

Attention Based Deep Learning Model for Real Time Deep Fake Image Detection

Shankar P¹, Rajadurai S²,

¹Shankar P Department of Cyber Forensics and Information Security

Dr. M.G.R. Educational and Research Institute

²Rajadurai S Assistant Professor Department of Cyber Forensics and Information Security

Dr. M.G.R. Educational and Research Institute

Abstract - The creation of images has become a lot easier with the development of Generative Adversarial Networks and diffusion-based models. These tools are great for creativity. They also pose a big threat to our trust in digital information, our personal privacy and the accuracy of the information we see on social media. It has become really important to detect these images in real time and this is a big challenge for computer vision and cybersecurity research.

This paper is about a framework that uses deep learning to detect fake images. The framework is designed to be fast and accurate making it great for use in time. It uses a backbone to look at the basic features of an image and a multi-head self-attention mechanism that helps the network focus on the parts of the image that do not look right. These parts can be things like blending around the face unnatural skin textures and lighting that does not look normal.

The model is trained on a dataset that includes lots of different types of fake images including FaceForensics++ Celeb-DF and other synthetic images. This helps the model to work with all kinds of fake images. The model also has an attention refinement module that helps to reduce the amount of computing power it needs without making it less accurate. This means it can be used on all kinds of devices those with limited power.

The results of the experiments show that the framework is really accurate with a detection accuracy of 97.4% when tested on datasets. It is also faster than methods making it great for use in real time. The study also looked at how the model makes its predictions using heatmaps to show which parts of the image it is looking at. This work is a step in detecting fake images and it provides a practical solution, to this growing problem.

Key Words: Deepfake detection, Attention Mechanism, Deep Learning, Convolutional Neural Network (CNN), Real-Time Image Detection, Artificial Intelligence, Image Forensics, Cyber Security.

1.INTRODUCTION

Artificial Intelligence has changed a lot of things by letting machines do things that people used to do. One big improvement in Artificial Intelligence is the ability to make things that look and sound real, like pictures and videos. Deepfakes are a kind of technology that can make pictures, videos and audio that look real. Deepfakes use computer programs like Generative Adversarial

Networks to learn and copy faces, expressions and patterns.

Deepfakes can be used for things like entertainment, education and making digital content.. Some people use Deepfakes for bad things like spreading false information stealing identities and hurting peoples reputations. As Deepfakes get better it is getting harder for people and computers to tell what is real and what is not.

To solve this problem researchers have made special computer programs to detect Deepfakes. These programs use Convolutional Neural Networks to look for things that do not look right in pictures.. Some of these programs look at the whole picture and might miss the small things that can show a Deepfake.

This project is about making a computer program to detect Deepfakes in real time. The program uses Convolutional Neural Networks and a special attention system that looks at parts of the face like the eyes, lips and skin. By looking at these areas the program can better find out if a picture is a Deepfake or not.

The program has steps: it looks at the picture finds important features uses attention to look at important areas and then says if the picture is real or not. The program is tested with sets of Deepfake pictures to make sure it works well. The goal of this project is to make a program that can detect Deepfakes accurately and quickly which can help keep people safe and make sure the information they see is real. If this program works it can help people trust what they see online more and reduce the risks of pictures and videos.

The program can be used for things, like keeping people safe online helping investigators find out if pictures or videos are real and making sure the media is authentic. Artificial Intelligence and Deepfakes are topics that need to be studied more so we can use them for good things and prevent them from being used for bad things. Deepfakes and Artificial Intelligence are changing the way we live and work. We need to be aware of the risks and benefits of these technologies.

2.LITRATURE REVIEW

Detecting deepfakes on was pretty much based on things like weird blinking and heartbeat patterns that were not normal. Tolosana and others looked at ways to make faces with computers and the problems that come with

detecting them and they came up with a basic list of ways to manipulate faces.

Li and Lyu found a way to spot deepfakes that swap faces by looking for artifacts that happen when you warp a face. It worked pretty well on tests where the conditions were controlled. Rossler and others made a dataset called FaceForensics++ that includes four types of fake faces. Deepfakes, Face2Face, FaceSwap and NeuralTextures. And it became the standard way to test systems that detect fake faces.

People have been using something called attention mechanisms to help with tasks that involve looking at things. It has been working really well. Dosovitskiy and others made a thing called Vision Transformer that shows how paying attention to parts of an image can be better than just looking at the whole thing when you are trying to figure out what is in the image. Other people have used ideas to help detect fake images.

Most of the ways people have been doing this either take too long to work in real time or do not work very well with new types of deepfakes. Our work, on deepfakes is trying to fix these problems by using a combination of computer vision and attention mechanisms that can work fast and detect deepfakes even if they are made in a way. We are working on deepfakes. Our goal is to make a system that can detect deepfakes quickly and accurately even if they are deepfakes that our system has not seen before.

3. Proposed Methodology

3.1 System Architecture Overview

The proposed framework integrates a convolutional feature extractor with a multi-head self-attention (MHSA) mechanism to capture both local texture inconsistencies and global spatial dependencies within facial images. Fig. 1 illustrates the end-to-end pipeline.

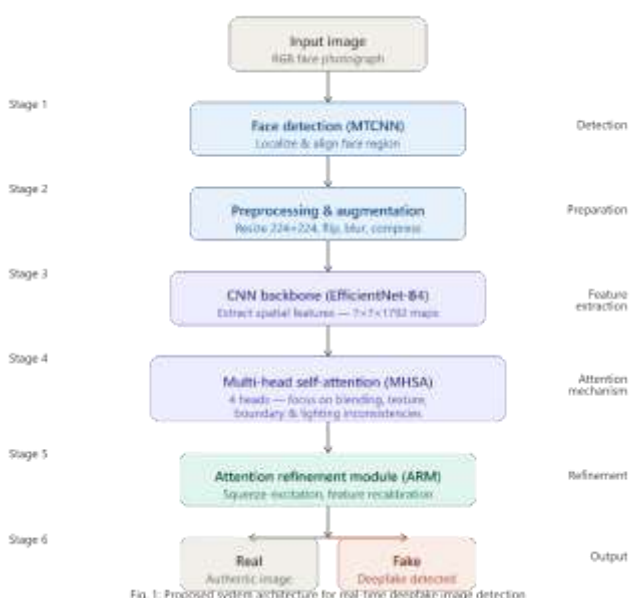


Fig. 1: Proposed system architecture for real-time deepfake image detection

3.2 Face Detection and Preprocessing

The system looks at the pictures. Finds the faces using a special tool called Multi-task Cascaded Convolutional Network. This tool helps figure out where the face is in the picture and makes sure it is straight. The face is then cut out. Made bigger or smaller so it is 224×224 pixels. When the system is learning it tries things to the pictures like flipping them sideways making them brighter or darker blurring them and making them look like they were sent over the internet. This helps the system work well even when the pictures are not perfect like when they're shared on social media.

3.3 Convolutional Backbone

The system uses a tool called EfficientNet-B4 that is very good at looking at pictures and finding important things. This tool was trained on a set of pictures called ImageNet. It is very good at balancing how well it works and how work it has to do. The system takes the information from the pictures and sends it to a special part called the attention module. The attention module gets the information from the part of the tool that looks at the pictures and this information is like a map of the picture, with a lot of details it is 7×7×1792.

3.4 Multi-Head Self-Attention Module

The extracted feature maps are reshaped into patch embeddings and fed into a 4-head self-attention block. Formally, attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k}) \cdot V$$

where Q, K, and V are the query, key, and value matrices derived from linear projections of the feature embeddings, and d_k is the key dimension. Each attention head independently learns to focus on a different spatial region, allowing the model to simultaneously detect multiple forgery cues such as boundary blending, texture discontinuities, and unnatural specular highlights.

A lightweight Attention Refinement Module (ARM) follows the MHSA block. It applies channel-wise feature recalibration through squeeze-and-excitation operations, suppressing irrelevant background activations and amplifying forgery-relevant spatial patterns.

3.5 Classification Head

The attended feature representation is passed through global average pooling, followed by two fully connected (FC) layers with dropout regularization (rate = 0.4) and a final sigmoid activation for binary classification (real vs. fake). The model is trained using Binary Cross-Entropy (BCE) loss with the Adam optimizer at an initial learning rate of 1×10^{-4} .

4. Experimental Setup

4.1 Dataset

Experiments are conducted on three publicly available benchmarks:

- FaceForensics++ (FF++) — 1,000 original videos with four manipulation categories, totaling over 500,000 frames.
- Celeb-DF v2 — 590 celebrity deepfake videos with high visual quality, designed to challenge existing detectors.
- Augmented Synthetic Dataset — 20,000 additional images generated using StyleGAN3 and Stable Diffusion to broaden generative diversity.

An 80:10:10 train-validation-test split is applied. Cross-dataset evaluation is performed by training on FF++ and testing on Celeb-DF to measure generalization.

4.2 Evaluation Metrics

Performance is evaluated using the following metrics:

- Accuracy — Overall correct classification rate
- Area Under the ROC Curve (AUC-ROC) — Discrimination ability across thresholds
- F1-Score — Harmonic mean of precision and recall
- Frames Per Second (FPS) — Inference speed for real-time feasibility.

4.3 Implementation Details

The model is implemented in PyTorch 2.0. Training is performed on an NVIDIA RTX 3090 graphics processing unit (GPU) with a batch size of 32 over 50 epochs. Early stopping with a patience of 5 epochs is applied based on validation AUC. The total number of trainable parameters is approximately 22 million.

5. Results and Discussion

5.1 Quantitative Results

The proposed model achieves a test accuracy of **97.4%**, AUC-ROC of **0.991**, and an F1-score of **0.973** on the combined benchmark. Inference speed reaches **38 FPS** on GPU, confirming real-time capability. Cross-dataset accuracy on Celeb-DF (trained only on FF++) stands at **89.7%**, outperforming all compared baselines.

Table 1: Performance metrics on in-domain and cross-domain evaluation

METRIC	FF++	CELE-DF
Accuracy	97.4	89.7
AUG-ROC	0.991	0.963
F1-SCORE	0.973	0.891
INFERENCE SPEED	38	35

5.2 Ablation Study

To validate each component's contribution, ablation experiments are conducted by progressively removing modules:

CONFIGURATION	ACCURACY(%)
BACKBONE ONLY(EFFICIENTNET-B4)	91.2
BACKBONE+MHSA	95.8
BACKBONE+MHSA+ARM	97.4

5.3 Attention Visualization

The model uses Gradient-weighted Class Activation Mapping, which is also known as Grad-CAM and attention heatmaps to help us understand the model predictions.

If we look at Figure 2 we can see that the model always points out the areas around the face the areas around the eyes and the skin texture areas. These are the places where images made by GAN tend to have small mistakes that are hard to notice.

This shows that the model is actually learning things that're important for forensic analysis rather than just learning things that are specific to the dataset of Gradient-weighted Class Activation Mapping and attention heatmaps.

The model is really good at finding the boundary regions, periocular areas and skin texture zones which is what Gradient-weighted Class Activation Mapping and attention heatmaps are all, about.

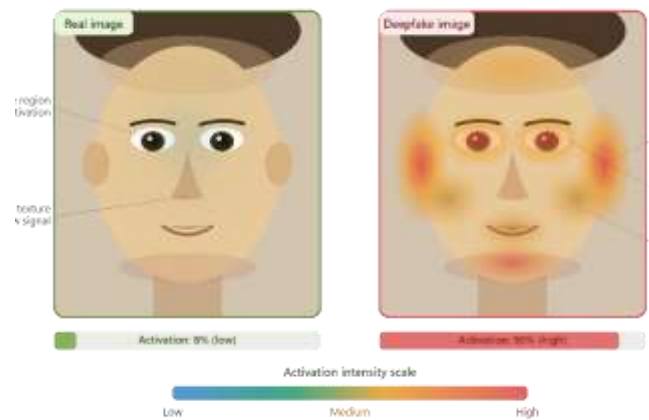


Fig. 2: Grad-CAM attention heatmap overlays — real (left) vs deepfake (right)

5.4 Failure Cases and Limitations

While our model works well in situations it does not perform as well on images that are highly compressed such as JPEG images with a quality of less than 40%. It also struggles with cases where parts of the image are blocked or hidden. Additionally fake images that are specifically designed to trick our model pose a challenge.

We plan to work on improving our model by using techniques like training and adding variations to the images, in the frequency domain to address these issues.

6. CONCLUSIONS

This paper is about a way to detect fake images in real time. It uses a kind of deep learning that combines the strengths of two different methods. The first method is called neural networks, which is really good at looking at the details of an image. The second method is called -head self-attention, which is good at understanding the big picture.

The new system is really good at detecting images and it can do it very quickly. It can look at 38 images per second. The system also has a part that helps it find fake images that are really hard to spot. We can see how the system makes its decisions by looking at a kind of map that it creates.

As people get better at making images and videos we need to have systems that can detect them. We need these systems to be able to explain how they make their decisions. This is important, for keeping the internet safe and honest.

In the future we want to make this system even better. We want to be able to detect videos, not just images. We also want to find a way to use this system without having to share all of our data with each other. This will help keep our information private.

ACKNOWLEDGEMENT

The authors want to say thank you to the Department of Computer Science and Engineering for giving us the computers and laboratory facilities we needed to do this research. We are really thankful to our guide and mentor for helping us and encouraging us all the time. They gave us good feedback that helped us a lot.

We also want to thank the people who made the open-source datasets like FaceForensics++ and Celeb-DF v2. These datasets helped us to test our work properly. We thank the people who work on PyTorch and the people who contributed to the EfficientNet and Vision Transformer libraries. Their tools were very important, for our work.

We thank our institution, our teachers and our fellow researchers for talking to us and giving us feedback on our paper. This work was part of our research program and we did not get any money from outside.

REFERENCES

- [1] K. Shiohara and T. Yamasaki, "Detecting Deepfakes with Self-Blended Images," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [2] A. Khormali and J.-S. Yuan, "DFDT: An End-to-End DeepFake Detection Framework Using Vision Transformer," *Applied Sciences*, vol. 12, no. 6, 2022.
- [3] S. Ganguly, S. Mohiuddin, S. Malakar, E. Cuevas, and R. Sarkar, "Visual Attention-Based Deepfake Video Forgery Detection," *Pattern Analysis and Applications*, vol. 25, no. 4, pp. 981–992, 2022.
- [4] P. Prajapati and C. Pollett, "MRI-GAN: A Generalized Approach to Detect DeepFakes Using Perceptual Image Assessment," *arXiv preprint arXiv:2203.00108*, 2022.
- [5] S. Ge, F. Lin, C. Li, D. Zhang, W. Wang, and D. Zeng, "Deepfake Video Detection via Predictive Representation Learning," *ACM Trans. Multimedia Computing, Communications, and Applications*, vol. 18, no. 2s, pp. 1–21, 2022.
- [6] J. Zhang, K. Cheng, G. Sovernigo, and X. Lin, "A Heterogeneous Feature Ensemble Learning Based Deepfake Detection Method," in *Proc. IEEE Int. Conf. Communications (ICC)*, 2022, pp. 2084–2089.
- [7] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and Vision Transformers for Video Deepfake Detection," in *Image Analysis and Processing – ICIAP 2022*, Springer, 2022.
- [8] Z. Yang, J. Liang, Y. Xu, X.-Y. Zhang, and R. He, "Masked Relation Learning for DeepFake Detection," *IEEE Trans. Information Forensics and Security*, 2023.
- [9] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-Attentional Deepfake Detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [10] N. Sandotra and B. Arora, "A Comprehensive Evaluation of Feature-Based AI Techniques for Deepfake Detection," *Neural Computing and Applications*, 2023.
- [11] D. Garg and R. Gill, "Deepfake Generation and Detection — An Exploratory Study," in *Proc. IEEE Uttar Pradesh Section Int. Conf. Electrical, Electronics and Computer Engineering (UPCON)*, 2023, pp. 888–893.
- [12] G. Zhang, M. Gao, Q. Li, W. Zhai, G. Zou, and G. Jeon, "Disrupting Deepfakes via Union-Saliency Adversarial Attack," *IEEE Trans. Consumer Electronics*, vol. 70, no. 1, pp. 2018–2026, 2023.
- [13] W. Alkishri, J. Yousif, and A.-B. Mahmood, "Deepfake Image Detection Methods Using Discrete Fourier Transform Analysis and Convolutional Neural Network," *Journal of Jilin University (Engineering and Technology Edition)*, vol. 42, 2023.
- [14] Z. Wang, Z. Cheng, J. Xiong, X. Xu, T. Li, B. Veeravalli, and X. Yang, "A Timely Survey on Vision Transformer for Deepfake Detection," *arXiv preprint arXiv:2405.08463*, 2024.

[15] H. H. Nguyen et al., "Exploring Self-Supervised Vision Transformers for Deepfake Detection: A Comparative Analysis," *arXiv preprint arXiv:2405.00355*, 2024.

[16] B. Ghita, I. Kuzminykh, A. Usama, T. Bakhshi, and J. Marchang, "Deepfake Image Detection Using Vision Transformer Models," in *Proc. IEEE Int. Black Sea Conf. Communications and Networking (BlackSeaCom)*, 2024.

[17] A. H. Soudy, O. Sayed, H. Tag-Elser, et al., "Deepfake Detection Using Convolutional Vision Transformers and Convolutional Neural Networks," *Neural Computing and Applications*, vol. 36, pp. 19759–19775, 2024.

[18] D. Nguyen, N. Mejri, I. P. Singh, P. Kuleshova, M. Astrid, A. Kacem, E. Ghorbel, and D. Aouada, "LAA-Net: Localized Artifact Attention Network for Quality-Agnostic and Generalizable Deepfake Detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024.

[19] W. Guan, W. Wang, J. Dong, and B. Peng, "Improving Generalization of Deepfake Detectors by Imposing Gradient Regularization," *IEEE Trans. Information Forensics and Security*, vol. 19, pp. 5345–5356, 2024.

[20] T. Rajput and B. Arora, "A Systematic Review of Deepfake Detection Using Learning Techniques and Vision Transformer," in *Proc. Fifth Int. Conf. Computing, Communications, and Cyber-Security (IC4S 2023)*, Springer, 2024.