Securing Digital Integrity: Proposed Comprehensive Framework for Deepfake Detection and BlockChain Validation

Anant Jain[1][0009−0000−2859−8911] anant.07019011921@ipu.ac.in

Adamya Gaur[1][0009−0005−6693−7984]adamya.04719011921@ipu.ac.in
Gauranshi Gupta[1][0009−0001−9918−3120]gauranshi.00719011921@ipu.ac.in
Shubhangi Mishra[1][0009−0002−1720−6921]shubhangi.09919011721@ipu.ac.in
Rahul Johari*[1][0000−0002−7675−8550] rahul@ipu.ac.in
Deo Prakash Vidyarthi[3][0000−0003−4151−0552]dpv@mail.jnu.ac.in

[1] SWINGER : **S**ecurity, **W**ireless, **I**oT **N**etwork **G**roup of **E**ngineering and **R**esearch, University School of Automation and Robotics(USAR), Guru Gobind Singh Indraprastha University, Delhi, India
[2] School of Computer and Systems Sciences, Parallel and Distributed System Lab JNU, Delhi, India

Abstract. This paper proposes a comprehensive solution to combat the growing threat of Deepfake technology, employing Convolutional Neural Networks (CNNs) and BlockChain. CNNs analyze video frames for anomalies indicative of Deepfake manipulation, while BlockChain ensures content integrity through timestamping and IPFS hashes. Leveraging transfer learning with EfficientNet-B1 architecture and dropout layers to prevent overfitting, the CNN model attains a 98.53 percent accuracy on testing data. Smart contracts are utilized to store social media content and AI verification results. The methodology integrates private storage on social media platforms, AI-based CNN verification, BlockChain timestamping, smart contract verification, and continuous improvement. By synergizing AI and BlockChain technologies, our approach aims to bolster defenses against deceptive content proliferation, enhancing the reliability of online media ecosystems.

Keywords: BlockChain · IPFS · Convolutional Neural Network · Deepfake

1 Introduction

Deepfake is a potent Artificial Intelligence-powered threat wherein Artificial Intelligence produces realistic fake content, often in the form of video or audio, impersonating real individuals. This technology enables malicious actors to manipulate and spread deceptive content, posing severe risks to individuals and

* corresponding author

society. Deep Fakes can deceive, defame, and incite misinformation, contributing to identity theft, fraud, and political manipulation.

AI algorithms, particularly Convolutional Neural Networks (CNNs), play a pivotal role in detecting fake content by analyzing video frames. CNNs excel at identifying anomalies within individual frames but detection of Deepfake gets harder when the input images are of low quality, as the distinguishing features visibility reduces. The dataset used in the experiment consists of low-quality videos, which further aids in creating a more robust model. During analysis, the network identifies patterns, textures, and facial features unique to the subject. Temporal dependencies were also considered as CNNs evaluated the consistency of these features across frames. By learning from extensive datasets, the algorithm distinguished between genuine and manipulated content, uncovering discrepancies in pixel-level details, and facial expression. This approach empowers AI to expose inconsistencies indicative of Deepfake manipulation, contributing to the battle against deceptive digital content.

BlockChain is a decentralized and tamper-resistant ledger technology that records transactions across a network of computers. It utilizes hashing techniques to ensure data integrity, immutability, and transparency. Each block in the chain contains a timestamp and a hash of the previous block, forming a secure and chronological sequence of records. In the context of combating fake content, BlockChain serves as a second layer of validation. Timestamps on the BlockChain provide an immutable record of when the original video was verified. By comparing timestamps between the original and suspected fake videos, inconsistencies can be detected, revealing potential manipulations. The use of IPFS (InterPlanetary File System) hashes on the BlockChain ensures content integrity. If the IPFS hash of the suspected fake video differs from the hash recorded for the original on the BlockChain, it indicates alterations. This dual validation, through timestamps and IPFS hashes, enhances content verification, reinforcing the authenticity of original content and detecting potential fakes.

2 Problem Statement

The rise of Deepfake technology poses a critical threat to individuals and society, allowing malicious actors to create malicious intent. This technology exposes individuals and society to severe risks, including deception, defamation, and the spread of misinformation, contributing to identity theft, fraud, and political manipulation. This research paper addresses the need for robust solutions to combat Deepfake by proposing a comprehensive approach that integrates advanced AI algorithms, specifically Convolutional Neural Networks (CNNs), with the tamper-resistant capabilities of BlockChain technology.

This research focuses on leveraging CNNs to analyze video frames, identifying patterns, textures, and facial features unique to subjects. Temporal dependencies are considered, enabling the algorithm to evaluate consistency across frames and expose inconsistencies indicative of Deepfake manipulation. Additionally, explored the integration of BlockChain as a second layer of validation, utilizing

timestamps and IPFS hashes to ensure content integrity. Through this research aimed to contribute to the ongoing battle against deceptive digital content and fortify the resilience of online media ecosystems.

3 Literature Survey

In [1], the author proposes a Deepfake detection system to combat manipulated videos on platforms like YouTube, mitigating risks of political manipulation and social unrest. Utilizing blockchain for transparency, multiple content providers collaborate, each employing their deep-learning model for detection. Integrated voting combines results for objectivity, with user involvement enhancing accuracy through collective intelligence. The system fosters cooperation among companies with competing interests, establishing a consortium contract environment to combat Deepfake threats effectively.

In [2], the challenges posed by Artificial Intelligence-Generated Content (AIGC) are addressed in terms of service latency, security, and trustworthiness. It introduces a novel BlockChain-empowered framework for managing the lifecycle of edge AIGC products. The proposed Proof-of-AIGC protocol safeguards ownership. Additionally, a multi-weight subjective logic-based reputation scheme assists AIGC producers in selecting trustworthy edge service providers. Numerical results validate the effectiveness of the approach, and the survey concludes by highlighting avenues for future research in the intersection of AIGC, BlockChain, and edge networks.

In [3], this paper reviews the proliferation of Deepfake technology in creating convincingly fake videos and images, often targeting public figures. It discusses notable research contributions and popular Deepfake tools, presenting two taxonomies for categorization. The analysis compares these models and tools based on algorithms, datasets, and accuracy. Despite advancements, challenges encountered include difficulties in detecting low-quality Deepfake, time-consuming implementation of detection techniques in real-world settings, and training Deepfake models on limited data for specific characters.

In [4], the author discusses that misinformation poses a significant threat across social, economic, and politics, fueled by the rapid dissemination facilitated by AI technologies like deep learning. To combat this, a novel approach integrates BlockChain and AI. BlockChain ensures content authenticity, while AI models detect fake news. Challenges include integrating these technologies seamlessly and ensuring transparency. This synergy marks a promising avenue, albeit with ongoing research to refine implementation and address potential limitations.

In [5], the author explores that Deepfake technology presents a critical challenge in discerning between genuine and fraudulent digital content online. Using AI, machine learning, and deep learning, Deepfake produces hyper-realistic yet deceitful material. Addressing this issue requires innovative approaches to reliably distinguish between authentic and fake multimedia. This paper reviews Deepfake technology, its associated threats, challenges, and future directions, emphasizing the need for robust detection methods. Critically, it highlights key

techniques for securing digital integrity, crucial amidst the evolving landscape of fake content dissemination.

In [6], the author offers a comprehensive examination of social media data privacy policies and introduces a trust index model leveraging blockchain technology to combat fake news and Deepfake. Privacy breaches on social media have become a pressing concern, especially highlighted during the COVID-19 infodemic. The proposed model aims to restore trust and reliability by providing transparent and secure information dissemination. Challenges include integrating blockchain seamlessly into social media platforms and ensuring user adoption. This model represents a crucial step towards fostering a more trustworthy online environment amidst rampant misinformation.

In [7], the surge in tempered electronic content facilitated by advanced artificial intelligence, particularly Deepfake, is addressed. The study highlights the inadequacy of current solutions in tracking digital media's history and provenance. It explores the potential of BlockChain technology, including innovations like Smart Contracts and Hyperledger Fabric, to combat the Deepfake problem. The paper proposes a BlockChain-dependent framework for preserving data integrity and preventing Deepfake. The study also delves into the current state of Deepfake video detection research, covering the generating process, detection algorithms, and benchmarks.

In [8], prevailing Deepfake literature predominantly emphasizes AI-assisted detection methods, overlooking the crucial aspect of content 'authentication.' This paper diverges by exploring BlockChain integration for tamper-proof evidence of genuine content, addressing a significant gap in current research. Despite scattered efforts in this direction, a comprehensive guide for the effective adoption of BlockChain against Deepfake is absent. The paper provides insights into potential use cases and solutions, offering a holistic guide to leverage BlockChain effectively in navigating the challenges posed by Deepfake technology.

In [9], the paper explores the influence of digital deception, like Deepfake and misinformation, via the Internet and social media. It suggests leveraging distributed ledger technologies, specifically BlockChain, to combat such deception. By offering transparent and immutable records in a peer-to-peer framework, DLTs enhance data authenticity and traceability. The survey delves into DLT applications, tackles associated challenges, and suggests future research directions to fortify resilience against cyber threats.

In [10], the paper addresses the proliferation of Deepfake, driven by advances in AI and deep learning, posing threats to reality and trust. Emphasizing the need for proof of authenticity, the paper critiques current solutions for lacking robust history tracking. Proposing a solution, it introduces an Ethereum smart contract framework using IPFS hashes to trace digital content back to its origin, applicable to various digital forms. The approach prioritizes traceability to trusted sources for content authenticity. The open-sourced smart contract code on Github encourages collaboration to combat challenges posed by fake digital content.

4 Methodology Adopted

4.1 Flowchart

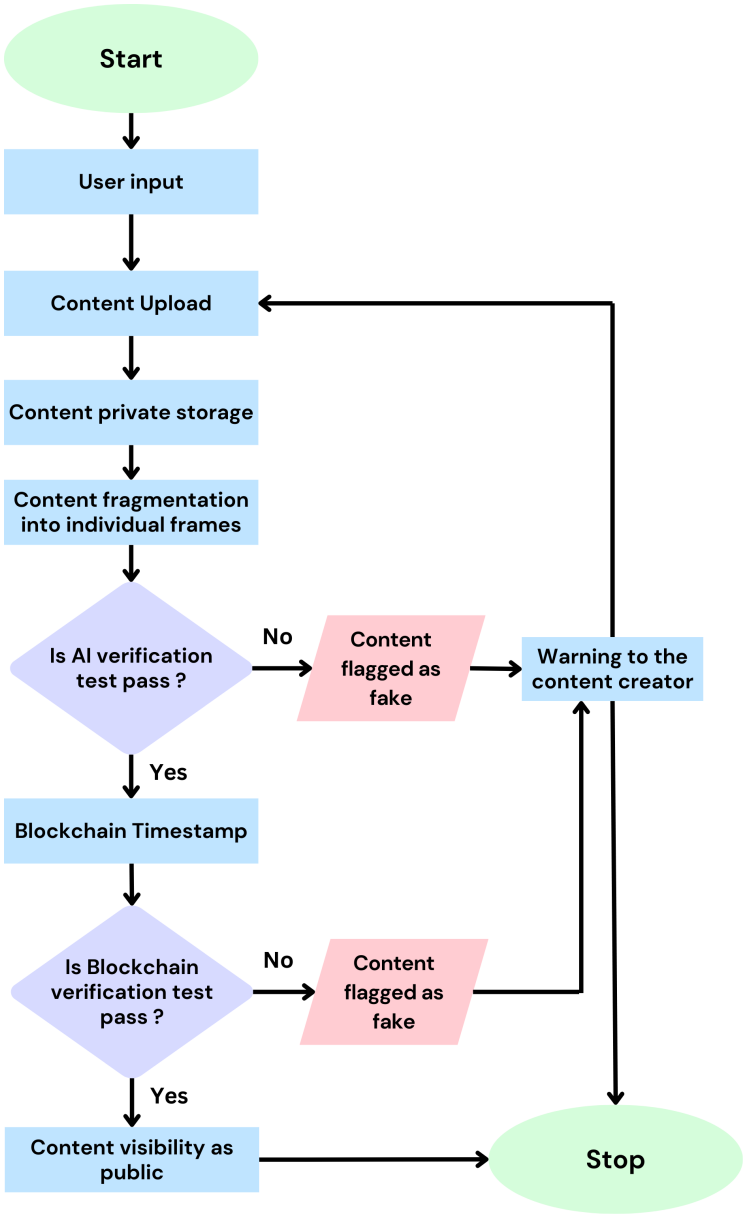

Fig. 1: Flow of the Deepfake content detection system

Fig 1 shows the architecture of AI and BlockChain-based model for Deepfake detection.

4.2 Simulation Environment

The proposed smart contract for the BlockChain was simulated and executed using the Remix IDE, a powerful development tool. Remix IDE facilitated the compilation, deployment, and testing of the smart contract through its user-friendly interface. Google Colab was employed to simulate the process of Deepfake detection. This cloud-based platform allowed for the execution of the Deepfake detection algorithm collaboratively and efficiently. The combination of Remix IDE for BlockChain simulation and Google Colab for Deepfake detection provided a comprehensive testing environment for the proposed solution.

4.3 Screenshot of BlockChain Simulation in Remix IDE

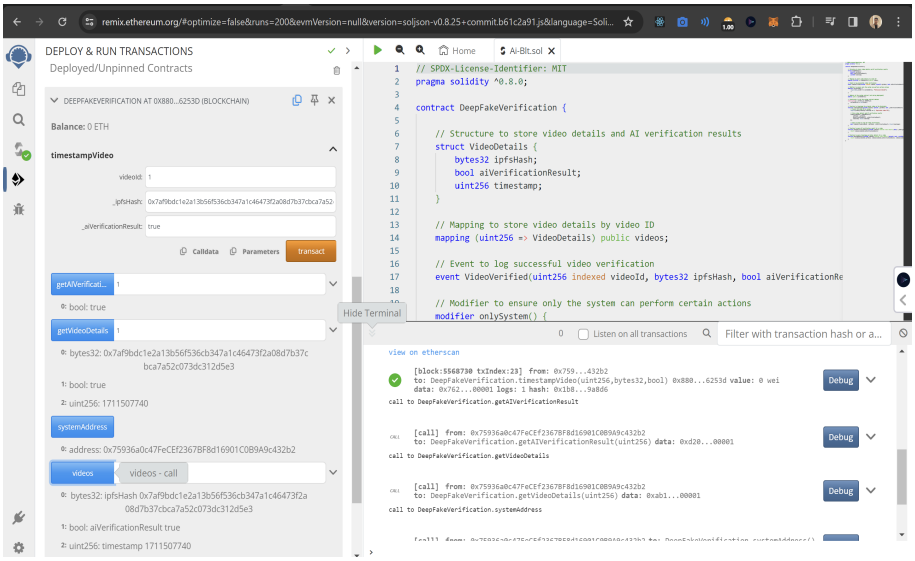

Fig. 2: Implementation of solidity code in Remix IDE

Figure 2 highlights three primary functions within the smart contract: timestampVideo(), responsible for storing the video hash and AI verification result on the BlockChain, along with several read functions such as getAIVerification() which takes videoId as input and returns true or false that AI verification for a video is done or not and getVideoDetails() that also takes videoId as input and returns video hash as output, enabling retrieval of results from the BlockChain.

4.4 Dataset and Pre-processing

Data used for the training of a model plays a crucial role in its performance. Data from DeepfakeTIMIT [11] and VidTIMIT [12] were used for training the model. It contains 10 face-swapped videos of each individual and no audio content. For creating the DeepfakeTIMIT dataset 16 similar-looking pairs of people were selected manually from the VidTIMIT database. As there were 10 videos per person in the VidTIMIT database, 320 videos corresponding to each version were generated which resulted in 620 total videos with faces swapped. Different methodologies were used for the generation of the Deepfake videos. Further, frames were extracted out from the videos, enabling the model to perform a frame-by-frame analysis of the Deepfake content. Figure 3 displays the real and fake frames of 5 such individuals. To create a robust model, the frames were

Fig. 3: Real and fake images of individuals[11][12]

further processed using different techniques in the training data. This included random horizontal, and vertical shifting and zoom in and out up to 10%. Shearing and horizontal flipping of images were also done. All the tensor values were standardized. The data was further divided into 80% training, 10% validation, and 10% testing sets.

5 Results

A transfer learning approach was used in the study, utilizing and further customizing EfficientNet-B1 [13] architecture. To prevent overfitting by the model, dropout layers were added to the model. Internally, the model is based on a Convolutional Neural Network (CNN), where a filter convolves over the image, leading to an internal representation that is useful for extracting important features. Figure 4 displays the architecture used.

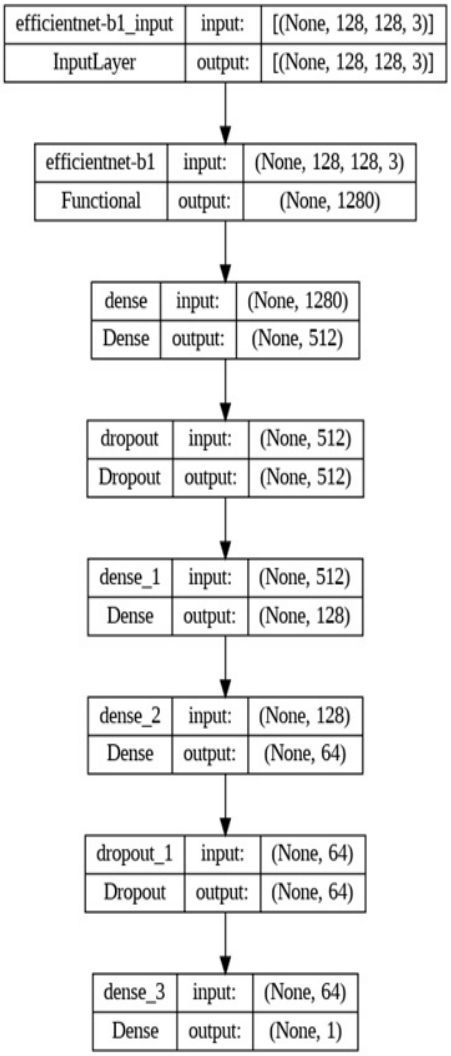

Fig. 4: Deep learning model architecture

In the output layer, the sigmoid activation function is used, which provides the likelihood of a frame being real or fake. It is calculated as:

$$S(x) = \frac{1}{1+e^{-x}}$$

Where e is Euler's number and x is the input value. Using the aggregate of these scores and threshold values, the final authenticity score is calculated. 5 fold cross-validation has been conducted on the model, to check for overfitting shown in Figure 5. The proposed model achieved a 98.53% accuracy score on the

testing set. Figure 6 displays the accuracy versus epoch and loss versus epoch line plots.

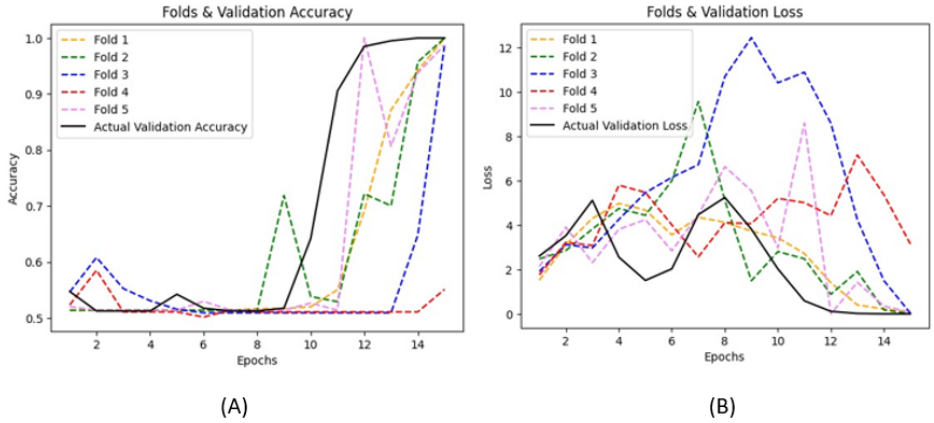

Fig. 5: 5 fold cross validation (A) Accuracy vs Epochs curves and (B) Loss vs Epochs curves for training and validation

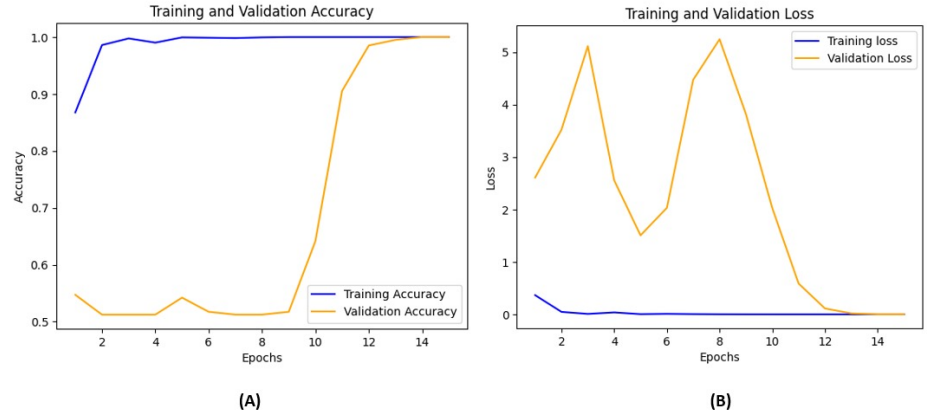

Fig. 6: (A) Accuracy vs Epochs curves and (B) Loss vs Epochs curves for training and validation

6 Conclusion and Future Work

A smart contract was successfully developed and tested to store timestamped videos, fetch AI test results, and retrieve video details. The smart contract, im-

plemented on Remix IDE, demonstrated robust functionality. Additionally, the Deepfake detection system, utilizing CNN on Google Colab, yielded an impressive accuracy of 98.53 percent. This integrated solution showcases the efficacy of combining BlockChain and AI in combating Deepfake, offering a reliable framework for authenticating and securing digital content.

In future work, the research will expand its focus beyond visual content to include audio content, as the current model primarily analyzes image frames from videos. Additionally, the study will be extended to develop a fully integrated Deep Fake detection system, combining BlockChain and AI technologies seamlessly. This integrated system will encompass all necessary components to detect and authenticate both visual and audio content, ensuring a comprehensive approach to combat Deepfake.

7 References

1. Choi, Nakhoon, and Heeyoul Kim. "DDS: Deepfake Detection System through Collective Intelligence and Deep-Learning Model in BlockChain Environment." Applied Sciences 13, no. 4 (2023): 2122.
2. Liu, Yinqiu, Hongyang Du, Dusit Niyato, Jiawen Kang, Zehui Xiong, Chunyan Miao, and Abbas Jamalipour. "BlockChain-Empowered Lifecycle Management for AI-Generated Content (AIGC) Products in Edge Networks." arXiv preprint arXiv:2303.02836 (2023).
3. Mukta, Md Saddam Hossain, Jubaer Ahmad, Mohaimenul Azam Khan Raiaan, Salekul Islam, Sami Azam, Mohammed Eunus Ali, and Mirjam Jonkman. "An investigation of the effectiveness of Deepfake models and tools." Journal of Sensor and Actuator Networks 12, no. 4 (2023): 61.
4. Seneviratne, Oshani. "Blockchain for social good: Combating misinformation on the web with AI and blockchain." In Proceedings of the 14th ACM Web Science Conference 2022, pp. 435-442. 2022.
5. Sharma, Mridul, and Mandeep Kaur. "A review of Deepfake technology: an emerging AI threat." Soft Computing for Security Applications: Proceedings of ICSCS 2021 (2022): 605-619.
6. Jing, Tee Wee, and Raja Kumar Murugesan. "Protecting Data Privacy and Prevent Fake News and Deepfake in Social Media via Blockchain Technology." In Advances in Cyber Security: Second International Conference, ACeS 2020, Penang, Malaysia, December 8-9, 2020, Revised Selected Papers 2, pp. 674-684. Springer Singapore, 2021.
7. Rashid, Md Mamunur, Suk-Hwan Lee, and Ki-Ryong Kwon. "BlockChain technology for combating Deepfake and protect video/image integrity." 24, no. 8 (2021): 1044-1058.
8. Yazdinejad, Abbas, Reza M. Parizi, Gautam Srivastava, and Ali Dehghantanha. "Making sense of BlockChain for AI Deepfake technology." In 2020 IEEE Globecom Workshops (GC Wkshps, pp. 1-6. IEEE, 2020).
9. Fraga-Lamas, Paula, and Tiago M. Fernandez-Carames. "Fake news, disinformation, and Deepfake: Leveraging distributed ledger technologies and

BlockChain to combat digital deception and counterfeit reality." IT professional 22, no. 2 (2020): 53-59.
10. Hasan, Haya R., and Khaled Salah. "Combating Deepfake videos using BlockChain and smart contracts." IEEE Access 7 (2019): 41596-41606.
11. Korshunov, Pavel, and Sébastien Marcel. "Deepfake: a new threat to face recognition? assessment and detection." arXiv preprint arXiv:1812.08685 (2018). https://zenodo.org/records/4068245
12. Sanderson, Conrad, and Brian C. Lovell. "Multi-region probabilistic histograms for robust and scalable identity inference." In Advances in Biometrics: Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings 3, pp. 199-208. Springer Berlin Heidelberg, 2009.
13. Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." In International conference on machine learning, pp. 6105-6114. PMLR, 2019.