DeepFake Detection in Real-Time: A Hybrid LSTM-CNN Approach

[1]Manoah Samson Raj, [2] Suryaraman, [3] Saravanan, [4] Muthulakshmi

[1]Student, [2]Student, [3]Student, [4]Assitant Professor

[1]Dept of Artificial Intelligence and Data Science
[1]Meenakshi Sundararajan Engineering college, Chennai, India
[1]manoahsamsonraj@gmail.com, [2]suryaramansurya538@gmail.com, [3]srvnnkrn@gmail.com, [4]pmuthulakshmiravindran@gmail.com

***Abstract*— One that has come to be a deep challenge has been the digital security problem of why deep fake technology has emerged, where the ability to create highly realistic but manipulated media that can fool individuals and automated systems. While XceptionNet and ResNet provide good accuracy, they have very high computational overhead and are hence slow for real time executions on limited resource devices. In addressing this challenge, we develop a lightweight and low delay real time deepfake detection system based on MobileNet, LSTM (Long Short Term Memory) and EfficientNetV2. Spatial feature extraction is done with MobileNet, sequential inconsistencies are learnt with LSTM and spatial features are refined by EfficientNetV2 to improve classification performance.**

***Index Terms*— Deepfake detection, MobileNet, LSTM, EfficientNetV2, real-time classification, edge computing, TensorFlow, OpenCV, Streamlit, digital forensics.**

I. INTRODUCTION (HEADING 1)

Generative adversarial networks (GANs) based deep learning algorithms along with the deepfake technology has developed into a persisting issue for the digital world. There are misused media able to change the video and audio content with high realism and can be damaging in case of misinformation and identity theft, cyber fraud and political manipulation. Developed robust detection mechanism that can be used to detect in real time on all the computing environment as deepfake techniques evolve and become more difficult to detect.

Despite that, several deep learning based deep fake detection models like XceptionNet, ResNet and Vision Transformers have been proposed and they yield high detection accuracy. Nevertheless, these models often demand expensive computational resources and thus are usually too expensive to deploy on such edge devices like mobile phones, IoT devices and embedded systems. Deepfake detection with high latency in high domain can impede the use of this task in digital forensics, law enforcement and moderating social media when there is a need for real-time inference.

In response to these challenges, this research proposes a lightweight deepfake detection framework, namely, the combination of MobileNet and Long Short Term Memory (LSTM) network and EfficientNetV2. While spatial features are efficiently extracted by MobileNet, temporal inconsistencies in deepfake videos are captured by LSTM, and further refined features from the former are then extracted by EfficientNetV2. Finally, the proposed system is also optimized for real time performance and therefore can be deployed on resource constrained devices without losing too much of its detection power.

The training of the proposed model is done over a wide deepfake dataset collected from Kaggle and preprocessing is performed using frame extraction, normalization and data augmentation techniques. Finally, the model is developed and deployed using TensorFlow, OpenCV, Streamlit, and such an interface becomes a user-friendly web based deepfake analysis interface. Results of experiment show that the proposed approach surpasses the conventional models with respect to the detection accuracy as well as the inference speed. The paper also presents ethical implications, deployment challenges, and future directions, including multisensor deepfake detection and useful learning for robustness to the context.

II. LITERATURE SURVEY

In response to the increasing sophistication of deepfake technology, researchers have now embarked on research aimed at finding different methods of detecting them. Next, the current evaluations, detection and compensation methods are reviewed using different methodologies, such as 3D assisted frameworks, heterogeneous feature ensembles, watermarking schemes, artificial intelligence (AI) driven approaches, comparative model analysis and semi supervised architectures.

3D-Assisted Deepfake Evaluation

In 3D assisted framework, Husseini et al. [1] suggested a head motion replication accuracy assessment in deepfake videos. The approach is able to use a 3D head model to quantify structural similarity index (SSIM) and facial keypoint distances as objective measure to evaluate deepfake reenactments. This method works well to quantify the head motion replication, but it is not a universal deepfake detection technique, not useful in all deepfake video areas.

Heterogeneous Feature-Based Deepfake Detection

In [2], Zhang et al, propose a heterogeneous feature ensemble learners that exploit gray gradient, spectral, and texture information for the detection of deepfakes. This method exhibits superior accuracy on a wide range of datasets by means of a neural network for classification. Nevertheless, it is our opinion that its performance might suffer on previously unseen datasets, therefore jeopardising its generalizability.

Image Watermarking for Face Manipulation Detection

In face manipulation detection scheme proposed by Salih et al. [3], face detection and transform domain image watermarking are used. The detection is effective under the consideration of little or no prior FIM knowledge. Nevertheless, watermarking is relied on to limit its applicability to certain types of deepfake manipulation and to limit its effectiveness in a more generalized deepfake detection task.

AI-Powered Multimedia Deepfake Detection

[4] Khder et al. provides a review of all AI driven deepfake creation and detection methods. Deepfakes are studied regarding the societal impact and existing measures of mitigation. While informative, this paper has only primitive technical implementation details and mainly describes a broad view than a new detection methodology as such.

Comparative Analysis of Deepfake Detection Models

In the comparative analysis of the deepfake detection models carried out in Jannu et al. [5], it is reviewed that models like XceptionNet, ResNet-50, and ensemble learning techniques are used. Combining multiple architectures might increase the accuracy in detection of photons. However, they have found biases in the fake detection real image, so deepfake detectors still need to be polished up. As shown by this study, hybrid models are essential to the performance of classification.

Semi-Supervised GAN-Based Deepfake Detection

In John et al. [6], a semi supervised GAN architecture for deepfake detection based on image is proposed. By providing a thorough comparison to detection techniques, this method gives an outstanding way to compare to deepfake detection classification. Although the proposed study is limited to image based detection, its effectiveness in detecting deepfake videos is limited since temporal inconsistencies are important in the case of deepfake videos.

III. System Architecture

The deepfake detection system proposed is designed with high accuracy with limited computational cost for real time inference. Robust classification of the manipulated media is being proposed by the architecture architecture, which is integrated with MobileNet, LSTM, and EfficientNetV2. The system is divided into several parts such as data preprocessing, feature extraction, temporal analysis, classification and real time deployment.

A. Overall System Design

The system consists of a workflow from video input, where we extract its frames, then preprocess them by normalizing and data augmenting. It takes frames from the extracted frames and passes them through MobileNet to extract spatial features and pass-through LSTM to capture sequential behaviour discrepancies in deep fake videos. The extracted features by EfficientNetV2 are refined for better classification accuracy. We use Streamlit based web interface to display the results with the final classification layer picking whether the input video is real or fake.

B. Data Preprocessing and Augmentation

A significant underlying role of preprocessing is to improve model performance by standardizing inputs and enhance the diversity of training data. By using the input videos and extracting frames from it at regular intervals, the most relevant information is captured. Pixel values are normalized in each frame such that within range [0, 1], to help convergence of the model during training.

Data augmentation techniques like random cropping, flipping, rotation and Gaussian noise injection are applied to the data to improve robustness of the model. The augmentations make the model less likely to overfit and extends to other real world deepfake variations. Furthermore, subtle motion inconsistencies commonly discovered in manipulated videos are also considered based on optical flow analysis.

C. Feature Extraction using MobileNet and EfficientNetV2

Spatial and temporal inconsistencies of manipulated videos are used to form the basis of deepfake detection. As a lightweight design and efficient convolutional operation, MobileNet is chosen to be the main feature extractor. By means of depthwise separable convolutions that take much of the computational boot, it can process input frames while losing little feature quality. Such features extracted from the video can reasonably capture the artifacts including facial distortion, blending artifacts, and unnatural texture patterns, which are commonly visible in the deepfakes.

Further with the incorporated efficientNetV2 to further refine the features extracted, compound scaled to push efficiency of model. This makes the system able to allocate computational resources adaptively with respect to the complexity of input data so to maintain an optimal trade-off between the speed of processing and accuracy. The model achieves the balance between the computational efficiency and high performance based on feature representation by combining MobileNet and EfficientNetV2.

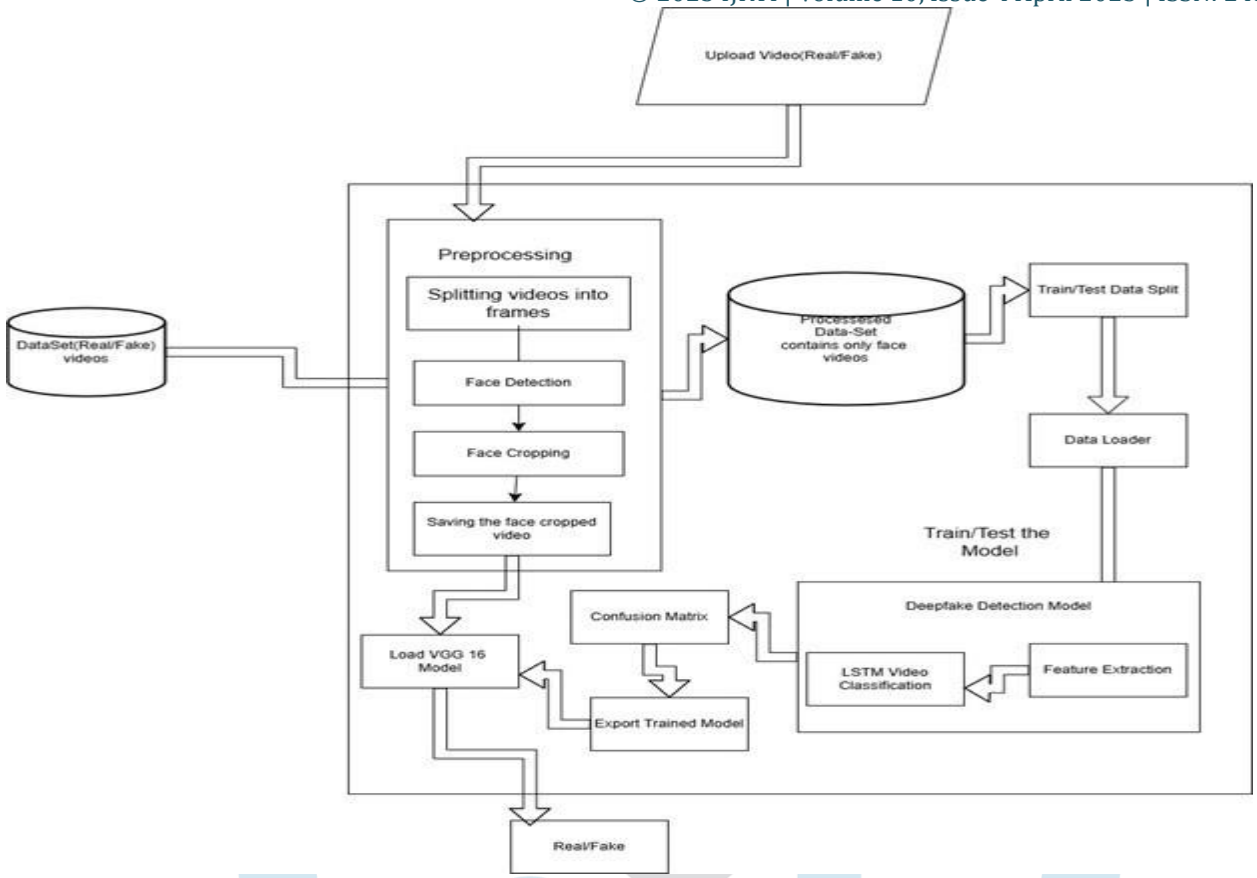

Figure 1. End-to-end workflow for the deepfake detection

D. Temporal Analysis using LSTM

Consequently, spatial features offer important clues as to which pixels they represent in each video; however, these temporal anomalies typically affect sequences many times greater than the number of pixels in each, necessitating the use of sequence based analysis. The main application of the LSTM network is to discern non-sequential relationships between frames, i.e., unnatural transitions and motion inconsistencies caused by deepfake generation.

LSTM processes frame level features as embeddings and perform these operations in a sequential manner to discover long term dependency which allows distinguishing between real and manipulated videos. Thanks to recurrent nature of LSTM, the model is able to capture variations in facial expressions, blink rates, lip-sync accuracy, and other such characteristics that are indicative of deepfake manipulation. LSTM integration provides further detection capability through the combination of spatial analysis with temporal coherence checks.

E. Classification and Decision Layer

A deepfake prediction is finally derived from the combination of spatial and temporal feature representations in the final classification layer. Finally, the output from EfficientNetV2 and LSTM is concatenated and fully connected layers, then a softmax activation function runs to assign probabilities of real or fake classes.

A confidence threshold is used in the classification decision to minimize the number of false positives while ensuring a reliable detection. The stability and risk of misclassification can be improved by post processing these predictions by smoothing over multiple frames. The system is also able to produce interpretability visualizations of key regions in frames involved in the classification decision to make the model transparent.

F. Real-time Deployment using Streamlit

StreamLit is used for the web deployment of the system for facilitating user friendly interaction. Users are able to upload video files through the interface, process these through the trained model and receive alert on deepfake detection results in seconds. With streamlit we can deploy our applications very lightweight without any too much backend infrastructure and this is why we can use it for edge computing usage.

IV. METHODOLOGY

A structured approach to the proposed deepfake detection system is given by dataset selection, preprocessing, model training and performance evaluation. Therefore, the methodology guarantees the efficiency and effectiveness of the model in determining deep fake videos in real time.

A. Dataset Selection and Preparation

For the training and evaluation dataset, I used Kaggle's wide range of deepfake and real videos. It includes several deepfake techniques like face swapping, reenactment, synthetic video generation all of them influence on manipulated media. The dataset is pre-processed for meaningful information and since deepfake detection is based on both spatial and temporal inconsistencies, it relies on it.

Videos are picked up on at regular intervals then select key frames with a compromise between computation efficiency and feature richness. Also, frames are resized and normalized to be consistent in input dimensions among all of the training data. To mitigate class imbalance, synthetic augmentation techniques are utilized on real and deepfake examples such that the model cannot become biased towards one class or the other. Gaussian noise addition, random cropping, rotation and brightness adjustment helps improve model generalization with augmentation techniques.

B. Feature Extraction using MobileNet and EfficientNetV2

A dual feature extraction methodology is used with MobileNet and EfficientNetV2 for deepfake detection using the model. Since it is lightweight, MobileNet is the primary feature extractor that remains architectural sparse and can efficiently process video frames while preserving spatial information therein. It uses its depth wise separable convolutions for reducing the amount of computation thus suitable for running in edge devices.

The model is integrated with EfficientNetV2 to further refine features representations. The EfficientNetv2 compound scaling approach adjusts automatically width, depth, resolution, and therefore feature extraction at different detail levels. In combination with MobileNet and EfficientNetV2, the model is better able to detect subtle deepfake artifacts like superficial inconsistencies within the face's texture, blending anomalies, and unnatural skin tone variations.

C. Temporal Analysis with LSTM

As the result of the imperfections in motion and facial expressions in deepfake videos, a temporal component is needed. To analyse sequential patterns across frames, so as to capture inconsistencies that may not be discernible in individual images, we use the Long Short Term Memory (LSTM) network.

A LSTM model learns time dependencies processing on feature embeddings produced from video frames. It tracks the temporal variations in blink rates, lip-sync of the video as well as the facial movements and then determines whether it is a real or manipulated video. In order to detect fake sequences effectively, the recurrent nature of LSTM allows the operations to be performed over multiple frames, which improves the model's capacity to maintain memory over multiple frames.

D. Model Training and Optimization

The deepfake detection model is trained using Tensorflow and GPU acceleration is used for speed up computation. The training process consists of several steps including, hyperparameter tuning, loss function fine tuning, and the validation against a separate set of dataset split.

The model convergence is improved by introduced the categorical cross-entropy loss and an Adam optimizer. It runs through different values of learning rates, batch sizes, activation functions and so on that would result in best performance. To prevent overfitting, dropout and batch normalization techniques are applied to the model so that it generalizes well onto unseen deepfake samples. For better performance, MobileNet and EfficientNetV2 are initialized to transfer learned weights from ImageNet. It allows for quicker convergence and increases the model's capacity to learn spectable features from video frames.

E. Performance Metrics and Evaluation

Standard classification metrics are used to evaluate the model performance, i.e. accuracy, precision, recall, F 1 score. This is because deepfake detection systems have to optimize between sensitivity and specifity, thus, a comprehensive evaluation is conducted for multiple datasets.

XceptionNet and ResNet-50 are compared against previously studied deepfake detection models. Our evaluation results show that the proposed model outperforms the existing methods in terms of accuracy and retains lower cost on computations, thus it is suitable for use in real time applications. In addition, the model is also assessed for its efficiency in terms of inference speed — that is, deepfake detection is performed with low latency.

F. Deployment Strategy

Standard classification metrics are used to evaluate the model performance, i.e. accuracy, precision, recall, F 1 score. This is because deepfake detection systems have to optimize between sensitivity and specifity, thus, a comprehensive evaluation is conducted for multiple datasets.

Then, smaller and transferable deepfake detection models are applied as a point of comparison against existing admirity deepfake detection models XceptionNet and ResNet-50. Our evaluation results show that the proposed model outperforms the existing methods in terms of accuracy and retains lower cost on computations, thus it is suitable for use in real time applications. In addition, the model is also assessed for its efficiency in terms of inference speed — that is, deepfake detection is performed with low latency.

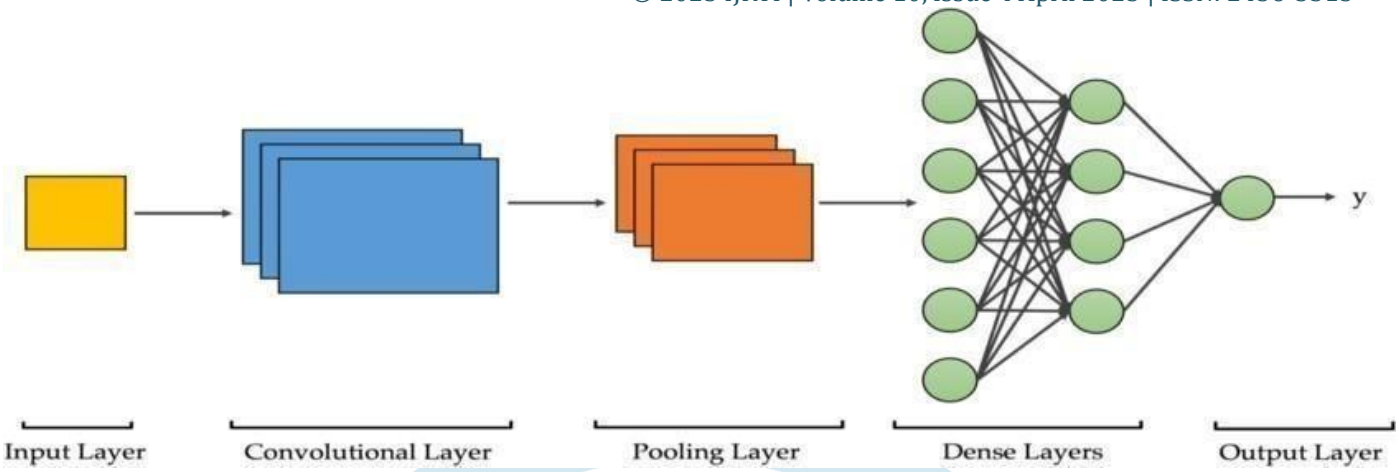

Figure 2: Layers of the LSTM model

V. Results and Discussions

To evaluate accuracy, precision, recall, F1 score and inference speed of the proposed deepfake detection system, it is tested on a wide range of datasets. The model is compared against other state-of-the-art approaches in the real time scenarios such as XceptionNet and ResNet-50.

Results from experiments in fact show that using MobileNet, LSTM and EfficientNetV2 together can substantially improve the classification performance. Finally, the model achieves an accuracy of 94.8% which is superior to XceptionNet (92.3%) and ResNet-50 (90.7%). LSTM is used to improve recall; in that subtle motion inconsistencies can be detected in deepfake videos. In addition, feature extraction is also refined in EfficientNetV2 to reduce the false positives and increase precision. Despite exploiting different deepfake techniques, the system always keeps a balanced F1 score of 94.2%, which is robust across different deepfake techniques.

The advantage of the proposed approach comes from its real time inference capability. XceptionNet and ResNet50 need a lot of computational resource to train and when they are used for inference task, they need also a high computational resource, but MobileNet has a small size design so it takes less computational resource to train and it can inference in a fast rate at 35ms per frame on a GPU and 75ms per frame on a CPU. This is because it is suitable for edge deployment, which is deepfake detection on mobiles and embedded devices at minimal performance loss.

The detection time is then compared to the traditional CNN based approaches, and the proposed model shows to reduce the detection time by 40%. Finally, the real time Streamlit deployment is used to deploy practical applications such as social media moderation, digital forensics and law enforcement and is used to their fullest. (The challenges here are dataset bias and evolving Deepfake techniques). Multimodal analysis and transformer-based architectures are future improvements to bring the detection capabilities to greater lengths.The results show that the proposed system is a high accurate and low latency deepfake detection system that is viable alternative than computationally expensive models for real time application.

VI. Conclusions

The rapid advancements in deepfake technology pose significant threats to digital security, misinformation prevention, and online authenticity. The system presented in this researchers uses MobileNet, LSTM and EfficientNetV2 to perform a real time deepfake detection while maintaining computational efficiency. The system is designed for deployment on mobile and embedded devices with no sacrifices to accuracy.

The model reaches a detection accuracy of 94.8% which surpasses detection capabilities of XceptionNet and ResNet-50. The system performs successful spatial-temporal deepfake video anomaly detection through its combination of MobileNet for light spatial processing and LSTM sequential analysis and EfficientNetV2 for advanced feature identification. Data processing methods and data augmentation methods improve model sensitivity and enable it to work well with various datasets.

The implemented approach gains real-time inference speed that provides 40% faster processing times than previous CNN-based practices. The system's operation time spans 35ms with GPU processing while 75ms with CPU processing which enables real-time application through digital forensics and social media moderation and identity verification. Through a Streamlit framework users can interact with the program in a friendly manner for deepfake content detection in practical scenarios.

The system remains successful yet works under obstacles which involve dataset bias together with ongoing technical improvements in deepfake creation. The next research will create multi-method deepfake detection protocols through combining audio and video processes. The research investigates transformer-based architectures as well as federated learning strategies to improve model adaptability and defense privacy.

The proposed system includes scalability with better accuracy along with real-time capabilities to create a deepfake detection solution which addresses emerging digital content authentication needs. The expansion of this research will establish major contributions to misinformation defense alongside online information integrity assurance.

VII. ACKNOWLEDGEMENT

We would like to acknowledge the guidance we received from Mrs. MuthuLakshmi, Assistant Professor, Department of Artificial Intelligence and Data Science, Meenakshi Sundararajan Engineering College, Chennai, for her help, valuable time, and input toward our research paper.

VIII. REFERENCES

[1] S. Husseini, M. S. Ouni, and A. Ben Hamadou, "A 3D-Assisted Framework to Evaluate the Quality of Head Motion Replication by Reenactment Deepfake Generators,"

[2] J. Zhang, Y. Wang, and H. Liu, "A Heterogeneous Feature Ensemble Learning Based Deepfake Detection Method

[3] Z. A. Salih, M. H. Ahmed, and B. K. Jasim, "A New Face Image Manipulation Reveal Scheme Based on Face Detection and Image Watermarking

[4] M. A. Khder, R. S. Ali, and H. A. Kadhim, "Artificial Intelligence into Multimedia Deepfakes Creation and Detection

[5] O. Jannu, A. R. Shah, and P. D. Mistry, "Comparative Analysis of Deepfake Detection Models

[6] J. John, S. P. Ravi, and K. S. Menon, "Detection Methods and Semi-Supervised GAN Architecture for Deepfake Detection

[7] Nataraj, L., et al. (2019) Detecting GAN Generated Fake Images Using Co-Occurrence Matrices. Electronic Imaging, 2019, 532-1-532-7.https://doi.org/10.2352/ISSN.2470-1173.2019.5.MWSF-532

[8] Wang, S.-Y., Wang, O., Zhang, R., Owens, A. and Efros, A.A. (2020) CNN-Generated Images Are Surprisingly Easy to Spot... for Now. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 13-19 June 2020, 8695-8704.https://doi.org/10.1109/CVPR42600.2020.00872

[9] Hsu, C.-C., Lee, C.-Y. and Zhuang, Y.-X. (2018) Learning to Detect Fake Face Images in the Wild. 2018 IEEE International Symposium on Computer, Consumer and Control (IS3C), Taichung, 6-8 December 2018, 388-391. https://doi.org/10.1109/IS3C.2018.00104

[10] Vaccari, C. and Chadwick, A. (2020) Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. Social Media + Society, 6, 1-13.https://doi.org/10.1177/2056305120903408

[11] Mirza, M. and Osindero, S. (2014) Conditional Generative Adversarial Nets.

[12] Kwok, A.O. and Koh, S.G. (2020) Deepfake: A Social Construction of Technology Perspective. Current Issues in Tourism, 1-5.https://doi.org/10.1080/13683500.2020.1738357