# Deep Fake Detection in Videos with explainability

Dharunkumar Udayakumar 205001031
Dharshan S 205001030

BE CSE, Semester 8

Dr. Sakaya Milton R
Supervisor

**Project Review: 1** (23 March 2024)
Department of Computer Science and Engineering
SSN College of Engineering

---

## 1 Abstract

The widespread use of social media and advancements in deep learning have led to the proliferation of deepfake videos, presenting significant challenges in combatting misinformation and preserving trust. In response, this research proposes a novel approach utilizing LSTM-based neural networks and pre-trained CNN architectures to detect deepfake videos. By leveraging temporal analysis and explainable AI techniques such as heat maps, our model aims to provide transparent and reliable detection capabilities.

Beyond technical advancements, the societal implications of deepfake technology are profound, affecting public trust, social cohesion, and democratic processes. Misinformation propagated through deepfakes can sow discord and undermine the credibility of institutions and individuals. Moreover, the accessibility of deepfake creation tools exacerbates these risks, enabling malicious actors to exploit vulnerabilities in media ecosystems. Through our research, we seek to address these challenges while promoting media literacy and responsible digital citizenship to mitigate the societal impacts of deepfake technology.

# 2 Introduction

## 2.1 Motivation

Advancements in mobile camera tech and widespread social media use have made creating and sharing videos easier. Deep learning has led to previously unimaginable technologies, like generative models producing lifelike content. These models are utilized in various fields such as text-to-speech for accessibility and generating medical imaging data.

Spreading of the Deep fakes over the social media platforms have become very common leading to spamming and peculating wrong information over the platform. Just imagine a deep fake of our prime minister declaring war against neighboring countries.These types of the deep fakes will be terrible, and lead to threatening, misleading of common people.

Deep learning models used to create such videos are called Generative Adversarial Network (GAN). The rise of the technology has led to a situation where anyone can create these deep fake videos without any prior knowledge on coding of Machine Learning by using Mobile apps. These mobile apps allow people to create deep fakes without strict authentication or age restriction.

## 2.2 Follow up on Review 1 suggestions

The suggestions given in review were:

- Make the explainability part more clear and precise on what will be the output and what information can be gained from the output
- Working of heat maps. heat maps highlight the regions of an input that are most influential in a model's decision-making process. They are generated by computing the gradient of the model's output with respect to the input pixels, identifying areas of high importance.
- Instead of removing video which are corrupted or of very large length, try to integrate equal number of frames from each video

## 2.3 Background

- Deepfake detection is a critical task aimed at distinguishing between authentic videos and manipulated content generated using deep learning techniques. With the rise of generative models like Generative Adversarial Networks (GANs),

which can produce highly realistic videos, detecting deepfakes has become increasingly challenging. However, researchers utilize various methods, including analyzing inconsistencies in facial expressions, blinking patterns, or artifacts unique to deepfake generation processes, to develop effective detection algorithms.

- ResNext is a state-of-the-art convolutional neural network (CNN) architecture known for its remarkable performance in image classification tasks. By employing a modular block structure with grouped convolutions, ResNext efficiently captures intricate features from input images, making it particularly suitable for tasks requiring sophisticated feature extraction, such as deepfake detection. Its ability to handle diverse input data and learn hierarchical representations contributes to its effectiveness in discriminating between authentic and manipulated videos.

- Long Short-Term Memory (LSTM) networks are a subtype of recurrent neural networks (RNNs) capable of learning patterns in sequential data with long-range dependencies. In the context of deepfake detection, LSTMs excel at analyzing temporal information in videos, allowing models to identify subtle inconsistencies or anomalies indicative of manipulation. Additionally, heat maps play a crucial role in model interpretability by highlighting the regions of an image or video that contribute most significantly to the model's decision-making process. By computing gradients of model outputs with respect to input pixels, heat maps provide valuable insights into the inner workings of the model, aiding researchers in understanding and improving deepfake detection algorithms.

## 2.4 Problem statement

The problem we are trying to tackle is to detect deepfake videos and provide reasoning for why a particular video is deemed real or fake. Explainable AI is a new domain that is utilized to understand the reasoning behind a model's predicted output. Our method employs an LSTM-based artificial neural network to process the sequential temporal analysis of the video frames, and a pretrained ResNeXt CNN to extract frame-level features. The ResNeXt Convolutional Neural Network extracts these features, which are then used to train the Long Short-Term Memory-based Recurrent Neural Network to determine whether the video is real or fake. Utilizing explainable AI techniques like heat maps, we generate a heat map that indicates which regions in the video are focused on by the model to make its decision. Additionally, we develop

a user-friendly application that allows users to upload a video and receive real-time predictions along with heat maps, enabling users to identify important parts of the video that are determined to be fake or real in real-time.

## 2.5 Purpose

Our project aims to address the rising concern of deepfake videos by developing an AI-based solution capable of both identifying such videos and providing insights into their authenticity. By combining LSTM-based neural networks for temporal analysis with a pre-trained ResNext CNN for feature extraction, our model can effectively discern between real and fake videos. Moreover, by leveraging explainable AI techniques like heat maps, we aim to generate heatmaps highlighting the specific regions within a video that influence the model's decision-making process. Ultimately, our goal is to empower users with a tool that not only detects deepfakes but also provides transparent explanations for its classifications, thereby enhancing trust and accountability in media content.

## 2.6 Scope of the Report

Our project focuses on developing a robust deepfake detection system with explainable AI capabilities, culminating in the creation of a user-friendly application. The primary goal is to accurately differentiate between authentic and manipulated videos, leveraging advanced AI algorithms for precise classification.

1. Deepfake Detection System: Our system employs state-of-the-art AI techniques, including LSTM-based neural networks and pretrained CNNs, to analyze temporal and spatial features of video frames and identify potential signs of manipulation.

2. Explainable AI Techniques: To enhance transparency and user trust, we integrate explainable AI methods such as heat maps, enabling users to visualize the key regions in videos that influence the model's classification decisions.

3. User-Friendly Application: We develop an intuitive web or mobile application interface that allows users to easily upload videos for analysis.

4. Real-Time Feedback: Upon uploading a video, users receive prompt feedback on its authenticity, accompanied by explanations derived from the heat maps.

## 2.7 Structure of the Project

Initially we spoke about the motivation behind the project and the impact that it will have in the society. Important needs for such detector is also mentioned. We covered

various research works that have been conducted on deepfake detection, detailing the efforts made in each paper. Section 4 provides a brief discussion on the research gaps and limitations. In Section 5, we elaborate on our solution to the problem. Section 6 outlines our expected outcomes. Section 6 addresses the feasibility of the project. Lastly, Section 7 concludes with implementation details. In section 7, we mentioned about out current results in the project mentioning about the model accuracy and its reliability.

# 3 Related work

The paper[1] proposes a temporal-aware system for detecting deepfake videos. The dataset includes 300 deepfake videos and 300 videos from the HOHA dataset. The system utilizes a convolutional LSTM structure, combining CNNs for feature extraction and LSTMs for sequence processing. With sequences as short as 40 frames, the system achieves over 97% accuracy in predicting deepfake fragments. Limitations include reliance on pre-trained models and potential challenges with evolving deepfake techniques.

The paper[2] introduces a Temporal Dropout 3-Dimensional Convolutional Neural Network (TD-3DCNN) for detecting deepfake videos. The data collection process involves three datasets: Celeb-DF(v2), DFDC, and FaceForensics++. The models used include TD-3DCNN, along with comparisons to state-of-the-art detectors like Two-stream NN, MesoNet, Head Pose, Visual Artifacts, Multi-task, and Warping Artifacts. Accuracy metrics such as ACC, AUC, and Logloss are employed for evaluation. TD-3DCNN achieves high accuracy across all datasets, surpassing previous methods. However, limitations may include scalability to emerging deepfake techniques and computational complexity.

The paper[3] introduces a novel method for detecting DeepFake videos by leveraging artifacts from the face warping process. They collect data from the internet and dynamically generate negative examples during training. They use CNN models including VGG16 and ResNet variants. The best model, ResNet50, achieves an AUC of 97.4% on the UADFV dataset and 99.9% on LQ videos in the DeepfakeTIMIT dataset. Limitations include vulnerability to various compression techniques and reliance on pre-designed network structures

The paper[4] introduces a novel approach for identifying deep fakes by analyzing eye blinking patterns, a critical factor in distinguishing between authentic and manipulated videos. The study employs the Long-term Recurrent Convolution Network (LRCN) to conduct temporal analysis on cropped frames focusing on eye blinking behavior. However, with the advancement of deepfake generation algorithms, the absence of eye blinking alone may no longer suffice for accurate detection. Therefore, additional parameters such as teeth enhancement, facial wrinkles, and eyebrow placement should be considered to enhance the reliability of deep fake detection methods.

In the paper[5] utilization of capsule networks for detecting counterfeit images and videos involves a technique aimed at identifying forged or manipulated content across various scenarios, including detecting replay attacks and computer-generated videos. The approach involves training the capsule network using random noise, which although resulted in favorable performance on their dataset, may lead to potential failure when applied to real-time data due to noise interference during training. Our proposed method suggests training the model on noise-free and real-time datasets to enhance its robustness and applicability in practical settings

The paper[6] initially dicusses about how deep fakes are created using a method called Generative Adversarial Network (GAN). It implemented image augmentation to improve the size of the dataset so that overfitting of the model can be avoided. It uses Xception Model and the speciality of the model is that it gets trained fast. Their proposed model architecture maintains a trade-off between the complexity of the model, including the training time of the model, prediction time for an input video, computational complexity, resources required for training our model, etc., and performance of the model on the DFDC dataset for deepfake video detection.

# 4 Research Gaps

Limitations of existing work are:

- Techniques which are very specific like eye movement detections are failing due to creation of advanced models
- Scalability and computational Complexity

- Reasoning behind model prediction and hypertuning of parameters for getting best result from the model

.

The above mentioned challenges are some important areas where work must be done in the case of deepfake detection.

# 5 Proposed Methodology

The proposed methodology is described below.

**Architecture:**

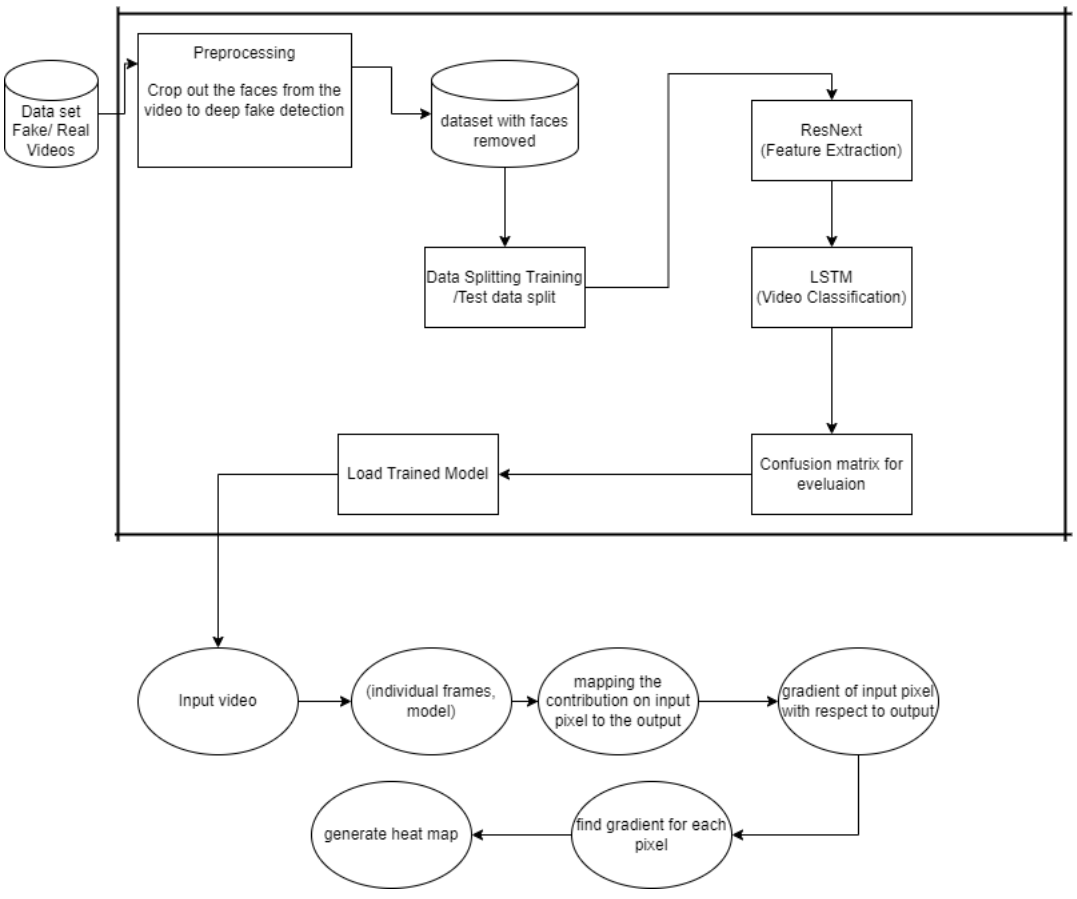


Figure 1: Schematic Representation of Deepfake Detection Model

## 5.1 Data Preprocessing

In the initial phase of our research, we collected a comprehensive dataset from the deepfake detection challenge dataset, comprising 3000 videos. However, upon closer inspection, we identified the presence of audio-altered files within the dataset. Given that our primary focus lies on visual deepfake detection, we made the decision to exclude all audio-altered videos to ensure clarity and maintain the integrity of our research objectives.

To streamline subsequent data processing and analysis, we recognized the importance of standardizing the number of frames per video. Given the computational constraints inherent in our research environment, we carefully determined a threshold based on the average number of frames across all videos. By doing so, we aimed to establish consistency in the dataset, facilitating smoother processing and analysis procedures.

Furthermore, prior to utilizing the videos for training purposes, we implemented a sophisticated face detection algorithm. This algorithm enabled us to precisely identify and isolate facial regions within each video frame. By focusing exclusively on facial features, we aimed to enhance the efficacy of subsequent model training and evaluation processes.

## 5.2 Data-set Split

Having preprocessed the dataset, we proceeded to partition it into distinct training and testing subsets. Employing a ratio of 70

To maintain a representative and unbiased dataset, we implemented a stratified split strategy. This approach ensured an equal distribution of both authentic and manipulated videos within each subset. By maintaining a balanced representation of real and fake videos, we aimed to mitigate potential biases and enhance the robustness of our model evaluations.

## 5.3 Training and Evaluation

With the dataset partitioned, we embarked on the training phase of our research. Leveraging the preprocessed dataset, we employed a state-of-the-art ResNext CNN model for feature extraction at the frame level. Renowned for its efficacy in handling complex data representations, the ResNext architecture offered optimal performance in feature extraction tasks.

Subsequently, we trained a Long Short-Term Memory (LSTM) network to classify

the videos based on the extracted features. The LSTM architecture, renowned for its proficiency in sequential data analysis, was particularly well-suited for tasks involving temporal analysis of video frames. By harnessing the complementary strengths of ResNext for feature extraction and LSTM for sequence processing, we aimed to develop a robust deepfake detection model capable of accurately discerning between authentic and manipulated videos.

### 5.4 Explainable AI

To enhance the transparency and interpretability of our model's predictions, we incorporated explainable AI techniques into our research framework. In particular, we employed the concept of heat maps to elucidate the decision-making process of our model. By mapping the contribution of each input pixel to the model's output prediction, heat maps provided valuable insights into the regions of the video that exerted the greatest influence on the classification process. This interpretability not only enhanced user understanding of the model's decisions but also facilitated the identification of potential areas of concern within the video content.

## 6 Expected Outcomes

- An efficient model for the prediction of deep fake videos using ResNext and LSTM
- heat maps for the video which comes under explainable AI to identify the reason behind the models decision by the way of heat maps which highlights the region in the video which was important for the model's prediction.
- An user friendly application where user can upload video and do the detection with real time generation of heat maps to know the reason for the models prediction.

# 7 Results

## 7.1 Working and performance of the trained model

- We used resNext model along with LSTM for feature extraction. The dataset contained over 1000 videos with a mix of real and fake videos. But there was a issue with the dataset, among all the videos only 110 were videos which are classified as REAL. Initially when we trained the model with the entire dataset, the accuracy was high as most of the validation videos had only FAKE videos.
- Accuracy and correctness of the model improved when we again trained the model but this time with equal no of real and fake videos. The accuracy finally came to be around 85
- The model architecture utilizes a pre-trained ResNeXt-50 32x4d model as the backbone convolutional neural network (CNN).
- Removed the last two layers of the ResNeXt model and adds an LSTM layer for sequence modeling, followed by dropout regularization to the LSTM output.
- Additionally, the model includes a fully connected linear layer for classification and applies adaptive average pooling to the input feature maps.
- **The approach to be followed to include saliency maps is:**
- **Modify the Forward Pass:** Save the feature maps (x) after passing through the ResNeXt backbone but before the average pooling layer. Pass the saved feature maps (x) to the LSTM layer.
- **Add a Method for Visualizing Feature Maps:** Implement a method that takes the saved feature maps (x) as input and visualizes them using suitable techniques such as heatmap visualization.This method could be invoked separately from the forward pass whenever feature map visualization is required.

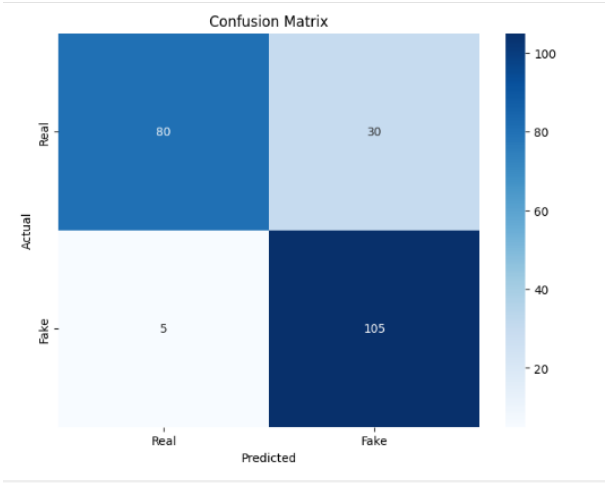


Figure 2: Confusion matrix

```python
class MyModel(nn.Module):
    def __init__(self, num_classes, latent_dim=2048, lstm_layers=1, hidden_dim=2048, bidirectional=False):
        super(MyModel, self).__init__()
        # Load pre-trained ResNeXt-50 32x4d model
        feature_extractor = models.resnext50_32x4d(pretrained=True)
        # Remove last two layers of the ResNeXt model
        self.feature_extractor = nn.Sequential(*list(feature_extractor.children())[:-2])
        # LSTM layer for sequence modeling
        self.sequence_model = nn.LSTM(latent_dim, hidden_dim, lstm_layers, bidirectional)
        # Dropout regularization
        self.dropout = nn.Dropout(0.4)
        # Linear layer for classification
        self.classifier = nn.Linear(2048, num_classes)
        # Adaptive average pooling to handle varying input sizes
        self.avg_pooling = nn.AdaptiveAvgPool2d(1)

    def forward(self, x):
        # Reshape input data
        batch_size, seq_length, c, h, w = x.shape
        x = x.view(batch_size * seq_length, c, h, w)
        # Feature extraction using ResNeXt
        x = self.feature_extractor(x)
        # Apply adaptive average pooling
        x = self.avg_pooling(x)
        # Reshape for LSTM input
        x = x.view(batch_size, seq_length, 2048)
        # LSTM forward pass
        x_lstm, _ = self.sequence_model(x, None)
        # Average pooling and linear classification
        return self.dropout(self.classifier(torch.mean(x_lstm, dim=1)))
```

Figure 3: Pseudo code

# 8 Issues Pertaining to Society, Health, Safety, Legal, Environment, and Culture

## 8.1 Societal Impact

The proliferation of deepfake videos poses significant challenges to societal cohesion and trust. Deepfakes have the potential to cause social unrest by influencing public opinion and exacerbating existing divisions within communities. The spread of misinformation through deepfakes can lead to polarization, undermining the foundation of trust in media and public figures. This erosion of trust can have far-reaching consequences for democratic processes, social stability, and public discourse.

## 8.2 Health Concerns

Exposure to manipulated content, particularly deepfake videos containing fabricated statements from public figures or maliciously altered visuals, can have detrimental effects on mental health. Individuals exposed to such content may experience heightened levels of anxiety, distress, and confusion. Additionally, the malicious use of deepfakes for purposes such as non-consensual pornography can lead to severe psychological trauma for victims, posing significant challenges for their well-being and recovery.

## 8.3 Safety Implications

The creation and dissemination of deepfake videos can have profound safety implications, particularly in legal contexts and public security. Deepfakes can be used to fabricate false evidence, leading to wrongful accusations and legal repercussions for innocent individuals. Moreover, manipulated videos depicting violent or criminal acts can incite real-world harm, jeopardizing public safety and security. Addressing these safety concerns requires robust detection mechanisms and effective regulatory frameworks to prevent the misuse of deepfake technology.

## 8.4 Legal Challenges

The rapid advancement of deepfake technology has outpaced existing legal frameworks, presenting formidable challenges for law enforcement agencies and policymakers. Current laws may lack clarity or specificity regarding the creation, distribution, and manipulation of deepfake content, making it difficult to prosecute offenders

or regulate the use of such technology. Legal issues surrounding privacy infringement, defamation, and intellectual property rights further complicate the landscape, necessitating comprehensive legislative measures to address emerging threats posed by deepfakes.

### 8.5 Environmental Impact

While not directly related to environmental concerns, the computational resources required for training deepfake detection models and processing large datasets can have indirect environmental implications. The energy-intensive nature of deep learning algorithms and the massive computational infrastructure needed for model training contribute to carbon emissions and resource consumption. As the scale of deepfake detection efforts continues to grow, it becomes imperative to consider the environmental footprint of these activities and explore strategies for minimizing their ecological impact.

### 8.6 Cultural Ramifications

Deepfake videos have the potential to distort cultural narratives, perpetuate harmful stereotypes, and undermine historical authenticity. Manipulated visuals depicting public figures or cultural icons can alter societal perceptions and erode the integrity of cultural heritage. Additionally, the proliferation of deepfake technology poses challenges to the preservation of cultural authenticity and identity. Addressing these cultural ramifications requires awareness, education, and collaborative efforts to promote media literacy and safeguard cultural integrity in the digital age.

## 9 Timeline

The timeline of the project is outlined below:

| | |
|---|---|
| January | Dataset collection, Literature Survey |
| February | Preprocessing and training |
| March | Implementation of explainability using heat maps |
| April | User interface and app creation |

Table 1: Timeline of the project

# References


[1] D. Güera and E. J. Delp *Deepfake Video Detection Using Recurrent Neural Networks*, 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018.

[2] Daichi Zhang1,2, Chenyu Li1,2, Fanzhao Lin1,2, Dan Zeng3and Shiming Ge *Detecting Deepfake Videos with Temporal Dropout 3DCNN*, Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)

[3] Yuezun Li, Siwei Lyu *Exposing DeepFake Videos By Detecting Face Warping Artifacts* University at Albany, State University of New York, USA

[4] Yuezun Li, Ming-Ching Chang and Siwei Lyu *"Exposing AI Created Fake Videos by Detecting Eye Blinking*, arXiv:1806.02877v2.

[5] ansal, Nency , Aljrees, Turki , Yadav, Dhirendra , Verma, Gyanendra *Real-Time Advanced Computational Intelligence for Deep Fake Video Detection*Applied Sciences. 13. 3095. 10.3390/app13053095

[6] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies,Matthias Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images" in arXiv:1901.08971.

[7] Yuezun Li , Xin Yang , Pu Sun , Honggang Qi and Siwei Lyu "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics" in arXiv:1909.12962

[8] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. arXiv:1702.01983, Feb. 2017

[9] J. Thies et al. Face2Face: Real-time face capture and reenactment of rgb videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2387–2395, June 2016. Las Vegas, NV.

[10] Yuezun Li, Siwei Lyu, "ExposingDF Videos By Detecting Face Warping Artifacts," in arXiv:1811.00656v3.

[11] Yuezun Li, Ming-Ching Chang and Siwei Lyu "Exposing AI Created Fake Videos by Detecting Eye Blinking" in arXiv:1806.02877v2.

[12] Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen " Using capsule networks to detect forged images and videos " in arXiv:1810.11

[13] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6.

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, June 2008. Anchorage, AK