Deepfake detection using Eulerian video magnification and Deep Learning

Gideão P. Abreu[1], Jefferson O. Andrade[1], Karin S. Komati[1]

[1]*Graduate program in Applied Computing (PPComp)*
Instituto Federal do Espírito Santo (IFES), Campus Serra
Av. dos Sabiás, 330 - Morada de Laranjeiras, 29166-630, Serra-ES, Brazil
gideaoabreu@gmail.com, {jefferson.andrade, kkomati}@ifes.edu.br

Abstract. The Internet witnesses millions of video views every minute in the contemporary landscape of extensive data proliferation and ubiquitous social media use. In this context, the burgeoning advancements in deepfake technologies introduce security and privacy concerns. These technologies facilitate manipulating video and audio content to an extent where someone can seamlessly replace one person's visage with another's, or entirely synthetic videos can be crafted using real individuals' voices and appearances, potentially deceiving viewers. Consequently, the malicious exploitation of deepfakes has caused apprehension due to its potential adverse societal repercussions, including, but not limited to, harm infliction, extortion, and reputational jeopardy. This study seeks, within the realm of deepfake detection, to evaluate the efficacy of Eulerian video magnification (EVM), a technique that accentuates subtle cues and motions, typically imperceptible to the naked eye. To this end, we propose the use of hybrid architectures comprising Convolutional Neural Networks (CNN) and Vision Transformers (ViT) with magnification techniques. Subsets of the FakeAVCeleb dataset, including authentic and manipulated videos, will train, validate, and test the model. The evaluation of the model will employ metrics such as accuracy, precision, recall, and the F1 score, with results compared to the existing literature.

Keywords: Convolutional Neural Networks, Vision Transformer, FakeAVCeleb.

1 Introduction

Techniques for forging fake videos, known as deepfakes, have been gaining notoriety on the Internet, allowing, for example, replacing a person's face in a video with another face [1]. The manipulations allow that head movements, facial expressions, lighting, and lip synchronization while speaking are as in the original video, but with a different face [2]. When used maliciously, deepfakes have the potential to harm, intimidate, extort, cause psychological damage, and compromise reputation [3]. Being a technology with great potential for influence, it can lead to personal losses, public panic, and even threats to peace between countries [4].

Thus, there is a demand to investigate methods capable of effectively distinguishing real videos from fake ones, given the severity of the malicious use of AI-generated videos [5]. Efforts from the academic community and private institutions have been made [6]. An example of these efforts is the Kaggle Deepfake Detection Challenge, organized by Facebook, aimed at finding new ways to detect deepfakes and combat this emerging threat [7]. At the beginning of these efforts, the most promising approaches indicated using CNNs (Convolutional Neural Networks), particularly EfficientNet [8]. The main approaches explored changes in spatial properties between different frames [9], as CNNs have initial layers that evaluate local spatial connections.

A more recent method of deepfakes detection uses the ViT (Vision Transformer) architecture, initially proposed by Dosovitskiy et al. [10]. The input image is divided into smaller fixed-size patches and transformed into input sequences. The self-attention mechanism evaluates the importance of elements, learns what is most important, and relates them, forgetting the less important ones. In the end, there is a series of transformer blocks (decoder) and a classifier, which, in this case, determines whether the video is a deepfake or not. The combination of technologies has produced promising results and led to significant improvements in classification performance using a combination of CNN and ViT (hereinafter referred to as CNN+ViT) [11].

Videos can contain subtle signals, such as movements and color changes, that reveal information about physiological functions, and the Eulerian video magnification (EVM) technique facilitates the enhancement of these signals, making them perceptible [12]. Fei et al. [5] presented a model that achieved promising results in identifying deepfakes by using the Eulerian motion magnification technique as a preprocessing step, amplifying the discrepancies in facial movement between authentic and falsified videos, achieving an average accuracy of 99.25% in four subsets of the FaceForensics++ dataset. They used an architecture based on a combination of CNN and LSTM. LSTM (Long Short-Term Memory) is a type of RNN (Recurrent Neural Network) that can remember certain parts of an input and forget others from one time step to the following [13]. Analogously to Fei et al. [5], this work intends to use the Eulerian motion magnification technique as a preprocessing step for frames, but we will use an architecture combining CNN and ViT.

This work hypothesizes that the use of the Eulerian video magnification technique will improve the deepfake classification metrics of using the CNN+ViT architecture. The experiments will compare the classification metrics of the methods in architectures with and without Eulerian video magnification techniques. The datasets for the training, validation, and testing of the proposed models will be subsets of the FakeAVCeleb dataset, which contains original videos and videos with deepfake techniques applied. The model evaluation will be performed using the accuracy metric, in addition to precision, recall, and the F1 score, and we will compare the results obtained with those of other works in the field.

The structure of the article is as follows: Section 2 explores related work. Section 3 introduces the FakeAVCeleb dataset. Section 4 describes the experimental methodology conducted. Section 5 details the results obtained and their analysis. Finally, Section 5 concludes the article with final considerations and presents suggestions for future work.

2 Related Work

The work of Fei et al. [5] presents a preprocessing method and a temporal-aware framework that can effectively distinguish real videos from AI-generated videos. The proposal involves enhancing facial movements with Eulerian magnification to differentiate real videos from fake ones. For this purpose, they used a CNN+LSTM architecture to take advantage of spatial and temporal features, identifying clues within the frame and in the relationship between different frames. They used four subsets of the FaceForensics++ dataset for method validation: DeepFakes, FaceSwap, Face2Face, and NeuralTexture. The results were competitive, contrasting with traditional pixel-based forensic approaches, with an average accuracy of 99.25%. One of the main conclusions highlighted that the AI-synthesized videos at the time generally left many clues regarding movement due to the high complexity of maintaining consistency in the temporal domain.

The work of Das, Negi, and Smeaton [2] aimed to identify whether a video is a deepfake using the result of EVM as a preprocessing method. They used three different techniques: SSIM, LSTM, and Heart Rate Estimation (HRE). As datasets, they use 400 DFDC Kaggle subsets and 30 self-authored datasets, generating five video datasets by altering the EVM parameters. SSIM and LSTM techniques showed that analyzing videos without preprocessing them was more accurate than that of videos preprocessed with EVM. The results of the heart rate estimation technique show that the original and fake videos were indistinguishable from the heart rate estimation technique. They concluded that the heart rate estimates between the original and deepfake videos were very close, with subtle changes only in the first decimal place, which was insufficient to detect deepfakes.

The work of Ciftci and Demir [14] aimed to detect the source of deepfake videos. A detection model based on clues left in the video movements was proposed, with an architecture similar to C3D. They use phase-based motion magnification and deep motion magnification methods. They performed model evaluations on the Forensics++ and FakeAVCeleb datasets. The results achieved an accuracy of 97.17% on Forensics++ and 94.03% on FakeAVCeleb. They also performed experiments with varying magnification levels, architectures, phase windows, minimum number of frames, and skin tones.

The work of Coccomini et al. [11] analyzes different solutions based on combinations of CNNs with different types of ViT. They used pre-trained EfficientNet B0 and Wodajo CNN as feature extractors. As datasets, they used FaceForensics++ and DFDC to train and evaluate the models. They proposed three models, and the proposed Convolutional Cross ViT achieved the best result with the EfficientNet B0 backbone, obtaining a mean accuracy of 80% in FaceForensics++ and an F1 score of 88.0% in DFDC. They also presented an effective voting scheme to deal with multiple faces in a video in the same video shot.

The work of Khalid, Tariq, and Woo [15] presents the FakeAVCeleb dataset, an audio-video multimodal deepfake detection dataset. After splitting the FakeAVCeleb dataset into train and test sets, considering racial and gender bias, they trained all three models mentioned in the following. The standard ResNet50 baselines were trained in audio and video datasets separately and then tested on the test dataset. Later, they trained the proposal ResNet50-based multimodal deepfake detector and tested it on test data, consisting of real and deepfake video and audio data for each fake type separately. They report the precision, recall, and F1 scores. The multimodal ResNet50 easily outperforms the detectors trained only on a single type of data, i.e., audio or video. The average F1 scores come out to be 40.98%, 88.50%, and 94.29% for ResNet50 (Audio), ResNet50 (Video), and multimodal

ResNet50 deepfake detector.

3 FakeAVCeleb dataset

For this study, we used subsets of the FakeAVCeleb dataset by Khalid, Tariq, and Woo [15] to train, validate, and test our proposed deepfake detection model. This dataset contains a mix of real and fake videos of celebrities, providing a diverse range of content. The FakeAVCeleb dataset has 500 authentic videos and 19,500 deepfakes videos. One of its main features is that it is a multimodal audio-video dataset, with the categories: Real audio and real video (500 videos), Real audio and fake video (9,000 videos), Fake audio and real video (500 videos), Fake audio and fake video (10,000 videos). They also divided the dataset into five ethnic groups: African (Black), South Asian (Indian), East Asian, Caucasian American, and Caucasian European. This variation was done to reduce bias, with balanced selection based on criteria such as gender, ethnicity, and age. In Figure 1, on the left are samples from the dataset that contain real videos, and on the right are samples with fake videos.

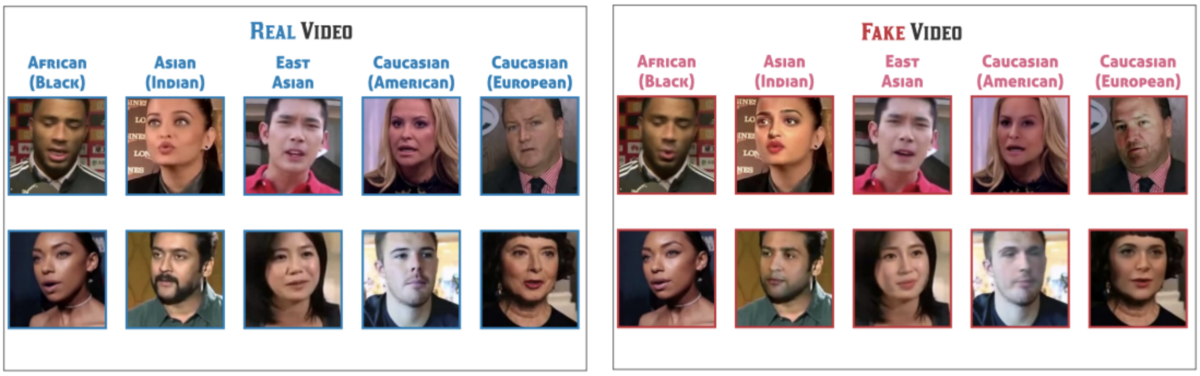


Figure 1. Samples from the dataset.

For this experiment, the two subsets we used are Caucasian European RealVideo-RealAudio and Caucasian European FakeVideo-RealAudio, and we only used video images, ignoring audio information.

4 Methods

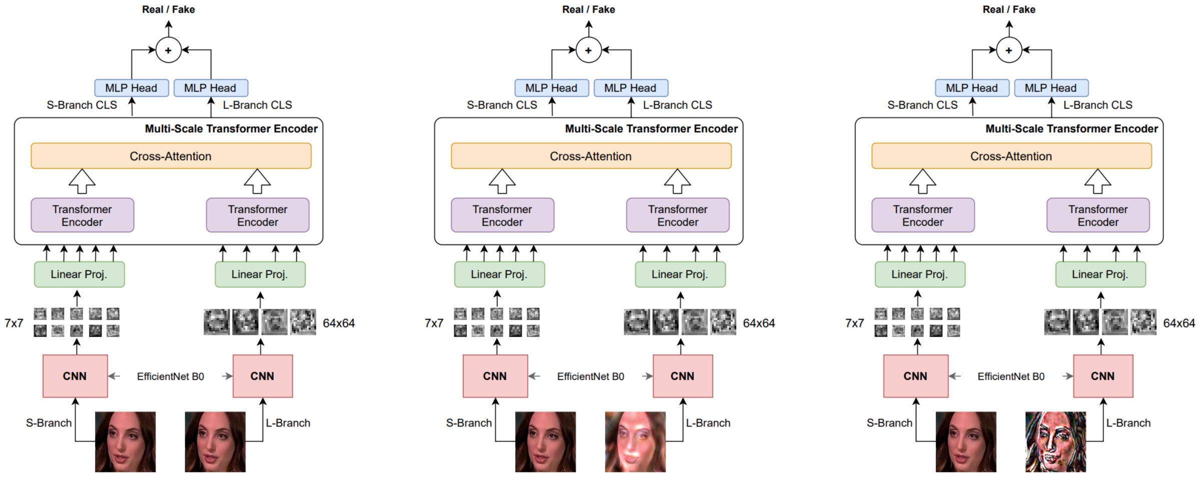
We propose the use of the Eulerian video magnification technique with a CNN+ViT model. The CNN+ViT model with which we choose to experiment is the Convolutional Cross ViT proposed by Coccomini et al. [11]. The model uses two distinct branches of Vision Transformers: the S branch, which deals with smaller patches, and the L branch, which has a wider receptive field, dealing with larger patches.

The Figure 2 (a) shows how the original architecture is, where they used the same images in the S and L branches of the model. We propose to add the Eulerian video magnification technique to the frame in the L branch, Figure 2 (b) and (c) shows the proposed architectures, and Table 1 specifies the different frames treatments.

The new architectures consist of the following: **EVM preprocessing module**: The application of the technique of Eulerian video magnification to enhance subtle cues between the frames. We use the PyEVM library, by Göhler [16], with the parameters shown at Table 1. **CNN Module**: A pre-trained EfficientNet B0 backbone extracts spatial features from the input frames. **ViT Module**: The extracted features are fed into a Vision Transformer to capture long-range dependencies and contextual information. How we use the Convolutional Cross ViT by Coccomini et al. [11] as the CNN+ViT model, there are two distinct branches of ViT, one that deals with smaller patches and another that has a wider receptive field to dealing with larger patches.

To ensure consistency and improve the performance of our model, we performed the following preprocessing steps on the datasets shown at Figure 3. The **input** is the video to classify as Real or Fake.

1. **Step 1** - Face detection: We used Multitask Cascaded Convolutional Networks (MTCNN) [17] from the FaceNet PyTorch library [18] to detect faces in each frame.
2. **Step 2** - Frames extraction: We decomposed the videos into individual frames using a Python script with the OpenCV library.



(a) Original architecture (Control) (b) Architecture EVM1 (c) Architecture EVM2

Figure 2. The different architectures used in the experiment. These images are adaptations of Convolutional Cross ViT by Coccomini et al. [11], and Khalid, Tariq, and Woo [15], with the characteristics of this study.

Table 1. Different treatments in the experiment

Treatment	Modification	Parameters of Eulerian video magnification			
		Mode	Amplification	Levels	Frequency ranges
Control	None				
EVM1	Apply EVM	COLOR	20	3	0.4Hz to 3Hz
EVM2	Apply EVM	MOTION	20	3	0.1Hz to 1Hz

3. **Step 3** - Crop faces: We cropped the frames to the box of each face using a script Python with the OpenCV library.
4. **Step 4** - Resize: We resized each cropped image of the faces to 224x224 pixels using a Python script with the OpenCV library.
5. **Step 5** - Application of EVM: We used three different treatments in the experiment. The first one is the Control, where we do not apply EVM, then the unmodified frame is returned. In the second and third ones, EVM1 and EVM2, we apply the EVM techniques using the PyEVM library of Göhler [16], with parameters from Table 1.
6. **Step 6** - Images are sent to the CNN + ViT model, in this study, to the Convolutional Cross ViT by Coccomini et al. [11]. In the Control architecture, the unmodified frame is passed to the S and L branches of the model, as shown at the Figure 2 (a). In the EVM1 architecture, the model uses the unmodified frame on the S branch, and the frame with EVM1 treatment is passed to the L branch, as shown in Figure 2 (b). Moreover, in the EVM2 architecture, the model uses the unmodified frame on the S branch, and the frame with EVM2 treatment is passed to the L branch, as shown in Figure 2 (c).

The **output** predicts whether the input video is Real or Fake.

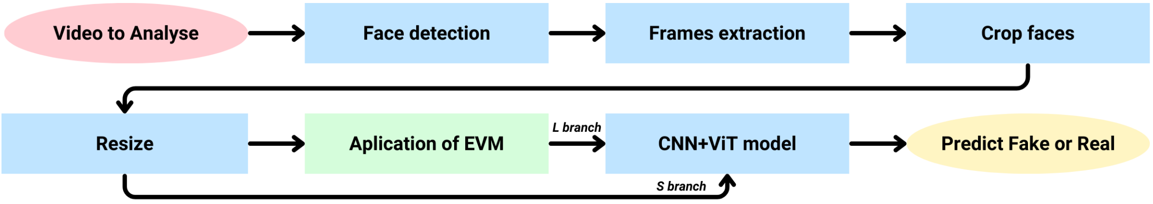


Figure 3. The steps of the detection on proposed architectures

The training procedure involved splitting the dataset into training, validation, and test sets in a 60:20:20 ratio.

We used a stochastic gradient descent (SGD) optimizer with a learning rate of 0.01 and a batch size of 16. We performed model training for 50 epochs at each architecture, with an early stopping of patience of 6 epochs based on validation loss to mitigate overfitting. Given the binary nature of the classification task (real versus fake), we utilized the PyTorch BCEWithLogitsLoss function, which combines binary cross-entropy loss with an integrated sigmoid activation layer. We used several metrics to evaluate the performance of trained models, including accuracy, precision, recall, and the F1 score. These metrics comprehensively assess the model's performance in detecting deepfakes.

We carried out the experiments on a machine equipped with two NVIDIA TITAN Xp with 12GB VRAM GPU each; Intel® Core™i9-9900K 3.60 GHz CPU; 32GB of RAM; and with the system, software, and libraries: Windows 10 Pro 22H2, Python 3.12.4, PyTorch 2.2.2, OpenCV 4.8.0, FaceNet PyTorch 2.5.3 and PyEVM 0.4.2.

5 Results and Discussion

We evaluated The performance of the proposed deepfake detection models using two subsets of the FakeAVCeleb dataset: Caucasian European RealVideo-RealAudio and Caucasian European FakeVideo-RealAudio. Table 2 summarizes the results of the different configurations, including Control (no EVM), EVM1, and EVM2 treatments. The metrics reported include accuracy, precision, recall, and the F1 score.

Table 2. Results of the different architectures in the experiment

Metric	Control	EVM1	EVM1 vs Control	EVM2	EVM2 vs Control
Accuracy	95,42%	**97,46%**	+2,04%	97,20%	+1,78%
F1	96,37%	**98,01%**	+1,64%	97,84%	+1,47%
Precision	98,35%	**98,80%**	+0,44%	97,27%	-1,09%
Recall	94,47%	97,23%	+2,77%	**98,42%**	+3,95%

The results indicate that applying the Eulerian video magnification (EVM) technique improves the model's performance in most metrics. The EVM1 configuration achieved the highest accuracy of 97.46%, a gain of 2.04% compared to the Control configuration, which had an accuracy of 95.42%. EVM2 also showed an increase of 1.78% in the accuracy score. The results suggest that the EVM technique can enhance the model's ability to distinguish between real and fake videos by amplifying subtle movements and inconsistencies.

Although EVM1 showed a slight improvement in the precision score of 0.44% compared to the Control treatment, EVM2 showed a decrease of 1.09%. However, EVM2 also showed an improvement in the recall of 3.95%. The decreasing precision and the increase in recall suggest that the proposed model with the EVM2 treatment is better at detecting true positives than at avoiding false positives compared to the Control treatment.

In addition to the architectures presented in Figure 2, we tested other configurations by applying the EVM1 and EVM2 treatments exclusively to the S branch and both branches. However, the results from these configurations were inferior to the Control treatment in all evaluated metrics, including accuracy, precision, recall, and the F1 score. They confirm that not all architectures will benefit from using Eulerian video magnification techniques. This statement can be reinforced by observing the results of performance improvements in the architecture by Fei et al. [5] and performance decreases in the architecture by Das, Negi, and Smeaton [2].

The numerical results are compared with the Control treatment, which was rigorously trained, validated, and tested with the same criteria as the EVM1 and EVM2 treatments, so neither would benefit nor be affected. It is not directly compared with other works because the datasets or input configurations used in the training, validation, and test differ. Like the 99.25% average accuracy obtained by Fei et al.[5] in the FaceForensics++ dataset, and Ciftci and Demir [14] who obtained an accuracy of 97.17% in FaceForensics++ and 94.03% in FakeAVCeleb with different input configurations.

In a separate test, the entire FakeAVCeleb dataset was processed on the Google Colab Pro + platform in the cloud with an NVIDIA A100 GPU. However, the persistence service on the disk exhibited instability due to the high volume of disk read/write operations during frame extraction and model training. Consequently, we did not explore these additional results further.

6 Conclusions

The findings indicate that EVM techniques, particularly in the proposed architecture EVM1, can enhance the performance of deepfake detection models in several metrics. The application's specific requirements should guide the choice between the proposed architectures EVM1 and EVM2. For scenarios where maximization of the detection of true positives is critical, EVM2 may be preferred despite its higher rate of false positives. In contrast, EVM1 offers a more balanced improvement in precision and recall, making it suitable for applications where both metrics are equally important.

The tests using only EVM, in the S and L branches, on the Convolutional Cross ViT architecture by Coccomini et al. [11] showed worse results compared to Control, collaborating with the findings of Das, Negi, and Smeaton [2] with other architectures. However, the combination of frames with and without Eulerian video magnification in the L and S branches shows the potential to improve the model's performance. Improvements were also observed using Eulerian video magnification in the architectures proposed by Fei et al.[5].

In conclusion, applying Eulerian video magnification techniques has proven to be a potential enhancement for deepfake detection in some architectures. It offers improvements in key performance metrics and demonstrates the possibility of application in real-world scenarios.

In this study, we did not use audio information; however, there is a significant interest in exploring its potential for multimodal detection in future research. Integrating audio data could provide additional insight and improve the accuracy of our detection methods. Another avenue for future work is to test different parameters in Table 1 for Eulerian video magnification. Using various parameter settings, we aim to optimize the magnification process and enhance the overall performance of our approach. In addition, future research will consider the inclusion of diverse ethnicities and genders. It will help us to ensure that the detection methods are robust and generalizable across different demographic groups. The interaction between these variables and the results will be statistically analyzed to understand their impact on the detection accuracy.

Acknowledgements. The authors would like to thank FAPES/UnAC (No. FAPES 1228/2022 P 2022-CD0RQ, No. SIAFEM 2022-CD0RQ) for the financial support provided through the Sistema UniversidaES. Professor Komati thanks CNPq for the DT-2 grant (nº 302726/2023-3) and project nº407742/2022-0, also thanks FAPES for project nº 1023/2022 P:2022-8TZV6.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] M. Koopman, A. Macarulla Rodriguez, and Z. Geradts. Detection of deepfake video manipulation. In *Proceedings of the 20th Irish Machine Vision and Image Processing conference (IMVIP)*, pp. 133–136, 2018.

[2] R. Das, G. Negi, and A. F. Smeaton. Detecting deepfake videos using euler video magnification. *Electronic Imaging*, vol. 33, n. 4, pp. 1–7, 2021.

[3] J. W. Seow, M. K. Lim, R. C. Phan, and J. K. Liu. A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing (Amsterdam)*, vol. 513, pp. 351–371, 2022.

[4] Y. Mirsky and W. Lee. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*, vol. 54, n. 1, pp. 1–41, 2021.

[5] J. Fei, Z. Xia, P. Yu, and F. Xiao. Exposing AI-Generated videos with motion magnification. *Multimedia Tools Appl.*, vol. 80, n. 20, pp. 30789–30802, 2021.

[6] D. Wodajo and S. Atnafu. Deepfake video detection using convolutional vision transformer. *CoRR*, vol. abs/2102.11126, 2021.

[7] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer. The deepfake detection challenge (DFDC) dataset. *CoRR*, vol. abs/2006.07397, 2020.

[8] L. Jiang, Z. Guo, W. Wu, Z. Liu, Z. Liu, C. C. Loy, S. Yang, Y. Xiong, W. Xia, B. Chen, P. Zhuang, S. Li, S. Chen, T. Yao, S. Ding, J. Li, F. Huang, L. Cao, R. Ji, C. Lu, and G. Tan. Deeperforensics challenge 2020 on real-world face forgery detection: Methods and results. *CoRR*, vol. abs/2102.09471, 2021.

[9] Y.-J. Heo, Y.-J. Choi, Y.-W. Lee, and B.-G. Kim. Deepfake detection scheme based on vision transformer and distillation. *CoRR*, vol. abs/2104.01353, 2021.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for

image recognition at scale. *CoRR*, vol. abs/2010.11929, 2021.

[11] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi. Combining efficientnet and vision transformers for video deepfake detection. In S. Sclaroff, C. Distante, M. Leo, G. M. Farinella, and F. Tombari, eds, *Image Analysis and Processing – ICIAP 2022*, pp. 219–229, Cham. Springer International Publishing, 2022.

[12] H. Lauridsen, S. Gonzales, D. Hedwig, K. Perrin, C. Williams, P. Wrege, M. Bertelsen, M. Pedersen, and J. Butcher. Extracting physiological information in experimental biology via eulerian video magnification. *BMC biology*, vol. 17, pp. 103, 2019.

[13] S. Russell and P. Norvig. *Inteligência Artificial*. Grupo GEN, 4 edition, 2022.

[14] U. A. Ciftci and I. Demir. How do deepfakes move? motion magnification for deepfake source detection. *arXiv:2212.14033*, 2022.

[15] H. Khalid, S. Tariq, and S. S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *CoRR*, vol. abs/2108.05080, 2021.

[16] V. Göhler. Eulerian Video Magnification for Python. https://github.com/vgoehler/PyEVM, 2020.

[17] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, vol. 23, n. 10, pp. 1499–1503, 2016.

[18] T. Esler. Face Recognition Using Pytorch. https://github.com/timesler/facenet-pytorch, 2023.