IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

FORENSICS OF DIGITAL IMAGES AND IDENTIFICATION OF SYNTHETIC MATERIALS

KONDA NAGA SUBBAREDDY
naganagasubbareddy@gmail.com
B.Tech CSE IBM (AI & DS)
C.T UNIVERSITY, PUNJAB

Abstract: The fast development of generative artificial intelligence has made synthetic media production to be democratized, allowing creating deep fakes that could be genuinely realistic and endanger the quality of digital evidence. In this paper, the author evaluates the problems facing digital forensics and the entire judicial framework, whereby the integrity of multimedia data is core to the ruling. Based on an extensive overview of existing types of detection methods, such as convolution neural nets, transformer architectures, and novel forensic paradigms, this study examines their relevance to the process of evidentiary authentication. The phenomenon of anti-forensics, whereby the adversarial methods are used to circumvent detection systems, is also explored in the study leading to an increase in technological arms race. By relying on current expert surveys and significant research projects, including the DETECTOR project conducted by the EU, the presented paper has suggested a multi-layered framework applicable to deep fake evidence authentication, combining technical detection methods with chain-of-custody and forensic examiner training and legal admissibility guidelines. These findings clarify the fact that despite the development of the detection algorithm, to discover a long-term solution, the systems level interventions, including the provenance infrastructure, regulatory models, and institutional adoptions are required to sustain the trust in digital evidence.

1 INTRODUCTION

Deep fake technology has become one of the most important problems in verifying digital evidence in the history of forensic research. After the technology was first referred to during the year 2017 under the name deep fake (a portmanteau of deep learning and fake), it became refined to accomplish more than just dodge facial manipulations; instead, the technology is now able to produce synthetic media that is capable of deceiving human eyes and even automated detection programs. Previously the preserve of special researchers with access to large amounts of computational power it has been democratized in the form of easy use applications and generative AI tools offering people the possibility to create compelling synthetic material at closest to no marginal cost. This technological transformation has a far-reaching impact on the employees of cyber security and digital forensics spheres. The basis statement of digital investigation that adequately preserved layers of digital artifacts can be trusted as valid evidence is becoming undermined by the fact that any multimedia recordings can be completely artificial or malicious in their implementation. As expresses in his critique of the so-called Generative AI Paradox, that societies can start to refuse to believe digital evidence outright as the simulated media streamline nearly everywhere and it is impossible to differentiate between digital and original recordings. The paper will highlight the intersect of deep fake technology and digital forensics in three main goals, to give a general understanding of deep fake generation algorithms and their implications

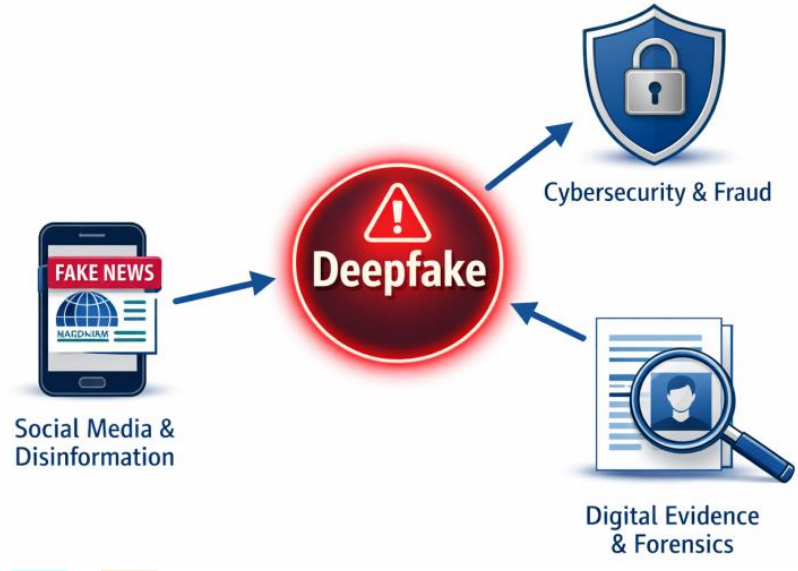

Figure 1: Deep fake Threat Landscape: This figure shows the three main areas that deep fake technology has affected the social/disinformation campaigns, cyber security/fraud applications, and integrity of digital evidence in the process of a forensic investigation. The interdependence shows how synthetic media generates chain of problems at the societal and technical levels.

to the field of evidence authentication, to critically examine existing detection algorithms and their application to forensic processes, and to offer a multi-layered framework of deep fake evidence authentication that would take into account technical, procedural, and legal aspects. This study will focus on visual deep fakes (images and videos) with special focus on face manipulations as faces are the most arduous targets of manipulations owing to their ample furors as identity, emotion, and biometric information carriers. Although audio deep fakes pose just as severe difficulties, they are discussed here as it is through the prism of audio-visual fusion and multimodal features.

2 BACKGROUNDS and Literature Review

2.1 The History of Deep fake Technology

Deep fake generation uses a deep learning architecture, which has shown an incremental increase of its ability to produce realistically synthesized content. Variation Auto encoders (VAE) and Generative Adversarial Networks (GAN) are the primary generative models in addition to diffusion models which are relatively new models. Each is a unique methodological solution with provide implications to the quality of generation and detestability levels. GANs, introduced by , popularized deep fakes in particular have been influenced by this especially harshly. The architecture has trained a generator and a discriminator through a mutual training process which gradually adds more and more actuality to the output. The refinements of GAN, namely ProGAN, StyleGAN, and StyleGAN2 have demonstrated an incredible ability to produce photorealistic faces that cannot be easily differentiated by people among authentic photographs. The most recent is the diffusion models, which involves a process that introduces noise to training images and is trained to reverse this corruption, now it is possible to manipulate the parameters of generation at a fine level

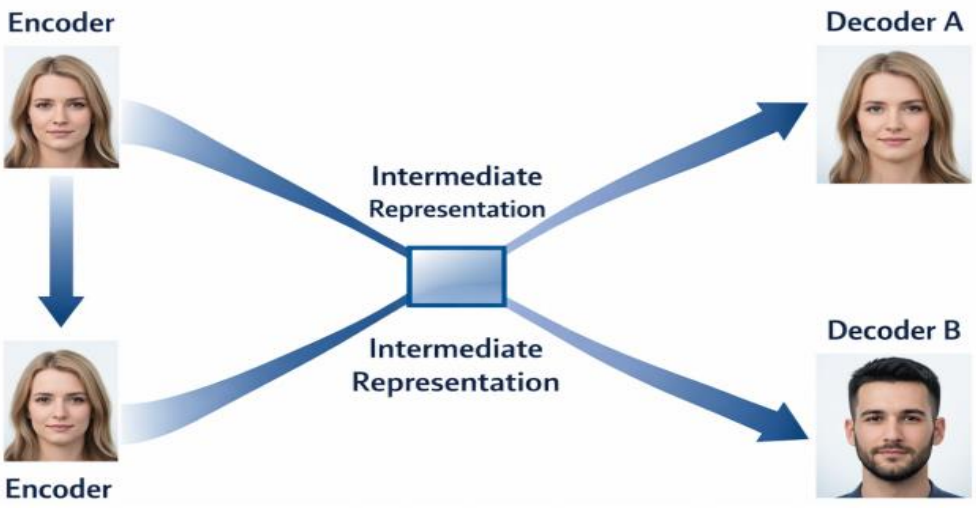

Figure 2: Deep fake Generation - Auto encoder Architecture: It is an hourglass architecture auto encoder that consists of an encoder (compression) and intermediate representation (latent space), and deck structures. The face swapping paired decoders permit the identity exchange between the target and source face.

The types of facial manipulations that can be performed with the help of these technologies are four major ones:

• Whole face synthesis results in completely novel faces that do not belong to anyone. This method is used in entertainment and creation of the digital avatar, as well as in the creation of the synthetic identity in order to commit fraud and run a disinformation campaign.

• Identity swap Identity switch It is applied to replace a face with a second one in images, photographs or videos. This is the best-known version of deep fake, which allows having situations when a person can sound or act in a way he did not do before.

• Attribute manipulation refers to the process of altering such individual features as age, sex, hair color or the presence of any accessories like glasses. Although these manipulations appear to be innocent, they may be manipulated to produce false evidence regarding the description of the witnesses or the appearance of a suspect.

I Face reenactment or expression swaps Face expression and movements are transferred onto a target individual to allow generating videos in which the emotional reaction of a person is completely artificial.

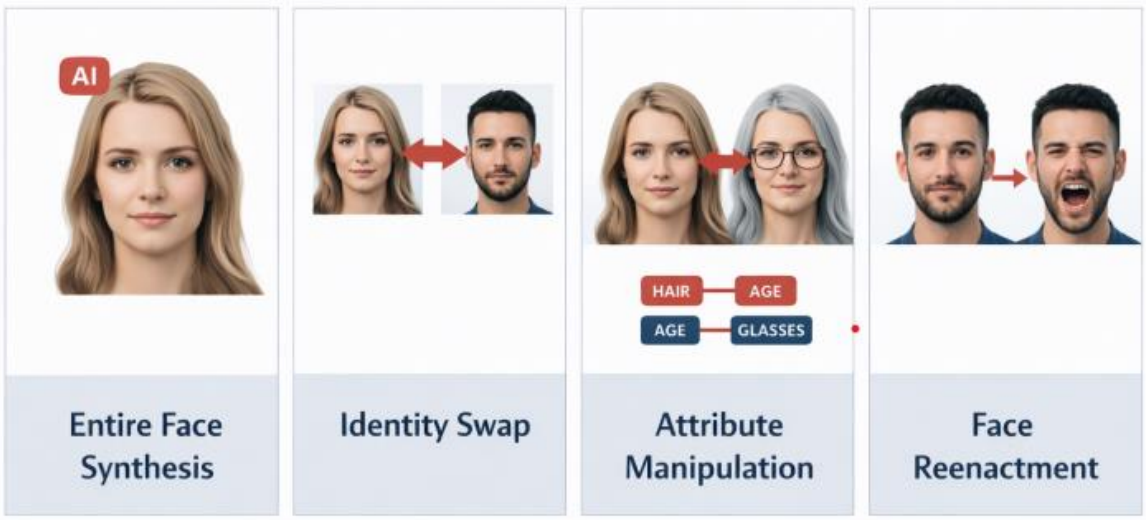

Figure 3: Taxonomy of Facial Manipulations: A taxonomy figure of a four-panel taxonomy illustrating the taxonomy of the kinds of facial deep fakes that include synthesizing a complete face, identity swap, attribute manipulation, or re-enacting a face. The type of each manipulation has its own challenge to forensic authentication.

2.2 The Forensic Challenge

The penetration of the deep fake technology into the threat environment has established what the European Union-based DETECTOR project defines as the increasing difficulty in determining the credibility of digital evidence by law enforcement agencies and criminal labs. Such a challenge exists on several levels: On technical level, the forensic examiners have to discriminate between genuine and artificial media, which is a more challenging task as a more sophisticated method of generation is introduced. Because the technology is adversarial, it would be possible that detection methods that are effective on a particular generation of deep fakes will not be effective on the next generation. From a procedural standpoint, the proven forensic protocols of gaining, preserving and analyzing evidence should be changed to suit synthetic media threats. No matter how much the need of chain of-custody documentation is true, on its own, it will never help to prove the authenticity of the content created synthetically at the time of its creation. On a legal front, the courts need to decide what is the standard of proof that is required in regards to the authenticity of digital evidence as well as the weight that should be placed on expert testimony concerning the findings of detection against the likelihood of false positives and false negatives.

2.3 The Anti-Forensics Dimension

One of the most worrying trends is the appearance of anti-forensic methods that are specifically aimed at avoiding deep fake's detection. As article in their overarching analysis, adversarial examples on deep fake detectors are an increasing danger to the accuracy of forensic examination. These attacks use the vulnerability of the detection models by crafting the perturbation based on carefully designed perturbations that lead to misclassification yet are not noticeable to human eyes. The adversarial environment includes white-box attacks, where the attackers are fully aware of the detection model, and black-box attacks, where detection has to be evaded by query-based exploration or transferable adversarial example. Defending methods include adversarial training whereby the models are trained on adversarial examples, as well as ensemble methods which fuse multiple detection methods and thus make them more resilient. This result in what researchers have termed as a vicious cycle, whereby there is increased detection that is countered by more advanced generation and evasion methodologies and that thus; there is a lack of stagnation in the forensic approach.

3 The State-of-The-Art Deep fake Detectors

3.1 Detection apropos Detection

Approaches Taxonomy the studies on deep fake detection have yielded a collection of a wide range of approaches that can be divided into various dimensions. In line with the guidelines created in the recent overall reviews, here the detection methods are categorized according to their inherent architectural paradigms and principles of operations.

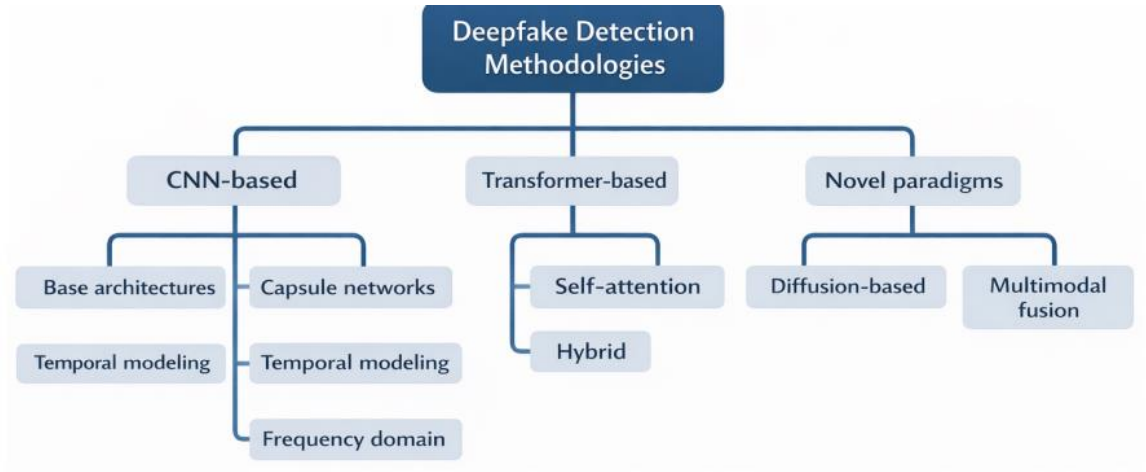

Figure 4: Detection Methodology Taxonomy: A hierarchical tree map representing the classification of detection approaches with CNN-based approaches at the base, their off-shoot transformer networks under them and the latest innovative paradigms.

3.2 Convolution Neural Network Solutions

Most of the deep fake detection systems have been based on Convolution Neural Networks (CNNs), because they have been shown to be effective in image classification activities. The first CNN-based detectors that were optimized to the binary classification task of real and fake face classification were implemented on object recognition architectures, e.g. XceptionNet and Efficient Net. CNN-based detectors become effective because they are capable of learning spatial details which distinguish between real and fake images. These properties are susceptible to some artifacts introduced during the generation process, including differences in eye reflections, sleekness in the frequency domain representation or other unnatural texture in facial boundaries. Extensions add the repeat-able neural network components of LSTM or GRU layers to process sequence of frames of video data. These methods take advantage of the fact that individual frames that are already very realistic can show temporal inconsistencies over time, e.g. irregular blinking within an image or changes in the geometry of faces across frames.

CNN-detection has been advanced to frequency domain techniques. These methods are capable of revealing artifacts that cannot be seen spatially, and are visible in the frequency domain since they convert images to frequency representations such as Discrete Cosine Transform or Fast Fourier Transform. Studies have shown that images created using GAN have unique frequency characteristics that can be used as effective features of detection.

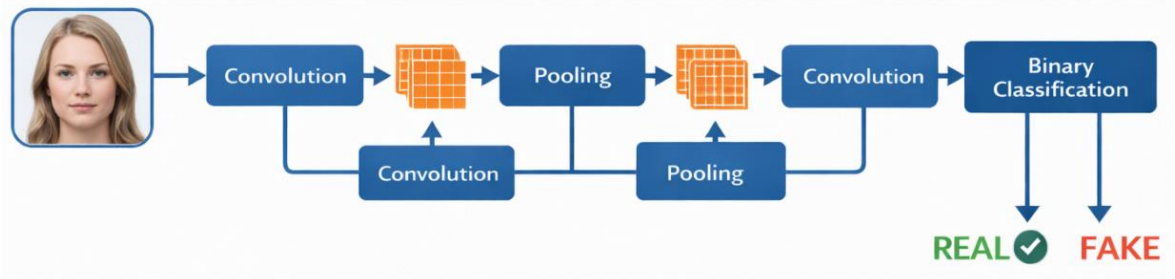

Figure 5: CNN-Based Detection Workflow: This is a flowchart that shows the process to be taken beginning with the input face image up to the final classification made to the real/ fake decision with the use of convolution layers, feature extraction, and feature usage. The hierarchical feature learning process is illustrated in the visualizations between layers in the form of the feature maps.

3.3 Transformer-Based Architectures

Transformer architecture which was initially used in natural language processing recently saw applications in deep fake detection with good outcomes. Self-attention mechanism which allows transformers to represent

long range dependencies in text is equally useful to capture global contextual relationships in images. In contrast to CNNs that operate on images by local receptive fields, and hierarchical pooling, transformers have the ability to pay direct attention to any region of an image, making it possible to detect manipulations involving both the concerted action of spatially separated facial features. This universal view adds the CNN-based methods and can be beneficial in the detection of particular categories of deep fakes. Research has identified that transformer-based detectors are competitive or even better at performance when compared to CNN baselines particularly on cross-dataset evaluation able to test the generalization to manipulation methods that have not been observed. This stalemate can be indicative of less dependence of transformers on dataset-specific artifact patterns.

3.4 Novel Detection Paradigms

New methods of detection are considering other paradigms other than the traditional supervised classification. Diffusion-based detection doesn't rely on the same deep fakes generative modeling technology, just in reverse: by attempting to reconstruct an image through diffusion, and measuring the quality of a reconstruction with fidelity as a metric, such techniques may be used as an estimate of whether a particular image is a deep fake or not . The other boundary is that of multimodal detection that recognizes that even advanced form of deep fakes must remain consistent across sound and video streams. Through an analysis of the synchronization patterns, the lips movements with the speech and the cross-modal features correlation, multimodal detectors can detect inconsistencies that cannot be measured by single modality analysis.

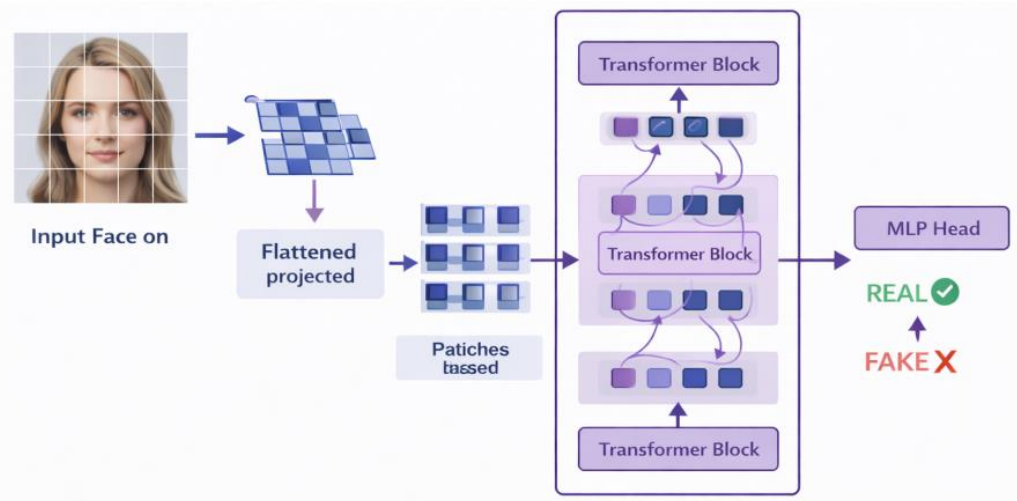

Figure 6: Transformer-Based Detection Architecture: Vision Transformer (ViT) architecture input face is divided into patches, patch embedding, transformer blocks with self-attention attachments and MLP head to make them classified. The model of self-attention allows modeling of global context in separated facial features that are spread in space.

4 DEEP FAKE FORENSICS IN PRACTICE: DIGITAL EVIDENCE Authentication.

4.1 Multimedia Evidence Forensic Workflow

To implement the deep fake detection mechanism into the forensic practice, it is necessary to modify the usual evidence handling protocols moving forward to tackle threats of synthetic media. The forensic workflow as it has been traditionally understood, that of identification, collection, preservation, analysis, and presentation, should be modified with considerations unique to deep fake at every phase. As part of the identification, investigators should understand that a digital media that might be relevant to a case can be

synthetic. This knowledge should not just be limited to obvious targets such as videos of high-profile individuals to include any evidence that may be used to benefit tactics and this includes surveillance video, videos of witnesses and any other digital documents.. The metadata and provenance information that could be used in future authenticity evaluation should be addressed during the collection stage. The camera origin, confirmation traces made by an editing software and file creation timestamps can be quite helpful but an investigator should also be aware that advanced actors can also create said metadata. Protection of possibly altered media is to be carried out in accordance with the regular forensic imaging guidelines and taking into consideration the requirement of keeping the material in formats that allow the further evidence in the form of detection. Compression and format conversion may introduce artifacts that are small and destruction of which is in the detection algorithm. The stage of analysis analysis is when the deep fake detection methods are implemented. Most importantly, the forensic examiners must not use a single detection method at the expense of the rest as they all have limitations and blind spots. The collection stage should take care of metadata and provenance information that can be utilized in subsequent authenticity assessment. Presentation The analysis of the results should happen with references to failure modes and confidence scores. Examiners should clearly describe what is and is not capable of various detection methods employed, not over-claim certainty, and be ready to deal with the questions of what is a false positive and how to evade detection.

Figure 8: Forensic Evidence Authentication Process: This is a 6 step flowchart whose flow steps model the forensic process: Identification, Collection, Preservation, Analysis, and Presentation, and deep fake specific sub-steps of each step: metadata analysis, multiple detector ensembles, legal admissibility.

5 CHALLENGES AND LIMITATIONS

The Adversarial Arms Race the adversarial relationship between the generative and detection is the most basic problem of deep fake forensics. As document, ”there is a vicious circle, with countermeasure designers creating new detection approaches and malicious actors reacting, with novel generative algorithms or misleading the detection systems by using both the power of deep fake generation models and adversarial methods at the same time.” This arms race is in a number of dimensions. Generation methods have better perceptual quality which minimizes the artifact which detection algorithms use. Anti-forensic tools are aimed at detection vulnerabilities. And the increased number of generation tools implies that attackers can model their work on the available detectors before putting them into practice, and in effect have their own red-team exercises. The lessons to forensic practice are grim: the tools used to detect a deep fake might not work against deep fakes in the future and the result of an authentication should not be trusted because of this temporal failure.

Generalization and Data set bias another fact that has remained a constant limitation in the study of deep fake detection is that it is difficult to generalize models trained on pre-existing datasets to new forms of manipulation. In line with the Springer review, it is necessary to generalize detection methods to learn circulating methods of forgeries in data, yet it is a problem that should be further investigated. This overgeneralization shows that deep learning models are likely to pick up data-contingent artifacts and not exactly fundamental properties of synthetic media. A detector pretrained with deep fakes produced by a GAN can work effectively on the same GAN output and fail miserably when presented with diffusion-based forgeries or entirely novel ways of manipulation.

6. CONCLUSION

Deep fake technology is a source of a fundamental challenge to the authentication of digital evidence and endangers the fundamental premises that multimedia forensics has been founded on. The issue, as this review has shown, is not solely technical in that it is procedural, legal, and epistemic, making all three factors influencing the aspect of determining the reliability of digital evidence during a court case. The current detection algorithms such as CNN where models are used, transformer models and new paradigms are energetic but there is always a problem with generalization, adversarial robustness and implementation in practice. It is this adversarial dynamic between the generation and detection that there is no permanent technical solution to the problem; instead, the forensic practice has to be molded to the constantly changing threats. The answer to this problem has to be also multi-dimensional. Detectors of technical nature should be woven into larger structures which include provenance infrastructure, controls, training of examiners and legal requirements. The models that are offered through international projects such as DETECTOR project, which combine the efforts of various stakeholders to consider the entire range of issues associated with deep fake forensics, are useful in creating the integrated approach. The deep fake age demands humility and determination, which is why it is needed by the cybersecurity and digital forensics world. Modesty to acknowledge that ultimate detection is impossible and that an atmosphere of trust in authenticity measurements should always be extended through recognition of the constraints. Persistence to keep on enhancing technical functions, perfecting processes and informing stakeholders on the potential and limitations of deep fakes forensics

REFERENCES

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

[2] 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.

[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. 19

[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. Guera, 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. Zollhofer, 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, "Highresolution 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-toimage 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.