. This scenario takes place, for example, if the attacker films a new video on his own, with a collaborative actor whose face is eventually replaced by the target face. Governmental bodies, enforcement agencies, the news industry, and also the man in the street are becoming acutely aware of the potential menace carried by such a technology. The scientific community is asked to develop reliable tools for automatically detecting fake multimedia.

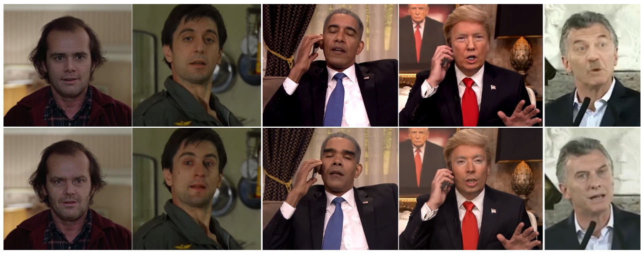

Fig. 1. Examples of deepfake manipulations from YouTube. Top: manipulated videos; bottom: original videos. It is worth noting that in the real videos Obama and Trump are impersonated by comic actors.

Actually, this is not a new problem. Image manipulation has been carried out since photography was born[1], and powerful image/video editing tools, such as Photoshop®, After Effects Pro®, or the open source software GIMP, have been around for a long time. Using such conventional signal processing methods, images can be easily modified, obtaining realistic results that can fool even a careful observer. Figure 2 shows some examples of skillfully manipulated images that have been disseminated on the Internet in recent years to spread false news, both on images[2] and videos[3]. In fact, research in multimedia forensics has been going on for at least 15 years [1], [2], and is receiving ever growing attention, not only from the academy, but also from major information technology (IT) companies and funding agencies. In 2016, the Defense Advanced Research Projects Agency (DARPA) of the U.S. Department of Defense launched the large-scale Media Forensic initiative (MediFor) to foster research on media integrity, with important outcomes in terms of methods and reference datasets.

Following the MediFor taxonomy, digital media verification should look for physical integrity, digital integrity, and semantic integrity. In the literature, several methods have been proposed, which expose physical inconsistencies, concerning

---

L.Verdoliva is with the Department of Industrial Engineering, Università Federico II di Napoli, Naples, Italy. E-mail: verdoliv@unina.it.

[1] https://www.dailymail.co.uk/news/article-2107109/Iconic-Abraham-Lincoln-portrait-revealed-TWO-pictures-stitched-together.html

[2] https://www.cnn.com/2018/03/26/us/emma-gonzalez-photo-doctored-trnd/index.html

[3] https://www.theguardian.com/world/2015/mar/19/i-faked-the-yanis-varoufakis-middle-finger-video-says-german-tv-presenter

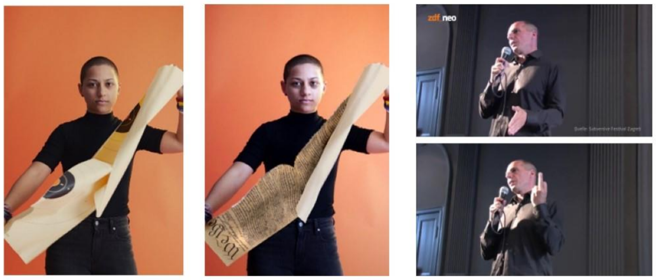

Fig. 2. Examples of fake multimedia where two different versions of an image/video can be retrieved from the web. In one image Emma Gonzalez (left) an american activist is tearing up the American Constitution, while in a video Yanis Varoufakis (right) a greek politician is giving the middle-finger gesture to Germany.

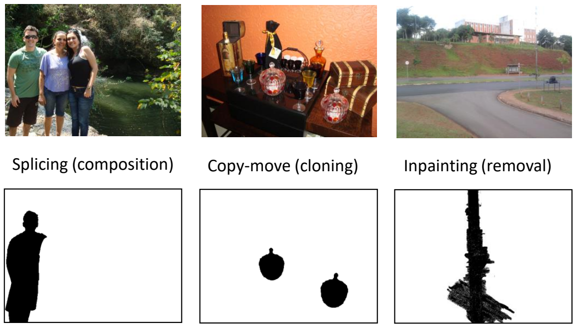

Fig. 3. Examples of image manipulations carried out using conventional media editing tools. Images come from the dataset of the 1st IEEE Image Forensics Challenge organized in 2013. From left to right: splicing (alien material has been inserted in the image), copy-move (an object has been cloned), inpainting (an object has been hidden by background patches).

for example shadows or illumination or perspective [3], [4], [5]. Modern sophisticated manipulations, however, are more and more effective in avoiding such pitfalls and methods which test digital integrity are by far more widespread and represent the current state of the art. Indeed, each image or video is characterized by a number of features, which depend on the different phases of its digital history: from the very same acquisition process, to the internal camera processing (e.g. demosaicing, compression), to all external processing and editing operations [6]. Digital manipulations tend to modify such features, leaving a trail of clues which, although invisible to the eye, can be exploited by pixel-level analysis tools. Instead, semantic integrity is violated when the media asset under analysis conveys information which is not coherent with the context or with evidence coming from correlated sources. For example, when objects are copy-pasted from images available on the web, several near-identical copies can be detected [7], [8], suggesting a possible manipulation. Moreover, by identifying the connections among the various versions of the same asset, it is possible to build its manipulation history (image and video phylogeny) [9], [10].

Despite the continuous research efforts and the numerous forensic tools developed in the past, the advent of deep learning, is changing the rules of the game and asking multimedia forensics for new and timely solutions. This phenomenon is also causing a strong acceleration in multimedia forensics research, which often relies itself on deep learning. There have been several reviews on this topic [11], [12], [6], [13], [14], however these last years have witnessed the advent of new methods. Hence, beyond reviewing the conventional media forensics approaches, a special attention will be devoted to deep learning-based approaches and to the strategies designed to fight deepfakes. The analysis will be restricted to passive methods and visual data-based solutions. That is, it will be assumed that no active strategy is in place to ensure integrity, and that a skilled attacker modified metadata to make them useless, otherwise they would provide precious information towards authenticity verification both for images and videos [15], [16], [17]. On the other hand, it is worth noting that metadata are routinely canceled when media assets are uploaded on a social network.

The review starts with a brief analysis of the most effective manipulation methods proposed in recent years (Section II). Then, integrity verification methods are described, beginning with conventional approaches (Section III), then moving to deep learning-based approaches (Section IV), to conclude with specific deepfake detection methods (Section V). In Section VI, a discussion of the state of multimedia forensics and its perspectives after the advent of deep learning is carried out. A list of the datasets most widespread in the field is presented in Section VII. Then, the further major themes of counterforensics (Section VIII) and fusion (Section IX) are considered. Finally, future research directions are outlined (Section X) and conclusions are drawn (Section XI).

II. FAKE CONTENT GENERATION

There are many ways to manipulate visual content, and new methods are proposed by the day. This Section will briefly review some of the most widespread and promising of them. Very common operations are adding, replicating or removing objects, as in the examples of Figure 3. A new object can be inserted by copying it from a different image (splicing), or from the same image (copy-move). Instead, an existing object can be deleted by extending the background to cover it (inpainting) like in the popular exemplar-based inpainting [33]. All these tasks are easily accomplished with widespread image editing packages. Then, some suitable post-processing, like resizing, rotation or color adjustment, may be required to better fit the object to the scene, both to improve the visual appearance and to guarantee coherent perspective and scale. In recent years, however, the same results are achieved, with better semantic consistency, through advanced computer graphics (CG) approaches and deep learning (see Figure 4, last column). Manipulations that do not require sophisticated artificial intelligence (AI) tools are sometimes referred to as "cheap fakes". Nonetheless, their impact in distorting reality can be very high. For example, by removing, inserting or cloning entire groups of frames one can completely change the meaning of a video. A simple frame-rate reduction was

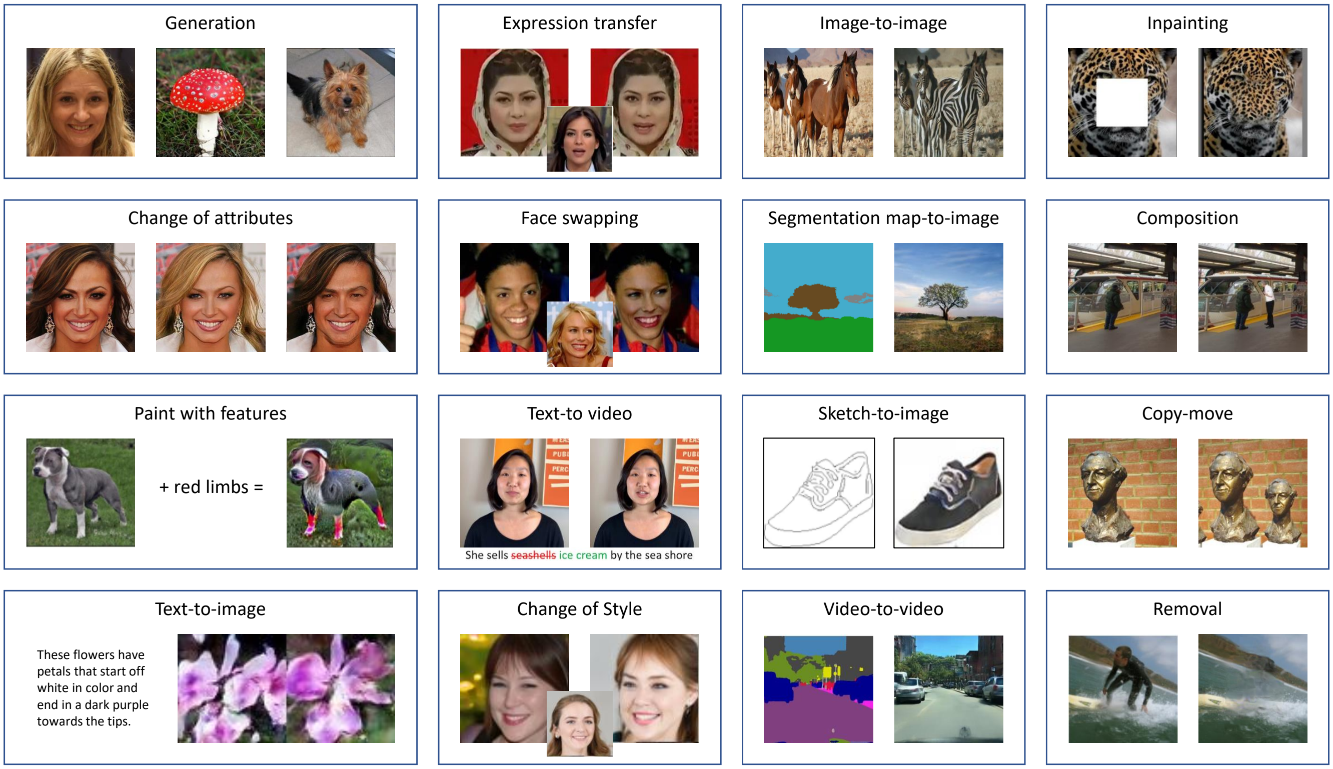

Fig. 4. Examples of image and video manipulations carried out using deep learning methods. Besides conventional manipulations, like composition [18], copy-move [19], object removal [20], inpainting [21], a large number of new tasks can be performed. These include content generation [22], [23], image/video synthesis from semantic labels [24], [25] or sketches [26] or text [27], changes of style and attributes [22], [28], [29], domain translation [30], up to expression transfer [25], face swapping typical of deepfakes [31] and talking-head video editing [32]. It is worth underlining that deep learning methods require no manual media editing on the part of the user, except for possible post-processing.

recently used to let Nancy Pelosi, Speaker of the U.S. House of Representatives, appear as drunk or confused[4].

Besides these "traditional" manipulations, concerning specific areas of the image or video, deep learning and computer graphics are now offering a large number of new ones. First of all, a media asset can be synthesized completely from scratch. To this end, autoencoders and generative adversarial networks allowed to develop successful solutions [34] especially for face synthesis, where a high level of photo-realism has been achieved [35], [22]. It is also possible to generate a completely synthetic image or video using a segmentation map as input [36]. Image synthesis is also achievable using only a sketch [37], [24] or a text description [27]. Likewise, the face of a person can be animated based on an audio input sequence [38], [39]. More often, the manipulation modifies existing images or videos. A well-known example is style transfer [26], [30], which allows to change the style of a painting, switch oranges to apples, or reproduce an image in a different season. Major efforts have been devoted to manipulating faces, for their high semantic value, and for the many possible applications. Methods have been proposed to change the expression of a face [40], [41], to transfer the expression from a source to a target actor [42], [43], or to swap faces [31]. Recently, it has been shown that effective face manipulation is feasible even without a huge amount of training photos of the targeted person [44]. It is even possible to animate the face of a still portrait and express various types of emotions [45]. Beyond faces, some recent work addressed motion transfer: the target person dances following the movements transferred from a source dancer [46]. In Figure 4 some examples of such manipulations are presented. One can easily observe how realistic they appear and the variety of possible automatic editing tools available nowadays.

III. CONVENTIONAL DETECTION METHODS

This Section reviews the major lines of research in multimedia forensics before the emergence of deep learning and deepfakes. The most popular approaches look for artifacts related to the in-camera processing chain (camera-based clues) or the out-camera processing history (editing-based clues) [47]. A defining property of the approaches proposed so far is the prior knowledge they rely upon, which impacts on their suitability for real-world applications. Following this perspective, first, blind methods will be described, where no prior knowledge is required. Then, the focus will shift on one-class methods, which need information only on pristine data, through a collection of images/videos taken from the camera of interest or, more in general, a large set of untampered data. Eventually, supervised methods will be considered, which rely on a suitable training set comprising both pristine and manipulated data.

---

[4] https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/

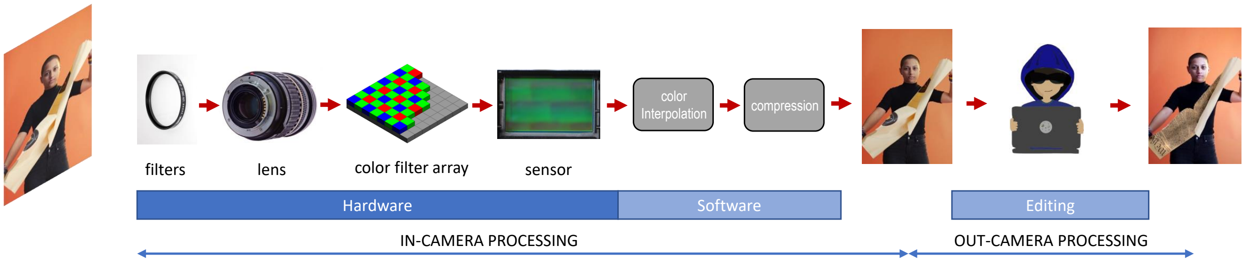

Fig. 5. An image is captured using an acquisition system whose basic components are represented in this figure. After undesired light components are reduced using optical filters, the lenses focus the light on the sensor. In order to extract the red-green-blue (RGB) components, a color filter array (CFA) is present. Each individual sensor element records light only in a certain range of wavelengths. Therefore, the missing color information at a pixel must be recovered from surrounding pixels, through a process known as color filter array interpolation or demosaicing. Then, a sequence of internal processing steps follow, including color correction, enhancement and, finally, compression. The implementation and parametrization of all these components differ based on the camera model and provide important clues that are exploited in the forensic multimedia analysis. Alterations carried out by the malicious user can also introduce artifacts that allow forensic analyses and detection.

A. Blind methods

Blind approaches do not use any external data for training or for other forms of pre-processing: they rely exclusively on the media asset under analysis, and try to reveal anomalies which may suggest the presence of an manipulation. In particular, they look for a number of specific artifacts originated by in-camera or out-camera processing (Figure 5). In fact, the image formation process inside a camera requires a number of operations, both hardware and software, which are specific of each individual camera and leave distinctive traces on the acquired image. For example, the demosaicing algorithm is typically different for different camera models. Therefore, when a manipulation involves the composition of parts of images acquired from different models, demosaicing-related spatial anomalies arise. Likewise, the out-camera editing process may introduce its own peculiar traces, as well as disrupt fingerprint-like camera-specific patterns, phenomena which both allow reliable detection of the attack. Of course, most of these traces are very subtle and cannot be perceived at a visual inspection. However, once properly emphasized, they represent a precious source of information to establish digital integrity (Figure 6).

1) Lens distortion: each camera is equipped with a complex optical system which cannot perfectly focus light at all different wavelengths. These imperfections can be used for forensic purposes. In [48] a method is proposed which exploits the lateral chromatic aberrations, off-axis displacements of the light components at different wavelengths that results in a misalignment between the color channels, while the method proposed in [49] relies on the aberrations generated by the interaction between lens and sensor. Improved versions of the method based on lateral chromatic aberrations, with a more efficient estimation of local displacements, are proposed in [50] and more recently in [51]. Finally, in [52] it is exploited the radial distortion that characterizes the wide-angle lens typically used for indoor/outdoor video surveillance.

2) CFA artifacts: most digital cameras use a color filter array (CFA), with a periodic pattern, so that each individual sensor element records light only in a certain range of wavelengths (i.e. red, green, blue). The missing color information is then interpolated from surrounding pixels, an operation known as demosaicing. This process introduces a subtle periodic correlation pattern in all acquired images. Whenever a manipulation occurs, this periodic pattern is perturbed. In addition, since CFA configuration and interpolation algorithms are specific of each camera model [53], [54], when a region is spliced in a photo taken by another camera model, its periodic pattern will appear anomalous. One of the first methods to exploit these artifacts was proposed by Popescu and Farid [55] back in 2005, based on a simple linear model to capture periodic correlations. Of course, periodic signals produce strong peaks in the Fourier domain. This can be used to distinguish natural images from computer generated images [55], [56], especially after high-pass filtering the image so as to extract more effective features [56], [57]. The problem can be also recast in a Bayesian framework, as proposed in [58], obtaining a probability map in output which allows for fine-grained localization of image tampering. In [59] the analysis is extended to take into account also pixel correlations across color channels.

3) Noise level and noise pattern: a more general approach is to highlight noise artifacts introduced by the whole acquisition process, irrespective of their specific origin. The analysis of *local* noise level may help reveal splicings, as shown in [60], because different cameras are characterized by different intrinsic noise. Local noise analysis has been proposed using statistical tools in the image domain [60], [61] or in the wavelet domain [62]. This approach is also at the basis of the so-called Error Level Analysis (ELA), widely used by practitioners for its simplicity. However, noise intensity alone is not very informative, and may easily provide wrong indications. Therefore, in [63] the high-pass noise residual of the image is used to extract rich features which better characterize

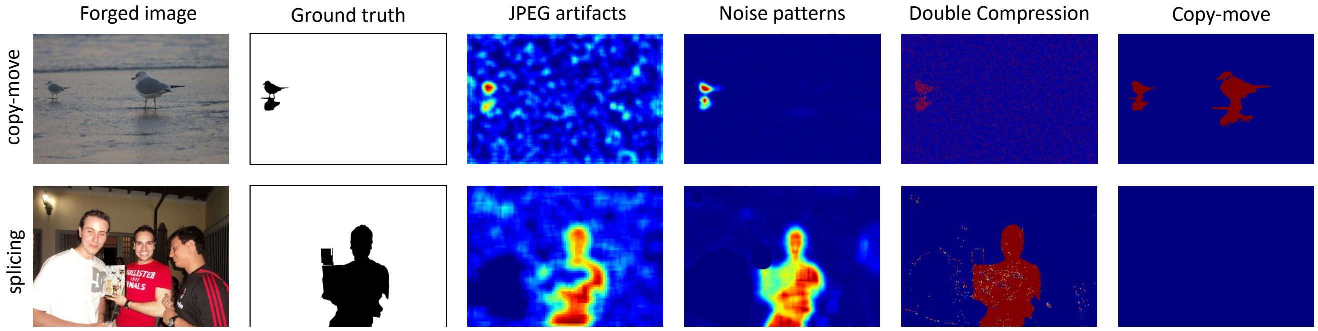

Fig. 6. Localization results of some blind methods for images with copy-move (top) and splicing (bottom). From left to right, manipulated image, ground truth, and localization heatmaps obtained with methods based on JPEG artifacts, noise patterns, double quantization artifacts, copy-move search. Of course, copy-move methods are not effective for splicing manipulations.

local neighborhoods. The expectation-maximization algorithm is then used for clustering these features and reveal possible anomalies. In all above methods, the noise residual is only used to detect possible anomalies.

Departing from this "agnostic" approach, the method proposed in [47] uses the noise residual to estimate the imaging model and define an intrinsic camera fingerprint. Inconsistencies with respect to the estimated model are then used to discover possible manipulations. A similar idea is extended to videos in [64], [65], [66] where the noise residuals of consecutive frames are analyzed and suitable features are extracted to discover traces of both intra-frame and inter-frame manipulations. Instead, in [67], the camera dependent photon shot noise is considered as an alternative fingerprint for static scenes.

4) Compression artifacts: exploiting compression artifacts, has long been a workhorse in image forensics. The many methods proposed in the literature, mostly for JPEG compressed images, can be classified based on the clues they rely upon. A first popular approach is to exploit the so-called lock artifact grid (BAG). Because of the block-wise JPEG processing, discontinuities appear along the block boundaries of compressed images, giving rise to a distinctive and easily detected grid-like pattern [68]. In the presence of splicing or copy-move manipulations, the BAGs of inserted object and host image typically mismatch, enabling detection. Several BAG-based methods have been proposed in the literature [69], [70], [71], some of which even in recent years [72].

Another major approach relies on double compression traces. In fact, when a JPEG-compressed image undergoes a local manipulation and is compressed again, double compression artifacts appear all over the image except in the forged region [73]. These artifacts change depending on whether the two compressions are spatially aligned or not, but suitable detection [74] and localization [75], [76] methods have been proposed for both cases. Another method relies on the so-called JPEG ghosts [77], arising in the manipulated area when two JPEG compressions use the same quality factor (QF). To highlight ghosts, the target image is compressed at all QFs and analyzed. Other methods [78], [79] look for anomalies in the statistical distribution of original DCT samples, assumed to comply with the Benford law.

A further approach is to exploit the model-specific implementations of the JPEG standard, including customized quantization tables and post-processing steps [80], [15]. In [81] model-specific JPEG features have been defined, the JPEG dimples, which depend on how coefficients are converted from real to integer: by the ceil, floor, or rounding operator. Also, chroma subsampling presents specific clues due to integer rounding [82].

Exploiting compression artifacts for detecting video manipulation is also possible, but is much more difficult because of the complexity of the video coding algorithm. Traces of MPEG double compression were first highlighted in the seminal paper by Wang and Farid for detecting frames removal [83]. In fact, the de-syncronization caused by removing a group of frames introduces spikes in the Fourier Transform of the motion vectors. A successive work [84] tried to improve the double compression estimation especially in the more challenging scenario when the strength of the second compression increases and proposed a distinctive footprint, based on the variation of the macroblock prediction types in the reencoded P-frames. This same artifact is exploited in [85] where its estimation is improved to detect traces of inter-frame tampering, and more recently in [86] to deal with video sequences that contain bi-directional frames.

5) Editing artifacts: the manipulation process often generates a trail of precious traces, besides artifacts related to re-compression. Indeed, when a new object is inserted in an image, it typically needs several post-processing steps to fit well the new context. These include geometric transformations, like rotation and scaling, contrast adjustment, but also blurring, to smooth the object-background boundaries. Therefore, many papers focus on detecting these basic operations as a proxy for possible forgeries. Some methods [60], [87] try to detect traces of resampling, always necessary in the presence of rotation or resizing by exploiting periodic artifacts. Other approaches focus on anomalies on the boundaries of objects when a composition is performed [88], or by blur inconsistencies [89].

A very common manipulation consists in copy-moving image regions to duplicate or hide objects. Of course, the presence of identical regions is a strong hint of forgery, but clones are often modified to disguise traces, and near-identical natural objects also exist, which complicate the forensic analysis. Studies on copy-move detection date back to 2003, with the seminal work of Fridrich [90]. Since then, a large literature

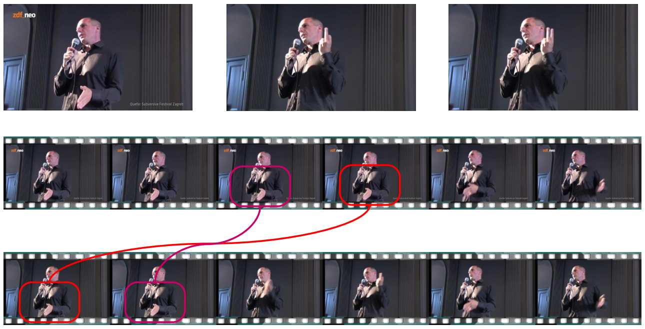

Fig. 7. The Varoufakis video copy-move. Three versions of the same video appeared on the web, with radically different content, as per the sample frames shown on the top row. With reference to the video on the left where the arm is down, a copy-move detector revealed that a sequence from one video (second row) was temporally flipped and copy-moved onto another one (bottom row). For the other two videos the detector was not able to tell apart the pristine from the forged one, since the discriminative region was too small.

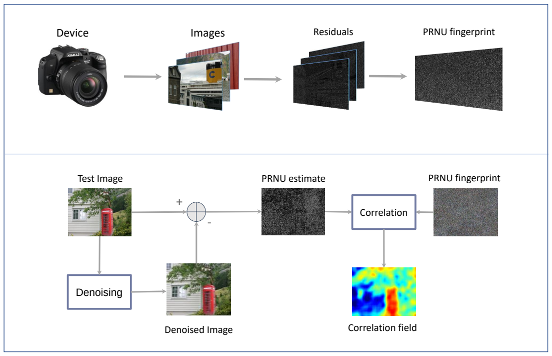

Fig. 8. PRNU-based forgery localization. Top: the device PRNU pattern is estimated by averaging a large number of noise residuals. Bottom: the image PRNU pattern is estimated through denoising, and compared with the reference pattern: the low correlation in the telephone booth region suggests a possible manipulation.

has grown on this topic. Effective and efficient solutions are now available which allow for copy-move detection even in the presence of rotation, resizing, and other geometric distortions [91]. Methods based on keypoints [92], [93] are very efficient, while dense-field methods [94], [95] are more accurate and deal also with occlusive attacks. In [96] dense-field methods have been shown to be effective also to detect inpainting. Extensions to video have been also proposed both for detection and localization [97], [98], the main issue being complexity. In the example of Figure 7, a method for 3D copy-move localization exposed a copy-move with flipping in one of the videos. As for inter-frame forgeries, local modifications can be detected based on the consistency of the velocity field [99].

B. One-class sensor-based and model-based methods

The camera sensor can provide a wealth of precious clues. In fact, due to manufacturing imperfections, the sensor elements present small deviations from their expected behavior. Such deviations form a noise-like pattern, stable in time, called photo-response non-uniformity (PRNU) noise. All images acquired by a given camera bear traces of its PRNU pattern, which can be therefore regarded as a sort of camera fingerprint. If a region of the image is tampered with, the corresponding PRNU pattern is removed, which allows one to detect the manipulation.

PRNU-based forgery detection was first proposed in [100] based on two steps: *i)* the camera PRNU pattern is estimated off-line from a large number of images taken from the camera, and *ii)* the target image PRNU is estimated at test time, by means of a denoising filter, and compared with the reference (see Figure 8). Clearly, this approach relies on some important prior knowledge, since a certain number of images taken from the source device, or the device itself, must be available. On the other hand, it is extremely powerful, as it can detect equally well all attacks, irrespective of their nature. The key problem is the single-image estimation at test time, since the PRNU pattern is a weak signal, easily overwhelmed by imperfectly removed image content. To reduce false alarms, in [101] a predictor is designed to adapt the decision threshold to the local image statistics, while in [102] disturbing non-unique artifacts are detected and removed. In [103] the strong spatial dependencies are modeled through a Markov Random Field so as to make joint rather than isolated decisions. Further variations rely on the use of guided filtering [104], discriminative random fields [105] and multiscale analysis [14].

It is worth noting that this approach can be also extended to blind scenarios, where no prior information about the camera is known provided a suitable clustering procedure identifies the images which share the same PRNU [96], [106].

An alternative to using PRNU is to base the analysis on camera *model* local features. Since cameras of the same model share proprietary design choices for both hardware and software, they will leave similar marks on the acquired images. Therefore, in [107] it was proposed to extract local descriptors from same-model noise residuals to build a reference statistical model. Then, at test time, the same descriptors are extracted in sliding-window modality from the target noise residual and compared with the reference. Strong deviations from the reference statistics suggest the presence of an attack. With respect to PRNU-based analysis, this approach cannot discriminate devices, but only models. On the other hand, model-related artifacts are much stronger than device-related PRNU, and provide more reliable information.

C. Supervised methods with handcrafted features

These methods are based on machine learning. Suitable features are first defined which help discriminating between pristine and manipulated images, and then a classifier is trained on a large number of examples of both types. It is worth underlining that features are *hand-crafted* by the forensic analyst, based on a deep understanding of the target manipulations.

Some features have been devised to detect specific artifacts, especially those generated by double JPEG compression [108], [102], [109] or related to the camera response function (CRF) [110], [111], [112]. However, more precious

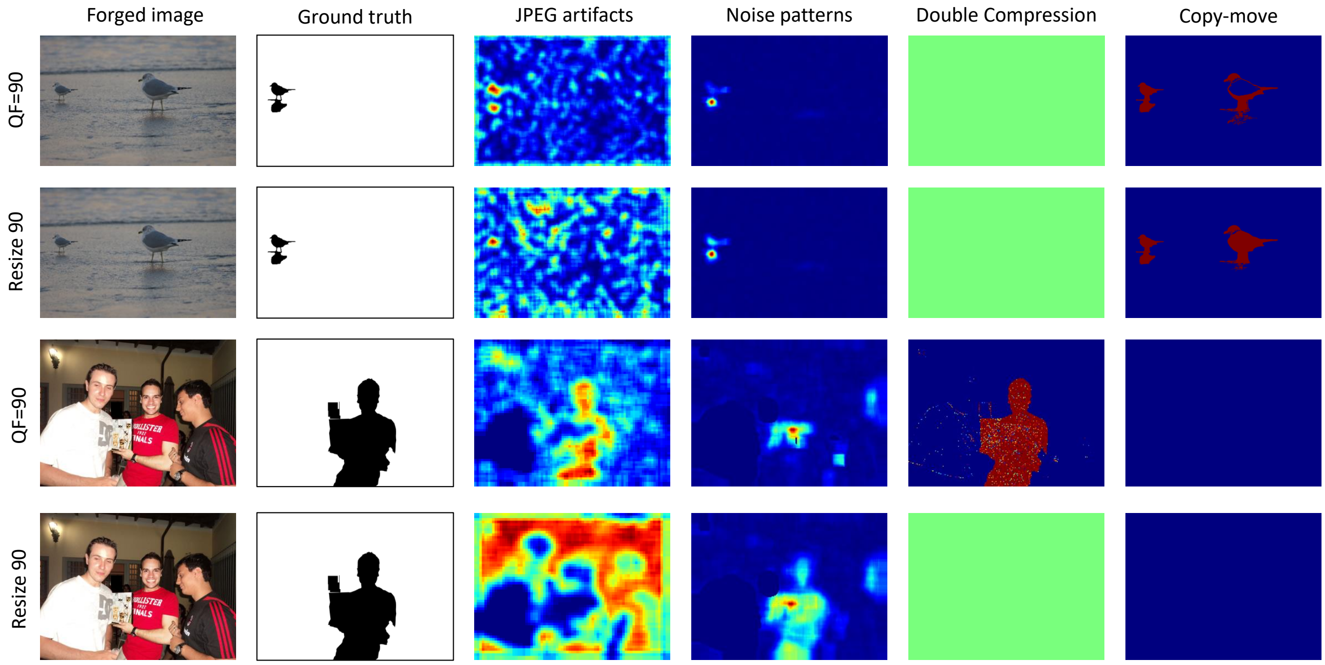

Fig. 9. Localization results of the same blind methods of Figure 6 after compression (QF=90) or resizing (scale=90%). Model-dependent methods, based on JPEG or double quantization artifacts, suffer a much stronger performance impairment than methods based on noise pattern anomalies.

are the *universal* features, based on suitable image statistics, which allow detecting many types of manipulation. Therefore, good statistical models for natural images may help selecting features that guarantee the highest discriminative power. As already observed, to highlight statistical anomalies caused by manipulations, one should first remove the high-level image content, to be regarded as noise [113]. Therefore, the most effective features are typically extracted from noise residuals, either in the spatial [114], [115] or in a transform domain [116], [117]. The pioneering work of Farid and Lyu [118], back in 2003, proved the potential of features based on high-order image statistics. These features capture subtle variations in the image micro-textures and prove effective in many application fields, such as computer graphics, biometrics, and steganalysis. Therefore, it is not by chance that the most popular such features, known as rich models [119], have been originally proposed for steganalysis and later applied with success in forensics. After passing the image through a set of high-pass filters, each one able to capture slightly different artifacts, the features are formed based on co-occurrence of selected neighbors. Then, an ensemble classifier is built. In 2013, two methods [120], [96] based on the fusion of these features and other forensic tools [103], [95] ranked first in both the detection and localization phases of the first IEEE IFS-TC Image Forensics Challenge.

D. Discussion

A major appeal of blind methods is that they do not require further data besides the image/video under test. However, methods based on very specific details depend heavily on their statistical model, and mostly fail when the hypotheses do not hold. With reference to Figure 6, for example, methods based on JPEG artifacts localize correctly both a copy-move and a splicing. However, if the image is slightly compressed (QF=90) or resized (scale 90%), as usual on social networks, their performance drop dramatically, as shown in Figure 9. Copy-move detectors, instead, are more reliable, even in the presence of post-processing, but can only detect cloning and some types of inpainting. On the contrary, methods based on noise patterns are quite general, and robust to post-processing, as they often do not depend on explicit statistical models but look for anomalies in the noise residual. Moreover, to improve reliability, they can be used in a supervised modality, as shown in Figure 10. The analyst can select a suspect region of interest for testing, using the rest of the image as an intrinsic model of pristine data.

As for machine learning-based methods, they can achieve very high detection results: in the 2013 challenge the accuracy was around 94% [120]. However, performance depends heavily on the alignment between training set and test data. It is very high when training and test sets share the same cameras, same types of manipulation, same processing pipeline, like when a single dataset is split in training and test or cross-validation is used. As soon as unrelated datasets are used, the performance drops, sometimes close to 50%, that is, random guess. Lack of robustness limits the applicability of learning based approaches to very specific scenarios.

IV. DEEP LEARNING-BASED APPROACHES

Recently, much attention has been devoted to deep learning-based methods, where features can be directly learnt from the data. Deep learning has proven successful for many computer vision applications, largely advancing the state-of-the-art. Is the same happening in multimedia forensics? How are deep learning ideas and architectures adapted to address the specific challenges of this field? This Section will describe deep

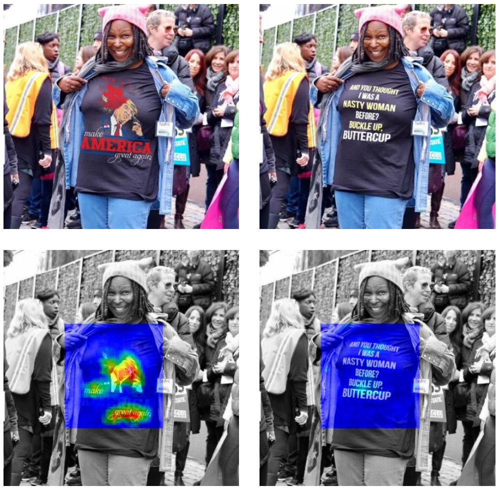

Fig. 10. Using single-asset anomaly-based methods in supervised modality. If only a well-defined region of interest (RoI) might have been manipulated, as with these two versions of Whoopy Goldberg, one can restrict the analysis to the RoI, and use the rest of the image as intrinsic model of pristine data. A strong anomaly appears in the left version, not in the right one.

learning-based methods proposed for the detection of generic manipulations. Then, next Section reviews methods for the detection of GAN-generated images and video deepfakes.

A. Supervised CNNs looking at specific clues

Some papers propose CNN architectures to detect specific artifacts generated by the editing process. Double JPEG compression, as seen in the previous Section, provides strong clues for authenticity verification. So, [121] uses the histograms of DCT coefficients as input to the CNN, while [122] extracts block-wise histogram-related statistical features, so as to enable also localization. To better exploit the peculiarities of forensics as well as the learning ability of CNNs, the approach proposed in [123] works on noise residuals rather than on image pixels and uses the first layers of the CNN to extract histogram-related features. Good results are achieved also when test images are compressed with quality factors (QFs) never seen in training and when conventional methods fail (second QF larger than the first one). To improve performance, [124] uses a multi-domain approach relying on both spatial domain and frequency domain inputs.

Double compression detection has been also extended to H.264 video analysis in [125], where a two-stream neural network is proposed to analyze separately intra-coded frames and predictive frames. A method specifically devised to detect sequence duplication in videos is proposed in [126]. First coarse-level matches between candidate clones are identified, and then a Siamese network based on the ResNet architecture [127] identifies fine-level correspondences.

Other useful editing inconsistencies arise with composition. When material from different sources is spliced together, artifacts can arise at the boundaries. In [128] a multi-task fully convolutional network is devised, which includes a specific branch to detect boundaries between inserted regions and background and another branch for the surface of the manipulation, while in [129] a segmentation network based on U-Net is proposed.

Deep learning methods have been also applied to detect copy-move forgeries. A first solution has been proposed in [130], where an end-to-end CNNs based approach implements the three main steps of a classic copy-move solution, i.e. feature extraction, matching and post-processing to reduce false alarms. This helps to jointly optimize all the modules and gives as output the predicted forgery localization maps. In [131] a different architecture is devised, which includes a multiscale feature analysis and a hierarchical feature matching, which seems to better adapt to different scenarios. Overall, deep learning based methods perform better on low resolution images with respect to conventional approaches, where probably parameters have been adapted to high resolution images. More interestingly it is the analysis carried out in [132], [133], where the problem of source-target disambiguation is faced. In fact, copy-move methods typically generate a map of the original object and its clone, and do not establish which is the forgerd region. To this end, both in [132] and in [133], a two-branch architecture is proposed followed by a fusion module, but the manipulation detection module in [133] focuses on the presence of interpolation artifacts and boundary inconsistencies.

Finally, in [134] a CNN based solution is devised to detect artifacts introduced by a specific Photoshop tool, Face-Aware Liquify, which performs image warping of human faces. The model is trained on fake images automatically generated by the very same tool. To increase robustness, data augmentation includes resizing (bicubic and bilinear), JPEG compression, and various types of histogram editing.

B. Generic supervised CNNs

Generic CNN detectors do not look for specific types of manipulations and artifacts. Of course, training such nets is very challenging, due to the variety of possible attacks and data histories.

A first group of methods take inspiration from the hand-crafted rich models features proposed in [119] and used with success in image forensics. In [135] and in [136], inspired by a similar solution used in [137] for steganalysis, a CNN is proposed with a constrained first layer that performs high-pass filtering of the image, that suppress the scene content and allow to work on residuals. In [135], the fixed rich-model filters are adopted, and the network is used only for feature extraction, followed by a SVM in charge of making the global decision. In [136], instead, constrained filter weights are learnt during training, and the CNN is used also for classification. A slightly different perspective is considered in [138] where a CNN architecture is built to exactly replicate the behaviour of the original rich-model classifier [119]. Then, minor architectural relaxations allow the network to be fine-tuned on domain-specific data and further improved. Therefore, a compact network is obtained which can be trained also on scarce data. All the methods described above allow for the detection of small patches. Localization is then possible by using a sliding window analysis on the whole image.

All these solutions rely mostly on low-level features, assuming that high-level features do not help detecting possible manipulations. However, imperfect image editing, like badly spliced material, may well leave traces, such as strong contrast or unnatural tampered boundaries. Hence, a two-stream network is proposed in [139] and extended in [140]. On the first path, rich model filters are used again to extract low-level features, while a second path relies on the RGB data to look for high-level traces.

Fixed high-pass filters in the first layer have been used also more recently in [141], and in combination with standard filters in [142]. However in both these papers a fully connected network is proposed in order to obtain as output the binary localization map, that hence accounts also for detection at image level. In [141] pixel-wise predictions are obtained by applying transpose convolutions. It is also observed that typically the area of a manipulated region is much smaller than the untouched pixels, and a focal loss is adopted to face this class imbalance, that assigns a modulating factor to the cross entropy term. This paper mostly focuses on manipulations created using deep learning based inpainting methods, while the objective of the work proposed in [142] is to detect every type of possible local manipulations. To this end, the training procedure is built so as to classify 385 image manipulation types, and to detect features related to local anomalies. Another generic solution to detect and localize image manipulations has been proposed in [143]. Resampling features are used to capture inconsistencies, long short-term memory (LSTM) cells highlight transitions between pristine and forged blocks in the frequency domain, and finally an encoder-decoder network segments the manipulation. Localization is indeed carried out in [144] by first introducing a process to generate forged (harder) examples during training in order to generalize across a large variety of possible manipulations. Then, a segmentation and refinement network is used so that the algorithm is forced to look at boundary artifacts.

In [145], [146], [147] the problem of image-level analysis for forgery detection is specifically addressed. Some methods [145], [146] train the CNN to extract compact features at the patch level, and then perform some forms of external aggregation to make image-level decisions. However, a patch level analysis does not allow to take into account at the same time both local (textural analyses) and global (contextual analyses) information. This latter requirement is not easily met, because CNNs accept in input patches that are much smaller than the whole image needed for contextual analysis. In computer vision, this problem is solved by resizing the image, but this process destroys the fine-grain structure of the image and hides important traces of manipulation [146], [147]. In [147] a gradient checkpointing is used to allow end-to-end training of both aggregation and feature extraction, allowing for their joint optimization, without any resizing. This helps analyzing the whole image through textural-sensitive features and highlight anomalies that would not appear at the patch level.

C. *One-class training*

Assembling a training set representative of all possible manipulations can be a prohibitive task. Hence, an alternative is to resort to a one-class approach and look for anomalies with respect to an intrinsic model of pristine data. Indeed, any manipulation is by definition an anomaly, and should be detectable as such. These methods possess then the desirable property to detect any type of manipulations.

In [148] a single-asset (blind) one-class method has been proposed. Expressive features are extracted from the noise residual through an autoencoder, and iterative feature labeling singles out two classes. The largest class defines the pristine model. Then, various criteria can be used to decide on whether the data of the second class are also pristine or else manipulated. In [149], the approach has been extended to videos by including a LSTM recurrent network to account for temporal dependencies. In [150], instead, GANs are used to learn features typical of pristine images, followed by a one-class SVM trained on them to determine their distribution, and eventually detect tampering of satellite images.

Several papers leverage the strong connection existing between source identification and splicing detection and localization. Indeed, in the presence of a splicing, the fact that different image parts are acquired by different camera models provides powerful forensic clues. In [151], a CNN is used to extract camera-model features from image patches, followed by clustering to detect anomalies. A similar approach is followed in [152]. First, a constrained network is used to extract high-level camera-model features, then, another network is trained to learn the similarity between pairs of such features. This work has been recently extended in [153] introducing a graph-based representation that better captures the forensic relationships among all image patches within an image. A Siamese network is also trained in [154] to decide whether two image patches have similar metadata. Once trained on pristine images with EXIF header, the network can be used on any image without further supervision. In [155], [156], [157] these concepts are exploited to extract a camera-model fingerprint, called noiseprint, similar to a PRNU-based device fingerprint. A denoiser CNN is trained in Siamese modality to tell apart similar (same camera model and same position) from different couples of patches. Once trained, the network is able to extract the image-size noiseprint, where artifacts related to camera model are emphasized. In [155] it is shown that noiseprints, thanks to their spatial sensitivity, can be used to detect splicing as well as several other manipulations, while in [157] the approach is extended to video forensics.

V. DEEPFAKE DETECTION

Human faces are by far the most expressive and emotionally-charged pieces of information that circulate on the web. The face is the main biometric trait of a person, a universal ID card, and a vehicle itself of non-verbal but powerful messages. Therefore, the appearance of artificial intelligence-powered tools that generate realistic faces of persons that do not exist, or modify in a credible way the attributes of faces in videos, has raised great alarm.

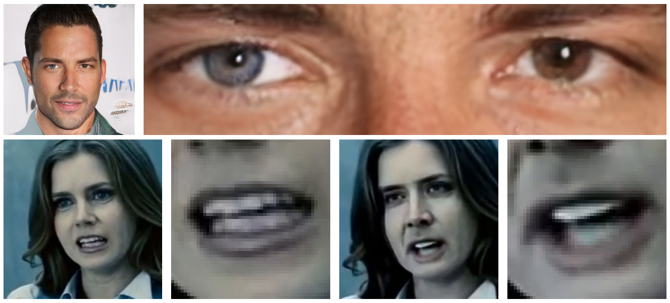

Fig. 11. Today's deepfakes sometimes exhibit some obvious asymmetries, such as eyes of different colors (top) or badly modeled teeth (bottom). However, such artifacts will likely disappear in the future.

However, computer generated faces already existed before the deep learning era. Research on distinguishing real from computer generated faces has been going on for years and represents a precious starting point. Indeed, fakes generated with CGI and deep learning have much in common, since they both lack the characteristic features that are typical of images and videos of human faces acquired by real cameras. In [158] face asymmetry is proposed as a discriminative feature to tell apart computer generated from real images of human faces. Then, in [159] the focus shifts to videos, and detection relies on the spatial-temporal deformations of a 3D model that fits the face. In particular, natural faces follow more complex and various geometric deformations then synthetic ones, and cause higher perturbations of the 3D model. Also, natural faces belong to living persons. So, the method proposed in [160] relies on the small variations in the appearance of the face due to the periodic blood flow caused by heart beating.

The following subsections review the work that has been devoted explicitly to detect local manipulations to images or videos or fully sinthetic media created using deep learning strategies. First, the methods based on handcrafted features will be described, then those relying themselves on deep learning will be analyzed.

A. *Methods based on handcrafted features*

A rather general approach is to look for high-level visual artifacts in the face. Methods following this approach try to highlight specific failures in the generation process which does not reproduce perfectly all the details of a real face. For example, in GAN-generated faces a mismatch may occur between the color of the left and right eye, as well as other forms of asymmetry, like a earring only on one side, or ears with markedly different characteristics. Deepfakes, instead, often present unconvincing specular reflections in the eyes, either missing or represented as white blobs, or roughly modeled teeth, which appear as a single white blob (see Figure 11). All these artifacts are exploited in [161], where simple features are built in order to capture them. [162] relies on eye blinking, which has a specific frequency and duration in humans, which is not replicated in deepfake videos. A solution based on a long-term recurrent network, working only on eye sequences, is designed to catch such temporal inconsistencies. In [163] and in [164] deepfakes are revealed by the lack of variations induced by heart beating, an idea already exploited in [160] for computer generated faces. However, in [163] the coherence of these biological signals is considered both spatially and along the temporal direction.

Other detection methods rely on face warping artifacts [165], face landmark locations [166] or head pose inconsistencies [167]. In [165] the approach exploits the fact that current deepfakes generation methods are able only to generate limited resolution images, that need to be further warped to match the original face in the source video. However, warping leaves peculiar traces that can be detected using a CNN that works on the face region and its surrounding areas. Instead, the observation made in [166] is that GAN-based face synthesis algorithms are able to generate a face with high level of realism and with many details, but lack an explicit constraint over the locations of these parts in a face. Hence, the locations of the facial landmark points, like the tips of the eyes, nose and the mouth, can be used as the discriminative features for verifying the authenticity of GAN images. This same problem is also present in deepfake videos and can be revealed by means of 3D head pose estimation [167].

Other common artifacts of GAN-generated images are related to how color is synthesized. In fact, the generator is constrained so as to limit the occurrence of saturated and under-exposed pixels [168], not infrequent in real images. Other disparities in color components are exploited in [169], in fact deep networks generate images in the RGB color space without any type of constraint on color correlations, and artifacts arise if looking at features in other spaces such as HSV and YCbCr, especially in the chrominance components.

A clear advantage of all these methods is that visual artifacts are not affected by resizing and compression. On the other hand, fake media that can be recognized also by human viewers represent less of a menace. Moreover, with the current pace of technology, it is very likely that next-generation deepfakes will overcome such imperfections and synthesize visually perfect fakes.

A different approach is followed in [170]. The idea is to protect individuals by acquiring some peculiar soft traits that characterize them and are very difficult to reproduce for a generator. In particular, it is observed that facial expressions and head movements are strongly correlated, and changing the former without modifying the latter may expose a manipulation. On the down side, to apply this approach, a large and diverse collection of videos in a wide range of contexts must be available for all individuals of interest.

B. *Methods based on deep learning*

With reference to GAN images, a first investigation has been carried out in [171], where several CNN architectures have been tested in a supervised setting to discriminate GAN images from real ones. Several solutions appear to be very effective, but the performance decreases significantly when training and test mismatch, or when data are compressed using the pipeline typical of social networks. In [172], as a preliminary step for forensic analyses, it is shown that each specific GAN architecture is characterized by its own artificial

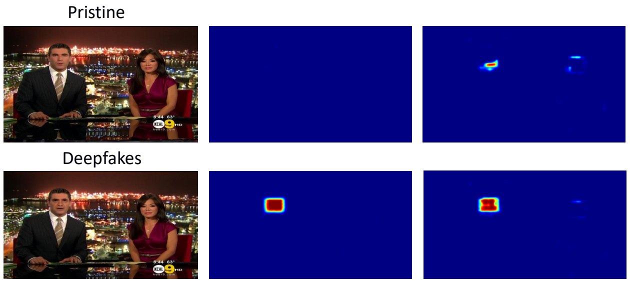

Fig. 12. Localization results provided by a deep network on a pristine video (top) and a deepfake (bottom). Excellent results are obtained with high-quality videos, while artifacts appear if the videos are compressed at low bit-rates.

fingerprint, present in all generated images, much like a real camera is characterized by its PRNU pattern. A similar goal is pursued in [173] through a suitably trained CNN. This last work also shows that these fingerprints persist across different image frequencies and patches, and are not biased by GAN artifacts. In [174], instead, a GAN simulator is proposed to reproduce common GAN-image artifacts, which manifest as spectral peaks in the Fourier domain. Then, a classifier is trained, which takes the spectrum as input. Another work [175] tries to attribute a test image to a specific generator in a white-box scenario. The basic idea is to invert the generation process. Once the original latent vector is recovered, it is fed again to the network, and the output is compared with the test image. A similar idea is also suggested in [176], where using projection-based methods the authors show that it is possible to detect that an image was synthesized by a specific network, even if it is of very high quality.

Switching to solutions for deepfake detection in videos, in [177] two simple architectures are proposed with a small number of levels and parameters that exploit mesoscopic features. A first solution (Meso-4) has four layers of convolutions and pooling and is then followed by a dense network with one hidden layer. The second solution (MesoInception-4) instead is based on a variant of the inception module that includes dilated convolutions. However, experiments carried out in [178] clearly show that, in a supervised setting, very deep general-purpose networks [179], [180], [181], outperform forensics-oriented shallow CNNs, as well as methods based on handcrafted features, especially in the presence of the strong compression typical of video codecs. For the detection task these methods use specific strategies to extract the faces from the frames, as input to the network. However, also localization strategies can be devised on the whole frame. In Fig.12 the pixel-level localization results of a deep network are presented using the approach described in [182], however this analysis can be accomplished in different ways [183]. Training on uncompressed data ensures perfect results on high-quality videos, while artifacts appear when the test video is strongly compressed. Further improvements can be obtained by including an attention mechanism, as done in [184]. Promising results are also obtained in [185] by using capsule-network architectures which require fewer parameters to train than very deep networks. It should be noted that these methods operate on still frames. Therefore, to boost performance, the scores obtained at frame level can be aggregated [177], or else an ensembling approach can be applied, using different color spaces as input [186].

Clearly, better results can be expected by strategies that take explicitly into account the temporal direction. In fact, even if current generation methods are very effective, they perform face manipulation on a frame-by-frame basis and hence may incorrectly follow the face movements. Several methods have been proposed in the literature to exploit this point. In [187], [188] a convolutional Long Short Term Memory (LSTM) network is used to exploit such dependencies and improve upon single-frame analysis. In [189], instead, a solution based on recurrent convolutional models has been proposed. Features are extracted at multiple levels and processed in individual recurrent networks, in order to exploit micro, meso and macroscopic features for manipulation detection. A significant improvement can be achieved even by using only 5 frames. In [190] the optical flow field is estimated to exploit discrepancies in motion across frames, but only some preliminary results are presented.

Despite these interesting results, it must be underlined that most of these detectors tend to overfit the training set, and perform badly on new data [191]. This is becoming a major issue in multimedia forensics, considering the fast growing number of deep learning-based manipulation methods. Therefore, new proposals should ensure good generalization ability, which calls for better validation of detectors, with multiple datasets and different types of manipulations. Focus on this problem is first found in [192], where an autoencoder-based architecture is proposed which adapts to new manipulations with just a few examples. This same approach has been followed in [193] by using a deeper network and including a segmentation task. An autoencoder is also used in [194], where a pixel-wise mask is used to enforce the model to learn intrinsic representation from the forgery region, so as to avoid to detect artifacts present in the training set. In [195] it is proposed a method based on incremental learning to reduce the burden and the risks of re-training networks on larger and larger datasets as new forms of manipulation appear. A different perspective is followed in [196], where a pre-processing step is introduced in order to reduce low level artifacts of GAN images and force the discriminator to learn more general forensic features. Instead, in [197] a careful pre- and post-processing and data augmentation are applied to improve transferability. The work shows that CNN-generated images share some common flaws that allow one to trace their origin even on unseen architectures, datasets and training methods. The main idea is to make a very large augmentation in the training step by means of several and different post-processing operations, like blurring and compression and combinations of them, even if they are not performed at test time.

Turning to videos, recently some interesting solutions have been considered to improve generalization. In [198] this is achieved using hierarchical neural memory networks. Beyond exploiting long-term dependencies, the method includes an attention mechanism and an adversarial training strategy. This helps to increase robustness to compression and to transfer learning to better deal with unseen manipulations. A completely different perspective is followed in [199]. In this work the focus is on the boundary of forged faces. In fact the observation is that they all share a common blending operation. Hence the generalization ability of the network increases, since it is not based on the artifacts of a specific face manipulation method. In [157] generalization is gained by training the method only on pristine videos and by extracting the camera fingerprint information (noiseprint) gathered from multiple frames. This can significantly improve the detection results on various types of face manipulations, even if the network never saw such forgeries in training.

VI. DEEP LEARNING IN MULTIMEDIA FORENSICS: CONSIDERATIONS

The review of previous Sections testifies of an exponential growth in the number of papers proposing deep learning methods for multimedia forensic problems. But, where are we now? Are forensic analysts winning or losing their "war"? Which lessons have we learned? Amidst all these proposals, it is not easy to extract truly innovative ideas, solid scientific trends, effective and robust solutions.

For data-driven methods, lacking theoretical models, experimental results take a fundamental role, and hence experimental protocols are extremely important. Let us consider supervised methods, for the time being, with perfectly aligned training and test sets, that is, disjoint training and test samples drawn from the very same dataset/distribution. Experimental evidence shows that, in this setting, deep learning methods work extremely well. In ideal conditions this is not so important, because then simpler approaches work just equally well. In challenging conditions, however, when forensic traces are weak, deep learning, especially very deep architectures, can provide a large performance gain with respect to conventional methods. In Table I, to compare approaches based on handcrafted features, deep networks, and very deep networks, accuracy results on DeepFakes videos are reported from [178]. In the presence of strong compression, there is a gap of about 15% between machine learning and deep learning, and 15% more using a very deep network. This can make a big difference in practical applications, since many standard processing steps tend to weaken forensic traces, as when video are routinely compressed and resized as soon as they are uploaded on a social network. This is certainly a remarkable achievement of deep learning methods.

However, a validation protocol which considers only perfect alignment is intrinsically weak, and falls short of its goal. Indeed, perfect alignment is a very favorable setting, which is easily obtained in simulations and rarely observed in real-world operations. Main causes of misalignment are: *i)* the target media asset has been generated or manipulated in ways that were never seen in the training set; *ii)* its processing history is not covered by training set samples. Both situations are extremely common. New forms of manipulations are invented by the day, and cannot be represented in the training set when they first appear. Moreover, images and videos can undergo a long sequence of transformations [200] and accounting for all of them is not reasonable. Working well with aligned training and test sets, possibly carved from the same dataset, is not very informative. Therefore, it is important to adopt stronger validation protocols, including experiments where training and test set are unrelated and account for realistic and challenging conditions, such as compression, multiple manipulations, unseen forgeries. In the absence of such strong validation, the results of the supervised deep learning methods do not yet appear completely convincing.

TABLE I
RESULTS OF CNN-BASED METHODS ON DEEPFAKES

	Accuracy		
	Uncompressed	High Quality	Low Quality
Handcrafted features	99.03%	77.12%	65.58%
Deep network	99.28%	90.18%	80.95%
Very deep network	99.59%	98.85%	94.28%

One-class methods seem to perform reasonably well also in challenging real-world conditions, and this looks as a promising approach to deal with such complex scenarios. Clearly, for one-class methods, no alignment problem exists. On the other hand, the fundamental question of what can be labelled as "pristine" remains open. Many operations are not malicious and should not be detected, yet they change the statistics of an image with respect to what is output by a camera. For example, one can apply histogram equalization to improve the appearance of a photo, but that is not a forgery. Even resizing has a different meaning based on the context: it is a sign of a manipulation when used to change the dimension of an object to perform splicing, while it is an innocent operation when used to save space. For these reasons, looking for local anomalies seems to be a good direction, which overcomes ambiguities. In general, the training phase is crucial to make a network work in the correct and desired way, and there is a strong need of large datasets that try to cover many different situations.

VII. DATASETS

For learning based approaches, it is of paramount importance to have good data for training. Moreover, to assess the performance of new proposals, it is important to compare results on multiple datasets with different features. The research community has made considerable efforts through the years to release a number of datasets with image and video manipulations. However, not all of them possess the right properties to support the development of learning-based methods. Many such methods, in fact, split a single dataset in training, validation, and test set, a practice that may easily induce some forms of polarization or over-fitting if the dataset is not built with care. In this Section, the most widespread datasets are described, and their features briefly discussed.

A. Images

Table II reports a list of datasets with manipulated images. Some of them are rather old and necessarily outdated, and some even present important flaws. It is surprising to see recent papers relying on unsuited datasets and presenting them as challenging testbeds.

The Columbia (color) dataset, presented in 2006 [201], is one of the first ones made available to the forensics community. It comprises 180 forged images with splicing, some examples of which are shown in Figure 13. Despite its merits, it presents several major problems: *i)* it is unrealistic, since the forgery is clearly visible; *ii)* the spliced region is not subject to any type of post-processing; *iii)* only uncompressed images are present; *iv)* only four cameras are used to take both the host images and the spliced regions. Moreover, it is not clear how to define the forged areas, given that both regions come from pristine data. For these reasons, this dataset should not be used both in the training and the testing phase, nor to perform fine-tuning, otherwise overly optimistic results will be observed.

Also widespread is the Casia dataset, proposed in 2013 [202]. In the first version (v1) splicings present sharp boundaries and are easily detectable. The second version (v2), however, is more realistic, and inserted objects are post-processed to better fit the scene. Nonetheless, it exhibits a strong polarization, highlighted in [203]. In fact, tampered images and pristine images are JPEG compressed with different quality factors (the former at higher quality). Therefore, a classifier trained to tell tampered and pristine images apart, may instead learn their different processing history, thereby working very well on test images from the same dataset, and very poorly on new unrelated images.

Another dataset with splicings is DSO-1 [5] a subset taken from the IEEE Image Forensics Challenge (unfortunately, the original datasets prepared for the challenge is not available anymore and the ground truths were never released by the organizers). Here, the manipulations are carried out with great care and most of them are realistic. Images are saved in the uncompressed PNG format, but most of them had been JPEG compressed before. Minor problems are the fixed resolution of images, and missing information on how the dataset was created, *e.g.,* how many cameras were used, which could help interpreting results.

Forgeries of various nature are present in the Realistic Tampering Dataset proposed by Korus in [204]. The manipulated images, all uncompressed, appear indeed very realistic, although there is only a small number of them. The dataset includes also the PRNU patterns of the four cameras used to acquire all images, enabling the use of sensor-based methods.

The Wild Web Dataset [205] is a collection of real cases from the internet. Therefore, there are no certified information on the manipulations, but the authors made a huge effort to gather different versions of the same images and to extract meaningful ground-truths.

Many datasets have been proposed specifically for copy-move forgery detection [91], [92], [206], [95], [207], the forgery most studied in the literature. Some of them are designed to challenge copy-move methods by adding multiple operations on the copied object, such as rotation, resizing, change of illumination. However, the more an object is modified, the more detectable it becomes for methods based on camera artifacts, since the distance between its statistical properties and those of the background increases.

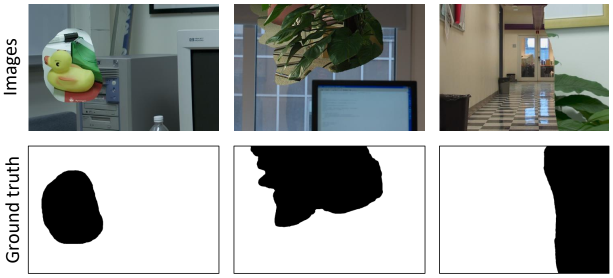

Fig. 13. Examples from the Columbia dataset. Top: images with splicing, bottom: ground truth. In all cases the inserted region is very large and obviously detectable.

There is also a realistic dataset to test double JPEG compression [76] and, recently, a synthetic dataset of single and double JPEG compressed blocks with 1,120 quantization tables, has been released, aimed at training deep networks [122].

To test algorithms in the wild, the U.S. National Institute of Standards and Technology (NIST) has released several large datasets [208]. The first one, NC2016, contains some redundancies: each spliced photo is presented four times, JPEG compressed at low and high quality, and with and without post-processing on the splicing boundaries. These multiple versions are meant to study how performance depends on such details. However, this feature is never exploited in the literature. On the contrary, several papers split the dataset in training and test carelessly, including the same image parts in both sets, and artificially boosting the performance. In subsequent years, NIST published three more datasets, NC2017, MFC2018, MFC2019, without the above redundancies. These datasets are very large and present a great variety of manipulations, resolutions, formats, compression levels, acquisition devices. Moreover, multiple manipulations are often carried out on the same image and even on the same objects. In several cases, a separate development dataset is associated with the main one to ensure correct training of learning-based methods. Overall, they represent a very challenging and reliable testbed for new proposals.

Recently, a very large dataset has been released in [209], called DEFACTO, which collects over 200,000 images with realistic manipulations, including splicings, copy-moves, removals and face morphing. Another very large dataset is the PS-Battles Dataset, where a collection of 102,028 images is presented, each containing the original image but also a varying number of manipulated versions [210]. For what concern facial modifications a dataset has been proposed in [139] which contains around 4,000 real and manipulated images, using two different face swapping algorithms. A large dataset of GAN-generated images using available software [192], [172] are available at [211].

B. Videos

Only a few datasets are available for experiments on videos, but their number has been growing rapidly in this last year. Creating high-quality realistic forged videos using standard

TABLE II
LIST OF DATASETS INCLUDING GENERIC IMAGE MANIPULATIONS

dataset	ref.	year	manipulations	# prist. / forged	image size	format
Columbia gray	[212]	2004	splicing (unrealistic)	933 / 912	128×128	BMP
Columbia color	[201]	2006	splicing (unrealistic)	182 / 180	757×568 - 1152×768	TIF, BMP
MICC F220	[92]	2011	copy-move	110 / 110	722×480 - 800×600	JPG
MICC F2000	[92]	2011	copy-move	1,300 / 700	2048×1536	JPG
VIPP	[76]	2012	double JPEG compres.	68 / 69	300×300 - 3456×5184	JPG
FAU	[91]	2012	copy-move	48 / 48	2362×1581 − 3888×2592	PNG, JPG
Casia v1	[202]	2013	splicing, copy-move	800 / 921	374×256	JPG
Casia v2	[202]	2013	splicing, copy-move	7,200 / 5,123	320×240 − 800×600	JPG, BMP, TIF
DSO-1	[5]	2013	splicing	100 / 100	2048×1536	PNG
CoMoFoD	[206]	2013	copy-move	260 /260	512×512, 3000×2000	PNG, JPG
Wild Web	[213]	2015	real-world cases	90 / 9,657	72×45 − 3000×2222	PNG, BMP, JPG, GIF
GRIP	[95]	2015	copy-move	80 / 80	1024×768	PNG
RTD (Korus)	[204]	2016	splicing, copy-move	220 / 220	1920×1080	TIF
COVERAGE	[207]	2016	copy-move	100 / 100	400×486	TIF
NC2016	[208]	2016	splicing, copy-move, removal	560 / 564	500×500 − 5,616×3,744	JPG
NC2017	[208]	2017	various	2667 / 1410	160×120 − 8000×5320	RAW, PNG, BMP, JPG
FaceSwap	[139]	2017	face swapping	1,758 / 1,927	450×338 - 7360×4912	JPG
MFC2018	[208]	2018	various	14,156 / 3,265	128×104 − 7952×5304	RAW, PNG, BMP, JPG, TIF
PS-Battles	[210]	2018	various	11,142 / 102,028	130×60 − 10,000×8558	PNG, JPG
MFC2019	[214]	2019	various	10,279 / 5,750	160×120 − 2624×19,680	RAW, PNG, BMP, JPG, TIF
DEFACTO	[209]	2019	various	– / 229,000	240×320 − 640×640	TIF
GAN collection	[172]	2019	GAN generated	356,000 / 596,000	256×256 − 1024×1024	PNG

editing tools is very time-consuming, hence, only a few small datasets are available on-line featuring classic manipulations, like copy-moves and splicings [97], [149], [98]. Many more, and much larger datasets include video manipulated with AI-based tools [215], [191], [178], [216], [217], [218], [219], [220] (Table III).

In [215] a face-swapping video dataset, DF-TIMIT, has been built, with 620 deepfake videos obtained with a GAN-based approach. The original data come from a database which contains 10 videos for each of 43 subjects. 16 couples of subjects were manually chosen from the database in order to generate videos with swapped faces from subject one to subject two and viceversa, producing both a low quality and a high quality video. In [191], instead, proposes the Fake Face in the Wild Dataset, FFW, comprising only 150 manipulated videos which, however, show a large variety of approaches, including splicing and CG faces, using both manual effort and completely automatic procedures. Finally, in [221], manipulated videos retrieved from the web have been collected in a dataset that includes 200 fake videos and 180 real videos. An extended version of this dataset also presents near-duplicates found on the web.

The first large dataset with automatically manipulated faces, FaceForensics++, has been proposed in [178]. It contains 1,000 original videos downloaded from the YouTube-8M dataset [222] and 4,000 manipulated videos obtained from them by using four different manipulation tools. Two of them are based on computer-graphics and two on deep learning, two perform changes of expression and two face swapping, Figure 14 shows a few examples. The dataset is available in uncompressed and H264 compressed format, with two quality levels, in order to stimulate developing methods robust to compression. Recently, Google and Jigsaw contributed the dataset with 3,000 more manipulated videos, created *ad hoc* using 28 actors [217]. Also in [216] a new deepfake video dataset has been introduced, called Celeb-DF. It comprises 5,639 manipulated videos, the real videos are based on publicly available YouTube video clips of 59 celebrities of diverse genders, ages, and ethnic groups. Forged videos are created by swapping faces for each pair of the 59 subjects using an improved deepfake synthesis method. Instead in [218] the first release of the dataset used for the Facebook DeepFake Detection Challenge (DFDC) is described. It is composed by 4,113 deepfake videos created using two different synthesis algorithms on the basis of 1,131 original videos featuring 66 enrolled actors. The final dataset made available for the Kaggle competition [219] (started on December 2019) is instead much larger. It comprises 100,000 manipulated videos and around 19,000 pristine ones. A very recent dataset has been built in [220], comprising 10,000 fake videos built using 100 actors and applying 7 perturbations, like color saturation, blurring and compression, with different parameters for a total of 35 possible post-processing so as to

TABLE III
LIST OF DATASETS INCLUDING VIDEO MANIPULATIONS

dataset	ref.	year	manipulations	# prist. / forged	frame size	format
DF-TIMIT	[215]	2018	deepfake	– / 620	64×64 – 128×128	JPG
FFW	[191]	2018	splicing, CGI, deepfake	– / 150	480p, 720p, 1080p	H.264, YouTube
FVC-2018	[221]	2018	real-world cases	2,458 / 3,957	various	various
FaceForensics++	[178]	2019	deepfake, CG-manipulations	1,000 / 4,000	480p, 720p, 1080p	H.264, CRF=0, 23, 40
DDD	[217]	2019	deepfake	363 / 3,068	1080p	H.264, CRF=0, 23, 40
Celeb-DF	[216]	2019	deepfake	– / 5,639	various	MPEG4
DFDC-preview	[218]	2019	deepfake	1,131 / 4,113	180p – 2160p	H.264
DFDC	[219]	2019	deepfake	19,154 / 100,000	240p – 2160p	H.264
DeeperForensics-1.0	[220]	2020	deepfake	50,000 / 10,000	1080p	–

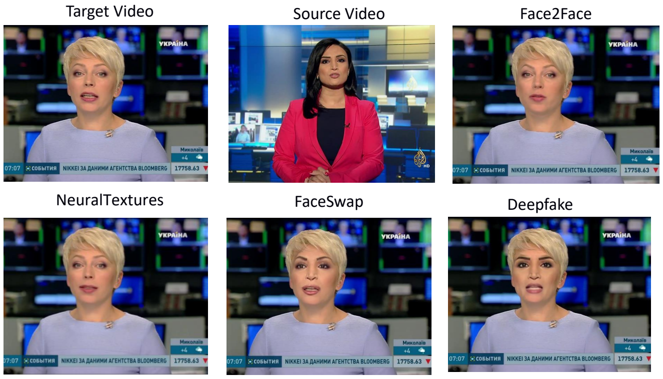

Fig. 14. Example manipulated videos from FaceForensics++. A single original video (top-left) is manipulated by four different tools (Face2Face, NeuralTextures, FaceSwap, DeepFake) using information drawn from a different source video.

better represent a real scenario.

C. Original media

Since some methods work on anomaly detection and are trained only on pristine data, it makes sense to describe also a list of useful publicly available resources based on datasets of authentic images and videos. Of course, they can also be used to simulate different types of manipulation and carry out initial experiments on a synthetic dataset. They are commonly used for source/camera identification, but this task is strictly related to media forgery detection. For example, in case of a composition, by identifying the origin of image patches, one can detect the presence of multiple sources.

The Dresden image database [223] is the most popular one. It contains over 14,000 JPEG images captured by 73 digital cameras of 25 different models. Raw images, instead, can be found in the RAISE dataset [224], composed by 8,156 images taken by 3 different cameras. A further dataset was released for the 2018 Kaggle competition [225] on camera model identification and is also available on-line. It is composed by 2,750 images from 10 different camera models. A dataset that contains SDR (Standard Dynamic Range) and HDR (High Dynamic Range) images has been presented in [226]. A total of 5,415 images were captured by 23 mobile devices in different scenarios under controlled acquisition conditions. The VISION dataset instead [227] includes 34,427 images and 1,914 videos from 35 devices of 11 major brands. Media assets are both in their original format and as they appear after uploading/downloding on various platforms (Facebook, YouTube, WhatsApp) so as to allow studies on data downloaded from social networks. Another mixed dataset, SOCRATES, is proposed in [228]. It contains 6,200 images and 680 videos captured using 67 smartphones of 14 brands and 42 models. Finally, this year the video-ACID database has been published [229], with over 12,000 videos from 46 physical devices of 36 different models.

VIII. COUNTERFORENSICS

In multimedia forensics, like in other security-related fields, one should always account for the presence of an adversary which actively tries to mislead the analyses. In fact, a skilled attacker, aware of the principles on which forensic tools rely, may enact a number of counter-forensic measures on purpose to evade detectors [230]. Forensic tools should prove robust to such attacks, as well as to all real-world conditions that tend to impair the performance observed in laboratory. Therefore, the many counterforensics methods designed to fool current detectors represent a precious help towards the development of multimedia forensics, since they highlight the weaknesses of current solutions and stimulate research for more robust ones [231].

A large body of literature concerns attacks targeted to specific forensic methods, which try to exploit their weaknesses. For example, some methods try to hide traces of resampling that manifest as strong peaks in the Fourier domain [232]. Also sensor traces, and especially PRNU fingerprints, are popular targets because of their importance in forensics. So, methods have been proposed to remove the true device fingerprint from an image, and also to inject the fingerprint of a different device in it [230]. Besides hindering source identification, these attacks can reduce the ability to discriminate manipulated from pristine data. As said before, however, they stimulated the design of more robust detectors [233], and then more powerful attacks [234], in an arms race typical of such two-player games. Attacks to traditional methods will not be explored,

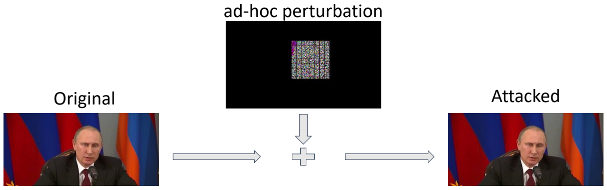

Fig. 15. Fooling deepfake detectors. A suitable inconspicuous adversarial noise pattern can be added to a deepfake video to mislead the CNN detector into classifying it as pristine.

TABLE IV
RESULTS ON FACE2FACE MANIPULATIONS IN THE PRESENCE OF ADVERSARIAL NOISE AND COMPRESSION

	Accuracy		
	Uncompressed	High Quality	Low Quality
no attack	99.93%	98.13%	87.81%
low attack	80.43%	94.83%	85.83%
medium attack	56.37%	89.93%	83.30%
strong attack	52.23%	82.00%	80.30%

referring the reader to a very recent review by Barni *et al.* [235], and focus instead on counterforensics for deep learning-based methods.

Back in 2014, in the context of image recognition, it has been shown [236] that convolutional neural networks are extremely vulnerable to adversarial attacks. By injecting a suitable inconspicuous noise pattern in the target image, the attacker can induce the network to change classification in any desired way (see Figure 15). Actually, this is not a new problem, the vulnerability of machine learning had been known for several years, especially in applications like spam and malware detection, and has been thoroughly studied [237] in the adversarial machine learning field. In any case, these alarming findings have spawned intense research on this topic in computer vision.

In multimedia forensics, some early papers [238], [239] focused on attacking machine learning detectors based on the rich model handcrafted features, obtaining mixed results with high complexity. Indeed, attacking CNNs seems to be easier and more effective [240]. Backpropagation provides a strong help to design gradient-based adversarial perturbations. However, differently from what happens in computer vision [241], it appears that adversarial noise designed to fool a specific CNN architecture does not transfer to different architectures trained for the same task [242], [243]. This is probably because the adversarial noise lives in the same space (high frequencies) where major forensic clues live. Lossy compression is a further intrinsic defence against adversarial attacks [244]. Indeed, strong lossy compression, ubiquitous in real-world scenarios, removes not only useful forensic traces but also adversarial noise, reducing the effectiveness of such attacks [245]. Table IV shows some experimental results on Face2Face manipulated videos [178]. Attacks crafted with FSGM [241] with various strengths ($\epsilon = 1, 2, 3$) have been applied only on faces (so as to not distort the visual quality, Figure 15) and the performance of a CNN-based detector are evaluated. With $\epsilon = 3$ the detector accuracy becomes close to 50% (random choice). However, if videos are compressed, a large part of the adversarial noise is removed, and the detector performance improves again.

Recently, some methods have been specifically developed to attack CNN detectors for multimedia forensics applications. In [246] a GAN-based architecture is proposed to hide the traces of median filtering in order to fool state-of-the-art CNN-based detectors. A different perspective is followed in [247], where a generative method has been proposed to falsify the camera model traces. The attack does not only fool camera model classifiers, but reduces also the power of forensic analyses based on traces of in-camera processing. Along this same direction, in [248] an autoencoder-based method is proposed to remove GAN fingerprints and impair the performance of systems designed to detect GAN-generated images. Instead [249] proposes a GAN-based architecture with a twofold goal, to inject traces of a real camera in synthetic images and, at the same time, reduce peculiar traces of GAN generation. Therefore, attacked images cannot be recognized anymore as computer-generated and are instead recognized as real images of the target camera. The attack takes place in a black-box scenario, with no information on the attacked detectors.

Notably, all these approaches preserve a very good image quality, with no perceivable visual artifacts, demonstrating the urgent need of stronger and more robust detection methods.

IX. FUSION

A system with the ambition to provide reliable decisions about the integrity of images and videos must necessarily integrate multiple tools, to cover most, if not all, the operating conditions of interest. In fact, each individual method works under suitable hypotheses, and can become completely useless when these do not hold. For example, a tool for copy-move detection will not help in case of a splicing. Fusing multiple tools, however, is not only important to widen the spectrum of detectable forgeries, but also to improve the detection capability with respect to each single one. In fact, the traces to be detected are usually extremely weak, and can be easily masked by intentional attacks, as described in the previous Section, and even by standard processing steps. Therefore, the integration of multiple tools designed to detect similar attacks with different approaches, may be expected to improve performance, and especially robustness, to both innocent and malicious disturbances. On the other hand, maximizing the number of clues is standard practice in investigative procedures.

The typical workflow of a forensic tool is to extract suitable low-level features from the original data, process them to obtain a scalar score or a probability vector, and eventually process the latter (score thresholding, max probability choice) to make the final decision. Fusion may take place at all three levels, called feature, measurement, and abstract level, with pros and cons. As observed in [250], working at the feature level presents serious drawbacks when a large number of tools

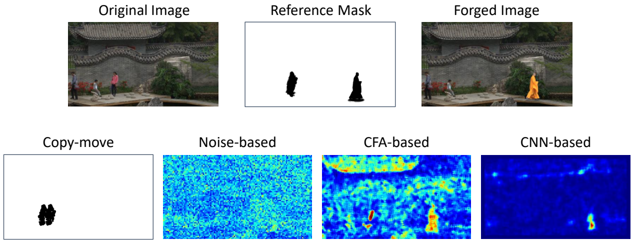

Fig. 16. Using information fusion to detect multiple manipulations. Different tools provide different pieces of information, which may locally agree, disagree, or being complementary. A human expert can likely make sense of all such clues, but transferring this expertise to a computer may be difficult.

are involved, especially for the number of features to deal with, and the problem of creating representative datasets for all possible situations of interest. On the other hand, with abstract-level fusion, precious information may have been already discarded, reducing the ability to exploit cross-tool dependencies. Working at the measurement level may be a reasonable compromise in between these two extremes. In any case, while fusion is certainly a major asset to improve performance, it is not obvious how to combine wildly different pieces of information in a sensible way, as shown by the example of Figure 16. Here, an expert can easily understand that manipulations occurred, and where, but transferring such a skill to a machine is not trivial. Another example is shown in Figure 17, where the two versions of Emma Gonzalez are analyzed using a CNN-based localization approach. For the first image the heatmap shows that there is no manipulation, while for the second image the forged area is highlighted. What is interesting is that a further manipulation was carried out by the malicious user, who also replicated some text inside the poster, which is detected by a copy-move based detector. Indeed, to create a realistic forgeries multiple modifications were needed and hence multiple traces were left. How to fuse these results is clearly not trivial.

Despite the intense research activity in multimedia forensics, there has been limited attention on the fusion of diverse and complementary tools. One of the first contributes is the study of Bayram *et al.* [113] back in 2006. Their results clearly show that detectors based on fusion dominate all individual base detectors in terms of accuracy, and the gain is more significant when operating at the abstract level. In [250], [251] fusion is addressed in a more systematic way, relying on the Dempster-Shafer theory (DST) of evidence [252] to meaningfully combine multiple tools at the measurement level. The DST provides a methodology to include concepts like uncertainty, reliability, and compatibility in the decision process. In fact, to attribute the correct importance to a detector score, one should take into account its overall reliability, the level of confidence associated with the specific decision and its compatibility with other detectors. Experiments in [250], [251] show that the DST approach, with reasonable fusion rules, outperforms consistently all individual tools and also abstract-level fusion, thanks to the use of richer information. It even outperforms machine-learning fusion in the absence of a training set well-aligned with the test set, a recurrent situation in multimedia forensics. In [253] several forensic tools, based on complementary hypotheses, are fused following various strategies, with results confirming that measurement-level fusion is more effective than abstract-level fusion. In [204], instead, contextual information is taken into account by means of prior Markovian models, while in [254] multi-scale fusion is investigated to improve the localization accuracy of tools operating in sliding-window modality.

It is worth mentioning that fusion is a common characteristic of all the winning approaches in several forensic challenges, like in the IEEE Forensics challenge [120], [96], the fraud detection contest on receipts [255] or the Camera Model Identification Challenge organized for the 2018 IEEE Signal Processing Cup and hosted on Kaggle [225]. More in general, in the data science community, it is standard practice to combine the probability vectors output by multiple deep networks, even by simple averaging. Empirically, this fusion improves robustness when moving from training to test data, more and more with the increasing number of networks, in line with the "wisdom of the crowd" principle.

X. FUTURE WORK

As clearly shown by this survey, in the last fifteen years there has been intense research on multimedia forensics, and great progresses. Nonetheless, many issues remain unsolved, new challenges appear by the day and, eventually, most of the road seems to be still ahead of us. This is not so surprising, though. For sure, the advent of deep learning has given extraordinary impulse to both media manipulation methods and forensic tools, opening new research areas. A more fundamental reason, however, is the two-player nature of this research field. The presence of skilled attackers is a guarantee that no tool will protect us forever, and new solutions will be always necessary to cope with unforeseen menaces. With this premise, it is important trying to identify the most promising areas for future research.

A first topic is fusion. As manipulations get smarter and smarter, individual tools will become ever less effective against a wide variety of attacks. Therefore, multiple detection tools, multiple networks, multiple approaches must be put to work together, and how to best combine all available pieces of information should be the object of a more sustained research. Besides multi-tool fusion, also multi-asset analysis should be pursued. More and more, the individual media assets should be analyzed together with all correlated evidence. A picture or a video used to spread a fake news should not be studied in isolation but together with the accompanying text [256], audio [257], and all available contextual information [258]. Also, the approach can be changed based on the availability of additional information, *e.g.*, metadata or near-identical versions of the image/video under test. Eventually, a whole array of semantic-level analyses should be pursued, as envisioned by the recent initiative launched by DARPA on Semantics Forensics.

Focusing more specifically on deep learning-based tools, the main technical issue is probably the (in)ability of deep networks to adapt to situations not seen in the training phase. This issue emerges in several circumstances. First of all, media

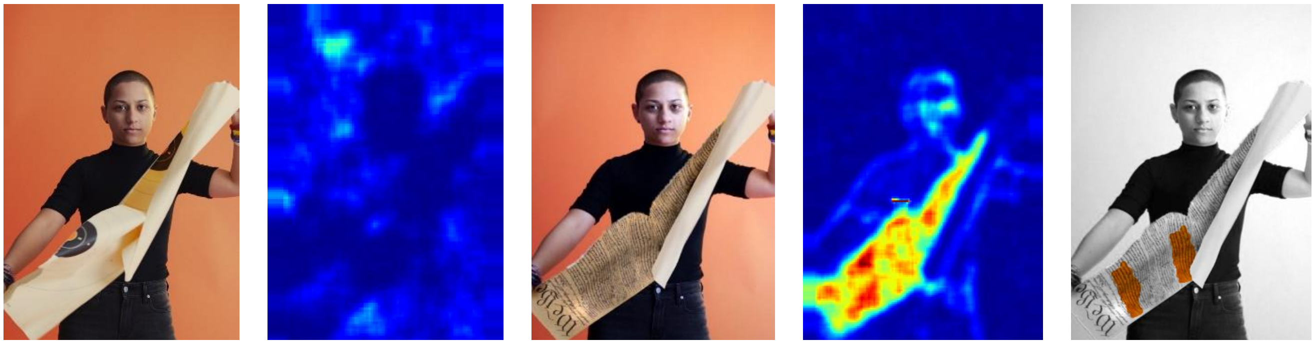

Fig. 17. From left to right: the pristine Emma Gonzalez image, the heatmap obtained using a CNN-based localization method, the manipulated image, and the corresponding heatmap obtained with the same CNN-based detector. The last image shows the output of a copy-move detector operating on the manipulated version. Therefore, multiple clues were found: splicing and cloning. Also in this case deciding the right fusion strategy is not easy.

assets undergo a number of innocent processing steps, like compression, resizing, rotation, re-capturing, etc., that modify significantly their high-order statistics, so precious for forgery detection. Besides innocent transformations, malicious ones should be also considered, designed specifically to disguise forensic clues. It is unlikely that all combinations of such transformations can be represented in a training set. Therefore, higher robustness should be pursued by other means. Also, to deal with the rapid advances in manipulation technology, deep networks should be able to adapt readily to new manipulations, without a full re-training, which may be simply impossible for lack of training data or entail a catastrophic forgetting phenomena.

Another hot issue for deep learning-based methods is interpretability. The black-box nature of deep learning makes it difficult to understand why a certain decision is made. A deep network may correctly classify a cat's picture as “cat”, but we do not know exactly which specific features motivated this decision. Of course, this is a serious issue for some forensic applications. For example, a judge would hardly base decisions only on statistical bases. More in general, being able to track the reasoning of a deep network would allow to improve its design and training phase, and provide higher robustness with respect to malicious attacks.

Lastly, we underline a resurgent of interest on active authentication methods [259], [260]. In past decades, a large body of research was produced on digital watermarking [261]. There is now much interest in blockchain technology [262], in criptography [263], and even new active methods have been proposed to ensure the integrity of digital media [264] or to protect individuals from becoming the victims of AI attacks [265]. As we said before, despite its long history, multimedia forensics appears to be still in full development, with high demands from industry and society and many answers yet to be given.

XI. CONCLUSION

Fifteen years ago multimedia forensics was a niche field of practical interest only for a restricted set of players involved in law enforcement, intelligence, private investigations. Both attacks and defences had an artisan flavor, and required painstaking work and dedication.

Artificial intelligence has largely changed these rules. High-quality fakes now seem to come out from an assembly line calling for an extraordinary effort on part of both scientists and policymakers. In fact, today's multimedia forensics is in full development, major agencies are funding large research initiatives, and scientists form many different field are contributing actively, with fast advances in ideas and tools.

It is difficult to forecast whether such efforts will be able to ensure information integrity in the future, or some forms of active protection will become necessary. This is an arms race, and one part is no smarter than the other. For the present time, a large arsenal of tools is being developed, and knowing them, the principles on which they rely, and their scope of application is a prerequisite to protect institutions and ordinary people.

XII. ACKNOWLEDGEMENT

We gratefully acknowledge the support of this research by a Google Faculty Award. In addition, this material is based on research sponsored by the Air Force Research Laboratory and the Defense Advanced Research Projects Agency under agreement number FA8750-16-2-0204. 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 the Air Force Research Laboratory and the Defense Advanced Research Projects Agency or the U.S. Government.

REFERENCES

[1] H. Farid, “Image forgery detection,” *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 16–25, 2009.

[2] ——, *Photo Forensics*. The MIT Press, 2016.

[3] M. Johnson and H. Farid, “Exposing digital forgeries in complex lighting environments,” *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 450–461, 2007.

[4] E. Kee, J. O'Brien, and H. Farid, “Exposing photo manipulation with inconsistent shadows,” *ACM Transactions on Graphics*, vol. 32, no. 3, pp. 28–58, 2013.

[5] T. de Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. Rocha, “Exposing digital image forgeries by illumination color classification,” *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 7, pp. 1182–1194, 2013.

[6] A. Piva, “An overview on image forensics,” *ISRN Signal Processing*, pp. 1–22, 2012.

[7] Y. Wu, W. Abd-Almageed, and P. Natarajan, "Deep matching and validation network: An end-to-end solution to constrained image splicing localization and detection," in *ACM International Conference on Multimedia*, 2017, pp. 1480–1502.

[8] Y. Lui, X. Zhu, X. Zhao, and Y. Cao, "Adversarial learning for constrained image splicing detection and localization based on atrous convolution," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2551–2566, 2019.

[9] Z. Dias, A. Rocha, and S. Goldenstein, "Video phylogeny: Recovering near-duplicate video relationships," in *IEEE International Workshop on Information Forensics and Security*, 2011.

[10] D. Moreira, A. Bharati, J. Brogan, A. Pinto, M. Parowski, K. W. Bowyer, P. Flynn, A. Rocha, and W. Scheirer, "Image provenance analysis at scale," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 6109–6123, 2018.

[11] A. Rocha, W. Scheirer, T. Boult, and S. Goldenstein, "Vision of the unseen: Current trends and challenges in digital image and video forensics," *ACM Computing Surveys*, vol. 43, no. 4, 2011.

[12] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacchi, and S. Tubaro, "An overview on video forensics," *APSIPA Transactions on Signal and Information Processing*, vol. 1, 2012.

[13] M. Stamm, M. Wu, and K. R. Liu, "Information forensics: An overview of the first decade," *IEEE access*, pp. 167–200, 2011.

[14] P. Korus, "Digital image integrity a survey of protection and verification techniques," *Digital Signal Processing*, vol. 71, pp. 1–26, 2017.

[15] E. Kee, M. Johnson, and H. Farid, "Digital image authentication from JPEG headers," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 1066–1075, 2011.

[16] M. Iuliani, D. Shullani, M. Fontani, S. Meucci, and A. Piva, "A video forensic framework for the unsupervised analysis of MP4-like file container," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 3, pp. 635–645, 2018.

[17] D. Güera, S. Baireddy, P. Bestagini, S. Tubaro, and E. Delp, "We need no pixels: Video manipulation detection using stream descriptors," in *ICML Workshops*, 2019.

[18] F. Tan, C. Bernier, B. Cohen, V. Ordonez, and C. Barnes, "Where and who? automatic semantic-aware person composition," in *IEEE Winter Conference on Applications of Computer Vision*, 2018.

[19] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: image synthesis using neural textures," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, 2019.

[20] R. Shetty, M. Fritz, and B. Schiele, "Adversarial scene editing: Automatic object removal from weak supervision," in *Conference on Neural Information Processing Systems*, 2018.

[21] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6721–6729.

[22] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[23] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *International Conference on Learning Representations*, 2019.

[24] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially adaptive normalization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346.

[25] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro, "Few-shot Video-to-Video Synthesis," in *Neural Information Processing Systems*, 2019.

[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[27] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text-to-image synthesis," in *International Conference on Machine Learning*, 2016.

[28] P.-W. Wu, Y.-J. Lin, C.-H. Chang, E. Chang, and S.-W. Liao, "RelGAN: Multi-Domain Image-to-Image Translation via Relative Attributes," in *International Conference on Computer Vision*, 2019.

[29] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, B. Tran, and A. Madry, "Adversarial robustness as a prior for learned representations," *arXiv preprint arXiv:1906.00945v2*, 2019.

[30] J. Y. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision*, 2017.

[31] Y. Nirkin, I. Masi, A. T. Tran, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," in *IEEE Conference on Automatic Face and Gesture Recognition*, 2018.

[32] O. Fried, A. Tewari, M. Z. abd A. Finkelstein, E. Shechtman, D. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala, "Text-based editing of talking-head video," *ACM Transactions on Graphics*, vol. 38, no. 4, 2019.

[33] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics*, vol. 28, no. 3, 2009.

[34] H. Huang, P. Yu, and C. Wang, "An introduction to image synthesis with generative adversarial nets," *arXiv:1803.04469v2*, 2018.

[35] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.

[36] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Conference on Neural Information Processing Systems*, 2018.

[37] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," *arXiv preprint arXiv:1905.08233v2*, 2019.

[38] S. Suwajanakorn, S. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: learning lip sync from audio," *ACM Transactions on Graphics*, vol. 36, no. 4, 2017.

[39] J. Chung, A. Jamaludin, and A. Zisserman, "You said that?" in *British Machine Vision Conference*, 2017.

[40] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[41] S. Qian, K.-Y. Lin, W. Wu, Y. Liu, Q. Wang, F. Shen, C. Qian, and R. He, "Make a face: Towards arbitrary high fidelity face manipulation," in *IEEE International Conference on Computer Vision*, 2019.

[42] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2016, pp. 2387–2395.

[43] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM Transactions on Graphics (TOG)*, 2018.

[44] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *International Conference on Computer Vision*, 2019.

[45] H. Averbuch-Elor, D. Cohen-Or, J. Kopf, and M. Cohen, "Bringing portraits to life," *ACM Transactions on Graphics*, vol. 36, no. 4, 2017.

[46] C. Chan, S. Ginosar, T. Zhouy, and A. Efros, "Everybody dance now," in *International Conference on Computer Vision*, 2019.

[47] A. Swaminathan, M. Wu, and K. J. R. Liu, "Digital image forensics via intrinsic fingerprints," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 1, pp. 101–117, 2008.

[48] M. Johnson and H. Farid, "Exposing digital forgeries through chromatic aberration," in *Proceedings of the 8th Workshop on Multimedia and Security*, 2006, pp. 48–55.

[49] I. Yerushalmy and H. Hel-Or, "Digital image forgery detection based on lens and sensor aberration," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 71–91, 2011.

[50] T. Gloe, K. Borowk, and A. Winkler, "Efficient estimation and large scale evaluation of lateral chromatic aberration for digital image forensics," in *Proc. SPIE*, vol. 7541, 2010.

[51] O. Mayer and M. Stamm, "Accurate and efficient image forgery detection using lateral chromatic aberration," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 7, pp. 1762–1777, 2018.

[52] H. Fu and X. Cao, "Forgery authentication in extreme wide-angle lens using distortion cue and fake saliency map," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 4, pp. 1301–1314, 2012.

[53] S. Bayram, H. Sencar, N. Memon, and I. Avcibas, "Source camera identification based on CFA interpolation," in *IEEE International Conference on Image Processing*, 2005.

[54] H. Cao and A. Kot, "Accurate detection of demosaicing regularity for digital image forensics," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 5, pp. 899–910, 2009.

[55] A. Popescu and H. Farid, "Exposing digital forgeries in color filter array interpolated images," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3948–3959, 2005.

[56] A. Gallagher and T. Chen, "Image authentication by detecting traces of demosaicing," in *IEEE CVPR Workshops*, 2008.

[57] A. Dirik and N. Memon, "Image tamper detection based on demosaicing artifacts," in *IEEE International Conference on Image Processing*, 2009.

[58] P. Ferrara, T. Bianchi, A. De Rosa, and A. Piva, “Image forgery localization via fine-grained analysis of CFA artifacts,” *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1566–1577, 2012.

[59] J. Ho, O. Au, J. Zhou, and Y. Guo, “Inter-channel demosaicing traces for digital image forensics,” in *IEEE International Conference on Multimedia and Expo*, 2010.

[60] A. Popescu and H. Farid, “Statistical tools for digital forensics,” in *International Workshop on Information Hiding*, 2004, pp. 128–147.

[61] S. Lyu, X. Pan, and X. Zhang, “Exposing region splicing forgeries with blind local noise estimation,” *International Journal of Computer Vision*, vol. 110, no. 2, pp. 202–221, 2014.

[62] B. Mahdian and S. Saic, “Using noise inconsistencies for blind image forensics,” *Image and Vision Computing*, vol. 27, no. 10, pp. 1497–1503, 2009.

[63] D. Cozzolino, G. Poggi, and L. Verdoliva, “Splicebuster: A new blind image splicing detector,” in *IEEE International Workshop on Information Forensics and Security*, 2015, pp. 1–6.

[64] C.-C. Hsu, T.-Y. Hung, C.-W. Lin, and C.-T. Hsu, “Video forgery detection using correlation of noise residue,” in *IEEE International Workshop on Multimedia Signal Processing*, 2008, pp. 170–174.

[65] P. Mullan, D. Cozzolino, L. Verdoliva, and C. Riess, “Residual-based forensic comparison of video sequences,” in *IEEE International Conference on Image Processing*, 2017.

[66] X. Ding, G. Yang, R. Li, L. Zhang, Y. Li, and X. Sun, “Identification of motion-compensated frame rate up-conversion based on residual signals,” *IEEE Trans. Inf. Forensics Security*, vol. 28, no. 7, pp. 1497–1512, 2018.

[67] M. Kobayashi, T. Okabe, and Y. Sato, “Detecting forgery from static scene video based on inconsistency in noise level functions,” *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 4, pp. 883–892, 2010.

[68] Z. Fan and R. de Queiroz, “Identification of bitmap compression history: JPEG detection and quantizer estimation,” *IEEE Transactions on Image Processing*, vol. 12, no. 2, pp. 230–235, 2003.

[69] W. Luo, Z. Qu, J. Huang, and G. Qiu, “A novel method for detecting cropped and recompressed image block,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[70] W. Li, Y. Yuan, and N. Yu, “Passive detection of doctored JPEG image via block artifact grid extraction,” *Signal Processing*, vol. 89, no. 9, pp. 1821–1829, 2009.

[71] Z. Lin, J. He, X. Tang, and C.-K. Tang, “Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis,” *Pattern Recognition*, vol. 42, no. 11, pp. 2492–2501, 2009.

[72] C. Iakovidou, M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, “Content-aware detection of JPEG grid inconsistencies for intuitive image forensics,” *Journal of Visual Communication and Image Representation*, vol. 54, pp. 155 – 170, 2018.

[73] J. Lukáš and J. Fridrich, “Estimation of primary quantization matrix in double compressed JPEG images,” in *Proc. of the 3rd Digital Forensic Research Workshop*, 2003.

[74] Y.-L. Chen and C.-T. Hsu, “Detecting recompression of JPEG images via periodicity analysis of compression artifacts for tampering detection,” *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 2, pp. 396–406, 2011.

[75] M. Barni, A. Costanzo, and L. Sabatini, “Identification of cut & paste tampering by means of double-JPEG detection and image segmentation,” in *IEEE International Symposium on Circuits and Systems*, 2010.

[76] T. Bianchi and A. Piva, “Image forgery localization via block-grained analysis of JPEG artifacts,” *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 1003–1017, 2012.

[77] H. Farid, “Exposing digital forgeries from JPEG ghosts,” *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 1, pp. 154–160, 2009.

[78] D. Fu, Y. Shi, and W. Su, “A generalized Benfords law for JPEG coefficients and its applications in image forensics,” in *Proc. SPIE, Security, Steganography, and Watermarking of Multimedia Contents IX*, 2007.

[79] C. Pasquini, G. Boato, and F. Pèrez-Gonzàlez, “Statistical detection of JPEG traces in digital images in uncompressed formats,” *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 12, pp. 2890–2905, 2017.

[80] H. Farid, “Digital image ballistics from JPEG quantization,” 2006.

[81] S. Agarwal and H. Farid, “Photo forensics from JPEG dimples,” in *IEEE Int. Workshop on Information Forensics and Security*, 2017.

[82] B. Lorch and C. Riess, “Image forensics from chroma subsampling of high-quality JPEG images,” in *Proc. of the ACM Workshop on Information Hiding and Multimedia Security*, 2019.

[83] W. Wang and H. Farid, “Exposing digital forgeries in video by detecting double MPEG compression,” in *Proceedings of the 8th Workshop on Multimedia and Security*, 2006, pp. 37–47.

[84] D. Vázquez-Padín, M. Fontani, T. Bianchi, P. Comesana, A. Piva, and M. Barni, “Detection of video double encoding with GOP size estimation,” in *IEEE International Workshop on Information Forensics and Security*, 2012, pp. 151–156.

[85] A. Gironi, M. Fontani, T. Bianchi, A. Piva, and M. Barni, “A video forensic technique for detecting frame deletion and insertion,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 6226–6230.

[86] D. Vázquez-Padín, M. Fontani, D. Shullani, F. Pérez-González, A. Piva, and M. Barni, “Video Integrity Verification and GOP Size Estimation via Generalized Variation of Prediction Footprint,” *IEEE Transactions on Information Forensics and Security, in press*, 2019.

[87] M. Kirchner, “Fast and reliable resampling detection by spectral analysis of fixed linear predictor residue,” in *10th ACM workshop on Multimedia and security*, 2008, pp. 11–20.

[88] J. Dong, W. Wang, T. Tan, and Y. Shi, “Run-length and edge statistics based approach for image splicing detection,” in *International workshop on Digital Watermarking*, 2006, pp. 177–187.

[89] K. Bahrami, A. Kot, L. Li, and H. Li, “Blurred image splicing localization by exposing blur type inconsistency,” *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 5, pp. 999–1009, 2015.

[90] J. Fridrich, D. Soukal, and J. Lukáš, “ Detection of copy-move forgery in digital images,” in *Proc. of the 3rd Digital Forensic Research Workshop*, 2003.

[91] V. Christlein, C. Riess, J. Jordan, and E. Angelopoulou, “An evaluation of popular copy-move forgery detection approaches,” *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1841–1854, 2012.

[92] I. Amerini, L. Ballan, R. Caldelli, A. D. Bimbo, and G. Serra, “A SIFT-Based Forensic Method for CopyMove Attack Detection and Transformation Recovery,” *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 3, pp. 1099–1110, 2011.

[93] E. Silva, T. Carvalho, A. Ferreira, and A. Rocha, “Going deeper into copy-move forgery detection: Exploring image telltales via multi-scale analysis and voting processes,” *Journal of Visual Communication and Image Representation*, vol. 29, pp. 16–32, 2015.

[94] S.-J. Ryu, M. Kirchner, M.-J. Lee, and H.-K. Lee, “Rotation invariant localization of duplicated image regions based on zernike moments,” *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 8, pp. 1355–1370, August 2013.

[95] D. Cozzolino, G. Poggi, and L. Verdoliva, “Efficient dense-field copy-move forgery detection,” *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 11, pp. 2284–2297, 2015.

[96] D. Cozzolino, D. Gragnaniello, and L. Verdoliva, “Image forgery localization through the fusion of camera-based, feature-based and pixel-based techniques,” in *IEEE International Conference on Image Processing*, 2014, pp. 5302–5306.

[97] P. Bestagini, S. Milani, M. Tagliasacchi, and S. Tubaro, “Local tampering detection in video sequences,” in *IEEE International Workshop on Multimedia Signal Processing*, 2013, pp. 488–493.

[98] L.D’Amiano, D.Cozzolino, G.Poggi, and L.Verdoliva, “A PatchMatch-based dense-field algorithm for video copy-move detection and localization,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 3, pp. 669–682, March 2019.

[99] Y. Wu, X. Jiang, T. Sun, and W. Wang, “Exposing video interframe forgery based on velocity field consistency,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 2674–2678.

[100] J. Lukàš, J. Fridrich, and M. Goljan, “Detecting digital image forgeries using sensor pattern noise,” in *Proc. SPIE*, 2006, pp. 362–372.

[101] M. Chen, J. Fridrich, J. Lukàš, and M. Goljan, “Imaging sensor noise as digital x-ray for revealing forgeries,” in *International Workshop on Information Hiding*, 2007, pp. 342–358.

[102] M. Chen, J. Fridrich, M. Goljan, and J. Lukàš, “Determining image origin and integrity using sensor noise,” *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 4, pp. 74–90, 2008.

[103] G. Chierchia, G. Poggi, C. Sansone, and L. Verdoliva, “A Bayesian-MRF approach for PRNU-based image forgery detection,” *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 4, pp. 554–567, 2014.

[104] G. Chierchia, D. Cozzolino, G. Poggi, C. Sansone, and L. Verdoliva, “Guided filtering for prnu-based localization of small-size image forgeries,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[105] S. Chakraborty and M. Kirchner, “PRNU-based forgery detection with discriminative random fields,” in *International Symposium on Electronic Imaging: Media Watermarking, Security, and Forensics*, 2017.

[106] D. Cozzolino, F. Marra, G. Poggi, C. Sansone, and L. Verdoliva, “PRNU-based forgery localization in a blind scenario,” in *International Conference on Image Analysis and Processing*, 2017, pp. 569–579.

[107] L. Verdoliva, D. Cozzolino, and G. Poggi, “A feature-based approach for image tampering detection and localization,” in *IEEE International Workshop on Information Forensics and Security*, 2014, pp. 149–154.

[108] J. He, Z. Lin, L., and X. Tang, “Detecting doctored JPEG images via DCT coefficient analysis,” in *European Conference on Computer Vision*, 2006, pp. 425–435.

[109] X. Jiang, P. He, T. Sun, F. Xie, and S. Wang, “Detection of double compression with the same coding parameters based on quality degradation mechanism analysis,” *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 1, pp. 170–185, 2018.

[110] Z. Lin, R. Wang, X. Tang, and H.-Y. Shum, “Detecting doctored images using camera response normality and consistency,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 1087–1092.

[111] T. Ng, “Camera response function signature for digital forensics part II: signature extraction,” in *IEEE International Workshop on Information Forensics and Security*, 2009, pp. 161–165.

[112] Y.-F. Hsu and S.-F. Chang, “Camera response functions for image forensics: An automatic algorithm for splicing detection,” *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 4, pp. 816–825, 2010.

[113] S. Bayram, I. Avcibaş, B. Sankur, and N. Memon, “Image manipulation detection,” *Journal of Electronic Imaging*, vol. 15, no. 4, pp. 1–17, 2006.

[114] X. Zhao, S. Wang, S. Li, J. Li, and Q. Yuan, “Image splicing detection based on noncausal Markov model,” in *IEEE International Conference on Image Processing*, 2013.

[115] H. Li, W. Luo, X. Qiu, and J. Huang, “Identification of various image operations using residual-based features,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 1, pp. 31–45, 2018.

[116] S. Lyu and H. Farid, “How realistic is photorealistic?” *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 845–850, 2005.

[117] Z. He, W. Lu, W. Sun, and J. Huang, “Digital image splicing detection based on Markov features in DCT and DWT domain,” *Pattern recognition*, vol. 45, pp. 4292–4299, 2012.

[118] H. Farid and S. Lyu, “Higher-order wavelet statistics and their application to digital forensics,” in *IEEE Workshop on Statistical Analysis in Computer Vision*, 2003, pp. 1–8.

[119] J. Fridrich and J. Kodovsky, “Rich models for steganalysis of digital images,” *IEEE Trans. Inf. Forensics Security*, vol. 7, pp. 868–882, 2012.

[120] D. Cozzolino, D. Gragnaniello, and L. Verdoliva, “Image forgery detection through residual-based local descriptors and block-matching,” in *IEEE International Conference on Image Processing*, 2014, pp. 5297–5301.

[121] Q. Wang and R. Zhang, “Double JPEG compression forensics based on a convolutional neural network,” *EURASIP Journal on Information Security*, pp. 1–12, 2016.

[122] J. Park, D. Cho, W. Ahn, and H.-K. Lee, “Double JPEG detection in mixed JPEG quality factors using deep convolutional neural network,” in *European Conference on Computer Vision*, 2018.

[123] M. Barni, L. Bondi, N. Bonettini, P. Bestagini, A. Costanzo, M. Maggini, B. Tondi, and S. Tubaro, “Aligned and non-aligned double JPEG detection using convolutional neural networks,” *Journal of Visual Communication and Image Representation*, vol. 49, pp. 153–163, 2017.

[124] I. Amerini, T. Uricchio, L. Ballan, and R. Caldelli, “Localization of JPEG double compression through multi-domain convolutional neural networks,” in *IEEE Computer Vision and Pattern Recognition Workshops*, 2017.

[125] S.-H. Nam, J. Park, D. Kim, I.-J. Yu, T.-Y. Kim, and H.-K. Lee, “Two-Stream Network for Detecting Double Compression of H.264 Videos,” in *IEEE International Conference on Image Processing*, 2019.

[126] C. Long, A. Basharat, and A. Hoogs, “A Coarse-to-fine Deep Convolutional Neural Network Framework for Frame Duplication Detection and Localization in Forged Videos,” in *IEEE CVPR Workshops*, 2019.

[127] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[128] R. Salloum, Y. Ren, and C. C. J. Kuo, “Image splicing localization using a multi-task fully convolutional network (MFCN),” *Journal of Visual Communication and Image Representation*, pp. 201–209, 2018.

[129] X. Bi, Y. Wei, B. Xiao, and W. Li, “RRU-Net: The Ringed Residual U-Net for Image Splicing Forgery Detection,” in *IEEE Computer Vision and Pattern Recognition Workshops*, 2019.

[130] Y. Wu, W. Abd-Almageed, and P. Natarajan, “Image copy-move forgery detection via an end-to-end deep neural network,” in *IEEE Winter Conference on Applications of Computer Vision*, 2018.

[131] J.-L. Zhong and C.-M. Pun, “An End-to-End Dense-InceptionNet for Image Copy-Move Forgery Detection,” *IEEE Transactions on Information Forensics and Security, in press*, 2019.

[132] Y. Wu, W. Abd-Almageed, and P. Natarajan, “BusterNet: Detecting copy-move image forgery with source/target localization,” in *European Conference on Computer Vision*, 2018, pp. 170–186.

[133] M. Barni, Q.-T. Phan, and B. Tondi, “Copy Move Source-Target Disambiguation through Multi-Branch CNNs,” *arXiv preprint arXiv:1912.12640v1*, 2019.

[134] S.-Y. Wang, O. Wang, A. Owens, R. Zhang, and A. Efros, “Detecting Photoshopped Faces by Scripting Photoshop,” in *International Conference on Computer Vision*, 2019.

[135] Y. Rao and J. Ni, “A deep learning approach to detection of splicing and copy-move forgeries in images,” in *IEEE International Workshop on Information Forensics and Security*, 2016, pp. 1–6.

[136] B. Bayar and M. Stamm, “A deep learning approach to universal image manipulation detection using a new convolutional layer,” in *ACM Workshop on Information Hiding and Multimedia Security*, 2016.

[137] Y. Qian, J. Dong, W. Wang, and T. Tan, “Deep learning for steganalysis via convolutional neural networks,” in *Proc. SPIE*, vol. 9409-0Y, 2015.

[138] D. Cozzolino, G. Poggi, and L. Verdoliva, “Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection,” in *ACM Workshop on Information Hiding and Multimedia Security*, 2017, pp. 1–6.

[139] P. Zhou, X. Han, V. Morariu, and L. Davis, “Two-stream neural networks for tampered face detection,” in *IEEE CVPR Workshops*, 2017, pp. 1831–1839.

[140] ——, “Learning rich features for image manipulation detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[141] H. Li and J. Huang, “Localization of deep inpainting using high-pass fully convolutional network,” in *IEEE International Conference on Computer Vision*, 2019, pp. 8301–8310.

[142] Y. Wu, W. AbdAlmageed, and P. Natarajan, “ManTra-Net: Manipulation Tracing Network For Detection And Localization of Image Forgeries With Anomalous Features,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[143] J. Bappy, C. Simons, L. Nataraj, B. Manjunath, and A. Roy-Chowdhury, “Hybrid LSTM and EncoderDecoder Architecture for Detection of Image Forgeries,” *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3286–3300, July 2019.

[144] P. Zhou and B.-C. C. X. H. M. Najibi1, “Generate, segment and refine: Towards generic manipulation segmentation,” *arXiv preprint arXiv:1811.09729v3*, 2019.

[145] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, “Distinguishing computer graphics from natural images using convolution neural networks,” in *IEEE Workshop on Information Forensics and Security*, 2017, pp. 1–6.

[146] M. Boroumand and J. Fridrich, “Deep learning for detecting processing history of images,” in *IS&T Electronic Imaging: Media Watermarking, Security, and Forensics*, 2018.

[147] F. Marra and D. Gragnaniello and L. Verdoliva and G. Poggi, “A Full-Image Full-Resolution End-to-End-Trainable CNN Framework for Image Forgery Detection,” *arXiv preprint arXiv:1909.06751*, 2019.

[148] D. Cozzolino and L. Verdoliva, “Single-image splicing localization through autoencoder-based anomaly detection,” in *IEEE Workshop on Information Forensics and Security*, 2016, pp. 1–6.

[149] D. D’Avino, D. Cozzolino, G. Poggi, and L. Verdoliva, “Autoencoder with recurrent neural networks for video forgery detection,” in *IS&T International Symposium on Electronic Imaging: Media Watermarking, Security, and Forensics*, 2017.

[150] S. Yarlagadda, D. Gera, P. Bestagini, F. M. Zhu, S. Tubaro, and E. Delp, “Satellite image forgery detection and localization using gan and one-class classifier,” in *IS&T International Symposium on Electronic Imaging: Media Watermarking, Security, and Forensics*, 2018.

[151] L. Bondi, S. Lameri, D. Güera, P. Bestagini, E. Delp, and S. Tubaro, “Tampering detection and localization through clustering of camera-based CNN features,” in *IEEE CVPR Workshops*, 2017.

[152] O. Mayer and M. Stamm, “Learned forensic source similarity for unknown camera models,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2018, pp. 2012–2016.

[153] ——, “Exposing fake images with forensic similarity graphs,” *arXiv preprint arXiv:1912.02861v1*, 2019.

[154] M. Huh, A. Liu, A. Owens, and A. Efros, “Fighting fake news: Image splice detection via learned self-consistency,” in *European Conference on Computer Vision*, 2018.

[155] D. Cozzolino and L. Verdoliva, “Noiseprint: A CNN-based camera model fingerprint,” *IEEE Trans. Inf. Forensics Security*, vol. 15, no. 1, pp. 14–27, 2020.

[156] ——, “Camera-based image forgery localization using convolutional neural networks,” in *European Signal Processing Conference*, Sep. 2018.

[157] D. Cozzolino, G. Poggi, and L. Verdoliva, “Extracting camera-based fingerprints for video forensics,” in *CVPR Workshops*, 2019, pp. 130–137.

[158] D.-T. Dang-Nguyen, G. Boato, and F. DeNatale, “Discrimination between computer generated and natural human faces based on asymmetry information,” in *European Signal Processing Conference*, 2012, pp. 1234–1238.

[159] D.-T. Dang-Nguyen, G. Boato, and F. De Natale, “3D-model-based video analysis for computer generated faces identification,” *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 8, pp. 1752–1763, Aug 2015.

[160] V. Conotter, E. Bodnari, G. Boato, and H. Farid, “Physiologically-based detection of computer generated faces in video,” in *IEEE International Conference on Image Processing*, Oct 2014, pp. 1–5.

[161] F. Matern, C. Riess, and M. Stamminger, “Exploiting visual artifacts to expose deepfakes and face manipulations,” in *IEEE WACV Workshop on Image and Video Forensics*, 2019.

[162] Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI created fake videos by detecting eye,” in *IEEE Workshop on Information Forensics and Security*, 2018.

[163] U. Ciftci, I. Demir, and L. Yin, “FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals,” *arXiv preprint arXiv:1901.02212v2*, 2019.

[164] S. Fernandes, S. Raj, E. Ortiz, I. Vintila, M. Salter, G. Urosevic, and S. Jha, “Predicting Heart Rate Variations of Deepfake Videos using Neural ODE,” in *ICCV Workshops*, 2019.

[165] Y. Li and S. Lyu, “Exposing deepfake videos by detecting face warping artifacts,” in *IEEE CVPR Workshops*, 2019.

[166] X. Yang, Y. Li, H. Qi, and S. Lyu, “Exposing GAN-synthesized faces using landmark locations,” in *ACM Workshop on Information Hiding and Multimedia Security*, June 2019, pp. 113–118.

[167] X. Yang, Y. Li, and S. Lyu, “Exposing deep fakes using inconsistent head pose,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[168] S. McCloskey and M. Albright, “Detecting GAN-Generated Imagery Using Saturation Cues,” in *IEEE International Conference on Image Processing*, 2019, pp. 4584–4588.

[169] H. Li, B. Li, S. Tan, and J. Huang, “Detection of deep network generated images using disparities in color components,” *arXiv preprint arXiv:1808.07276v1*, 2018.

[170] S. Agarwal and H. Farid, “Protecting world leaders against deep fakes,” in *IEEE CVPR Workshops*, 2018, pp. 38–45.

[171] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, “Detection of GAN-generated fake images over social networks,” in *1st IEEE International Workshop on Fake MultiMedia*, April 2018.

[172] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, “Do GANs leave artificial fingerprints?” in *2nd IEEE International Workshop on Fake MultiMedia*, March 2019.

[173] N. Yu, L. Davis, and M. Fritz, “Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints,” *International Conference on Computer Vision*, 2019.

[174] X. Zhang, S. Karaman, and S.-F. Chang, “Detecting and simulating artifacts in GAN fake images,” in *IEEE Workshop on Information Forensics and Security (WIFS)*, 2019.

[175] M. Albright and S. McCloskey, “Source Generator Attribution via Inversion,” in *CVPR Workshops*, 2019, pp. 96–103.

[176] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and Improving the Image Quality of StyleGAN,” *arXiv preprint arXiv:1912.04958*, 2019.

[177] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “Mesonet: a compact facial video forgery detection network,” in *IEEE International Workshop on Information Forensics and Security*, 2018, pp. 1–7.

[178] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in *International Conference on Computer Vision (ICCV)*, 2019.

[179] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[180] G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger, “Densely connected convolutional networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4700–4708.

[181] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1800–1807.

[182] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics: A large-scale video dataset for forgery detection in human faces,” *arXiv preprint arXiv:1803.09179*, 2018.

[183] J. Li, T. Shen, W. Zhang, H. Ren, D. Zeng, and T. Mei, “Zooming into Face Forensics: A Pixel-level Analysis,” *arXiv preprint arXiv:1912.05790v1*, 2019.

[184] J. Stehouwer, H. Dang, F. Liu, X. Liu, and A. Jain, “On the detection of digital face manipulation,” *arXiv preprint arXiv:1910.01717v1*, 2019.

[185] H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-forensics: using Capsule networks to detect forged images and videos,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[186] P. He, H. Li, and H. Wang, “Detection of fake images via the ensemble of deep representations from multi color spaces,” in *IEEE International Conference on Image Processing*, 2019, pp. 2299–2303.

[187] D. Güera and E. Delp, “Deepfake video detection using recurrent neural networks,” in *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2018.

[188] S. Sohrawardi, A. Chintha, B. Thai, S. Seng, A. Hickerson, R. Ptucha, and M. Wright, “Poster: Towards robust open-world detection of deepfakes,” in *ACM SIGSAC Conference on Computer and Communications Security*, 2019.

[189] E. Sabir, J. Cheng, A. Jaiswal, W. Abd-Almageed, I. Masi, and P. Natarajan, “Recurrent convolutional strategies for face manipulation detection in videos,” in *CVPR Workshops*, 2019.

[190] I. Amerini, L. Galteri, R. Caldelli, and A. D. Bimbo, “Deepfake Video Detection through Optical Flow based CNN,” *ICCV Workshops*, 2019.

[191] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch, “Fake face detection methods: Can they be generalized?” in *International Conference of the Biometrics Special Interest Group*, Sep. 2018.

[192] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, “ForensicTransfer: Weakly-supervised domain adaptation for forgery detection,” *arXiv preprint arXiv:1812.02510*, 2018.

[193] H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, “Multi-task learning for detecting and segmenting manipulated facial images and videos,” in *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2019.

[194] M. Du, S. Pentyala, Y. Li, and X. Hu, “Towards generalizable forgery detection with locality-aware autoencoder,” *arXiv preprint arXiv:1909.05999v1*, 2019.

[195] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, “Incremental learning for the detection and classification of gan-generated images,” in *IEEE Workshop on Information Forensics and Security*, 2019.

[196] X. Xuan, B. Peng, W. Wang, and J. Dong, “On the generalization of GAN image forensics,” in *Chinese Conference on Biometric Recognition*, 2019.

[197] S.-Y. Wang, O. Wang, Z. R, A. Owens, and A. Efros, “CNN-generated images are surprisingly easy to spot... for now,” *arXiv preprint arXiv:1912.11035v1*, 2019.

[198] T. Fernando, C. Fookes, S. Denman, and S. Sridharan, “Exploiting human social cognition for the detection of fake and fraudulent faces via memory networks,” *arXiv preprint arXiv:1911.07844v1*, 2019.

[199] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, “Face X-ray for More General Face Forgery Detection,” *arXiv preprint arXiv:1912.13458v1*, 2019.

[200] T. Bianchi and A. Piva, “Image and video processing history recovery,” in *Handbook of Digital Forensics of Multimedia Data and Devices*, T. Ho and S. Li, Eds. Wiley-IEEE Press, 2015.

[201] Y.-F. Hsu and S.-F. Chang, “Detecting image splicing using geometry invariants and camera characteristics consistency,” in *IEEE International Conference on Multimedia and Expo*, 2006, pp. 549–552.

[202] J. Dong, W. Wang, and T. Tan, “Casia image tampering detection evaluation database,” in *IEEE China Summit and International Conference on Signal and Information Processing*, 2013, pp. 422–426.

[203] G. Cattaneo and G. Roscigno, “A possible pitfall in the experimental analysis of tampering detection algorithms,” in *International Conference on Network-Based Information Systems*, 2014, pp. 279–286.

[204] P. Korus and J. Huang, “Evaluation of random field models in multimodal unsupervised tampering localization,” in *IEEE International Workshop on Information Forensics and Security*, december 2016, pp. 1–6.

[205] M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, "Large-scale evaluation of splicing localization algorithms for web images," *Multimedia Tools and Applications*, vol. 76, no. 4, pp. 4801–4834, 2017.

[206] S. G. D. Tralic, I. Zupancic and M. Grgic, "CoMoFoD - New Database for Copy-Move Forgery Detection," in *Proc. 55th International Symposium ELMAR*, 2013, pp. 49–54.

[207] B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, and S. Winkler, "COVERAGE a novel database for copy-move forgery detection," in *IEEE International Conference on Image processing*, 2016, pp. 161–165.

[208] H. Guan, M. Kozak, E. Robertson, Y. Lee, A. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith, and J. Fiscus, "MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation," in *IEEE WACV Workshops*, 2019, pp. 63–72.

[209] G. Mahfoudi, B. Tajini, F. Retraint, F. Morain-Nicolier, J. Dugelay, B. France, and M. Pic, "DEFACTO: Image and Face Manipulation Dataset," in *European Signal Processing Conference*, 2019.

[210] S. Heller, L. Rossetto, and H. Schuldt, "The PS-Battles Dataset – an Image Collection for Image Manipulation Detection," *CoRR*, vol. abs/1804.04866, 2018. [Online]. Available: http://arxiv.org/abs/1804.04866

[211] "GAN datasets," http://www.grip.unina.it/web-download.html.

[212] T.-T. Ng, and S.-F. Chang, "A data set of authentic and spliced image blocks," 2004.

[213] M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, "Detecting image splicing in the wild (web)," in *IEEE International Conference on Multimedia and Expo Workshops*, 2015.

[214] "MFC2019," https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2019-0.

[215] P. Korshunov and S. Marcel, "DeepFakes: a new threat to face recognition? assessment and detection," *arXiv preprint arXiv:1812.08685v1*, 2018.

[216] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF (v2): A new dataset for deepfake forensics," *arXiv preprint arXiv:1909.12962v3*, 2018.

[217] N. Dufour, A. Gully, P. Karlsson, A.V. Vorbyov, T. Leung, J. Childs and C. Bregler, "Deepfakes Detection Dataset," 2019.

[218] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," *arXiv preprint arXiv:1910.08854v2*, 2019.

[219] "Deepfake Detection Challenge," https://www.kaggle.com/c/deepfake-detection-challenge.

[220] L. Jiang, W. Wu, R. Li, C. Qian, and C. Loy, "DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection," *arXiv preprint arXiv:2001.03024v1*, 2020.

[221] O. Papadopoulou, M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, "A corpus of debunked and verified user-generated videos," *Online Information Review*, 2018.

[222] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8M: A largescale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[223] T. Gloe and R. Böhme, "The 'Dresden Image Database' for benchmarking digital image forensics," in *Proceedings of the 25th Annual ACM Symposium On Applied Computing*, vol. 2, Mar. 2010, pp. 1585–1591.

[224] D. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato, "RAISE: a raw images dataset for digital image forensics," in *6th ACM Multimedia Systems Conference*, 2015, pp. 219–1224.

[225] "Kaggle Camera Model Identification Challenge," https://www.kaggle.com/c/sp-society-camera-model-identification.

[226] O. Al Shaya, P. Yang, R. Ni, Y. Zhao, and A. Piva, "A New Dataset for Source Identification of High Dynamic Range Images," *Sensors*, vol. 18, 2018.

[227] D. Shullani, M. Fontani, M. Iuliani, O. A. Shaya, and A. Piva, "VISION: a video and image dataset for source identification," *EURASIP Journal on Information Security*, pp. 1–16, 2017.

[228] C. Galdi, F. Hartung, and J.-L. Dugelay, "Videos versus still images: asymmetric sensor pattern noise comparison on mobile phones," in *IS&T EI: Media Watermarking, Security and Forensics*, 2017.

[229] B. Hosler, X. Zhao, O. Mayer, C. Chen, J. Shackleford, and M. Stamm, "The Video Authentication and Camera Identification Database: A New Database for Video Forensics," *IEEE Access*, vol. 7, 2019.

[230] T. Gloe, M. Kirchner, A. Winkler, and R. Böhme, "Can we trust digital image forensics?" in *ACM international conference on Multimedia*, 2007, pp. 78–86.

[231] R. Böhme and M. Kirchner, "Counter-forensics: attacking image forensics," in *Digital Image Forensics*, H. Sencar and N. Memon, Eds. Springer, 2012.

[232] M. Kirchner and R. Böhme, "Hiding traces of resampling in digital images," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 582–592, 2008.

[233] M. Goljan, J. Fridrich, and M. Chen, "Defending against fingerprint-copy attack in sensor-based camera identification," *IEEE Trans. Inf. Forensics Security*, vol. 6, pp. 227–236, March 2011.

[234] F. Marra, F. Roli, D. Cozzolino, C. Sansone, and L. Verdoliva, "Attacking the triangle test in sensor-based camera identification," in *IEEE International Conference on Image Processing*, 2014, pp. 5307–5311.

[235] M. Barni, M. Stamm, and B. Tondi, "Adversarial multimedia forensics: Overview and challenges ahead," in *European Signal Processing Conference*, 2018, pp. 962–966.

[236] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[237] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[238] F. Marra, G. Poggi, F. Roli, C. Sansone, and L. Verdoliva, "Counter-forensics in machine learning based forgery detection," in *SPIE Media Watermarking, Security, and Forensics*, 2015.

[239] Z. Chen, B. Tondi, X. Li, R. Ni, Y. Zhao, and M. Barni, "A gradient-based pixel-domain attack against svm detection of global image manipulations," in *IEEE Workshop on Information Forensics and Security*, 2017.

[240] D. Güera, Y. Wang, L. Bondi, P. Bestagini, S. Tubaro, and E. Delp, "A counter-forensic method for CNN-based camera model identification," in *CVPR Workshops*, 2017, pp. 28–35.

[241] I. Goodfellow, J. Shlens, and C. Szegedy, "Intriguing properties of neural networks," in *Explaining and harnessing adversarial examples*, 2015.

[242] D. Gragnaniello, F. Marra, G. Poggi, and L. Verdoliva, "Analysis of adversarial attacks against CNN-based image forgery detectors," in *European Signal Processing Conference*, 2018, pp. 384–389.

[243] M. Barni, K. Kallas, E. Nowroozi, and B. Tondi, "On the transferability of adversarial examples against CNN-based image forensics," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[244] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, S. Li, L. Chen, M. Kounavis, and D. Chau, "SHIELD: Fast, Practical Defense and Vaccination for Deep Learning Using JPEG Compression," in *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery*, 2018, pp. 196–204.

[245] F. Marra, D. Gragnaniello, and L. Verdoliva, "On the vulnerability of deep learning to adversarial attacks for camera model identification," *Signal Processing: Image Communication*, vol. 65, 2018.

[246] D. Kim, H.-U. Jang, S.-M. Mun, S. Choi, and H.-K. Lee, "Median Filtered Image Restoration and Anti-Forensics Using Adversarial Networks," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 278–282, Feb 2018.

[247] C. Chen, X. Zhao, and M. C. Stamm, "Generative adversarial attacks against deep-learning-based camera model identification," *IEEE Trans. Inf. Forensics Security, in press*, October 2019.

[248] J. Neves, R.Tolosana, R. Vera-Rodriguez, V. Lopes, and H. Proenca, "Real or fake? spoofing state-of-the-art face synthesis detection systems," *arXiv preprint arXiv:1911.05351v2*, 2019.

[249] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "SpoC: Spoofing camera fingerprints," *arXiv preprint arxiv.org/abs/1911.12069*, 2019.

[250] M. Fontani, T. Bianchi, A. De Rosa, A. Piva, and M. . Barni, "A framework for decision fusion in image forensics dased on Dempster-Shafer theory of evidence," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 4, pp. 593–607, april 2013.

[251] P. Ferrara, M. Fontani, T. Bianchi, A. De Rosa, and A. Piva, "Unsupervised fusion for forgery localization exploiting background information," in *IEEE International Conference on Multimedia and Expo Workshops*, july 2015, pp. 1–6.

[252] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Ann. Math. Statist.*, vol. 38, pp. 325–339, 1967.

[253] D. Cozzolino, F. Gargiulo, C. Sansone, and L. Verdoliva, "Multiple classifier systems for image forgery detection," in *International Conference on Image Analysis and Processing*, september 2013.

[254] P. Korus and J. Huang, "Multi-scale fusion for improved localization of malicious tampering in digital images," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1312–1326, march 2016.

[255] C. Artaud, N. Sidére, A. Doucet, J.-M. Ogier, and V. Poulain D'Andecy Yooz, "Find it! fraud detection contest report," in *IEEE International Conference on Pattern Recognition*, 2018.

[256] K. Dhruv, G. J. Singh, G. Manish, and V. Vasudeva, "MVAE: Multimodal variational autoencoder for fake news detection," in *ACM World Wide Web Conference*, 2019.

[257] P. Korshunov, M. Halstead, D. Castan, M. Graciarena, M. McLaren, B. Burns, A. Lawson, and S. Marcel, "Tampered speaker inconsistency detection with phonetically aware audio-visual features," in *ICML Workshop on Detecting Audio-Visual Fakes*, 2019.

[258] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, and Y. Kompatsiaris, "Detection and visualization of misleading content on twitter," *International Journal of Multimedia Information Retrieval*, 2017.

[259] P. Singh and H. Farid, "Robust homomorphic image hashing," in *IEEE CVPR Workshops*, 2019.

[260] Y. Zheng, Y. Cao, and C.-H. Chang, "A PUF-based data-device hash for tampered image detection and source camera identification," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 620–634, 2020.

[261] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, *Digital Watermarking and Steganography*. Morgan Kaufmann, 2008.

[262] H. Hasan and K. Salah, "Combating deepfake videos using blockchain and smart contracts," *IEEE Access*, 2019.

[263] D. Boneh, A. Grotto, P. McDaniel, and N. Papernot, "How relevant is the Turing test in the age of sophisbots?" *IEEE Security & Privacy*, vol. 17, pp. 64–71, 2019.

[264] P. Korus and N. Memon, "Content authentication for neural imaging pipelines: End-to-end optimization of photo provenance in complex distribution channels," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8621–8629.

[265] Y. Li, X. Yang, B. Wu, and S. Lyu, "Hiding faces in plain sight: Disrupting AI face synthesis with adversarial perturbations," *arXiv preprint arXiv:1906.09288v1*, 2019.