

Robust Detection Of AI-Generated Fake Images Using Deep Learning

Ms. R. Femimol
Department of Computer Science
and Engineering,
Bharath Institute of Higher
Education and Research
femimol.cse@bharathuniv.ac.in

Thota Vishnu Vardhan
Department of Computer
Science and Engineering,
Bharath Institute of Higher
Education and Research
thotavishnu52@gmail.com

Thota Sri Jayadev Shashank
Department of Computer Science and
Engineering,
Bharath Institute of Higher Education
and Research
shashankthota01@gmail.com

Tirumani Lakshmi Ganesh
Department of Computer Science
and Engineering,
Bharath Institute of Higher Education
and Research
tirumaniganesh58@gmail.com

Thounaojam Lanchan Luwang
Department of Computer Science
and Engineering,
Bharath Institute of Higher Education and
Research
thounaojamlanchanluwang@gmail.com

Abstract- In Recent years have seen a tremendous advancement in artificial intelligence-based image creation technology, making it possible to create extremely realistic artificial pictures that are frequently unrecognizable from actual photographs. The risk of false information, digital manipulation, and misuse of visual media on internet platforms has increased due to the widespread availability of such tools. As a result, trustworthy techniques for differentiating authentic photographs from information produced by AI have grown in significance. Convolutional Neural Networks (CNN) and the EfficientNet architecture are used in this paper to present a deep learning-based method for identifying AI-generated photos. In order to detect subtle abnormalities and inconsistencies frequently found in artificial content, the suggested approach examines visual patterns and structural elements in images. The model is trained and evaluated using a dataset that includes both real-world photos and images produced by artificial intelligence. To enhance model performance, image preparation methods including scaling and normalization are used. In comparison to the unique CNN model, experimental evaluation shows that the EfficientNet model achieves better generalization and higher classification accuracy. The suggested method can help stop the growth of misleading artificial media and offers an automated and scalable way to confirm the authenticity of digital images.

Keywords- Conventional Neural Network, EfficientNet,, Deep Learning, Artificial Intelligence

creation. Current generative models, including Generative Adversarial Networks (GANs) and diffusion-based models, can be used to create very realistic images that are close to real photographs. Although these technologies have tremendous advantages in fields like digital content producing, entertainment, and virtual worlds, they also come with new aspects of misinformation and manipulation of digital media.

With the proliferation of AI image generation technologies, it has become more and more challenging to discern between genuine images and those generated through AI. Artificial intelligence (AI) has the potential to be abused to produce bogus visual content, disseminate fake information, or distort perceptions of people on the Internet. This has made the issue of authenticity of digital images an issue of concern in digital forensics and media Verification

In this project, a robust deep learning-based system is proposed for detecting AI-generated fake images. The system utilizes CNN and EfficientNet models to classify images as real or fake with high accuracy. Image preprocessing and data augmentation techniques are applied to improve model performance and generalization. The proposed system aims to enhance digital image authenticity and provide an effective solution for applications such as social media monitoring, digital forensics, cybersecurity, and content verification.

Conventional approaches to verifying image authenticity are usually based on manual verification, or other basic forensic approaches, including metadata analysis and artifact discovery. These methods are however not effective with high-level AI generated images. Convolutional Neural

I. Introduction

The artificially intelligent technology has been developing very fast over the past few years, especially in image

Networks (CNNs) and other deep learning methods have demonstrated high performance in image analysis tasks because they are able to automatically learn complex visual features.

This paper suggests a deep learning-based architecture to identify AI-generated images based on CNN and EfficientNet networks.

II. LITERATURE SURVEY

Due to the fast evolution of artificial intelligence, it has generated sophisticated generative models that can create very realistic synthetic pictures. The more such images can

After the collection of the dataset, the images go through preprocessing phase. Another significant procedure in the data preparation in the deep learning model training is image preprocessing. At this phase, we reduce images to a consistent resolution to make them consistent in the dataset. The values of pixels are scaled to allow the model to work effectively with the data. Moreover, the data augmentation methods like rotation, flipping, and scaling can also be used to enhance the variety of the dataset and train the model more resiliently.

Once the data has been preprocessed, a deep learning model is trained with the prepared data. Convolutional Neural Network (CNN) is employed in this study because it is effective in classifying images. The CNNs can automatically detect significant features of images with multiple convolution and pooling operators. These layers assist the model to identify complex features like textures, shapes, edges, and structural anomalies found in synthetic images.

In the training process, CNN model learns to recognize the distinguishing features between real and AI generated images. The model modifies its internal parameters in an optimization procedure to reduce errors in classification. Through this training process, this system is able to create a trusted classification model that is able to identify subtle differences between authentic and synthetic images.

Evaluation of the model is done after successful training with a separate test dataset of images that had not been used in training the model. This assessment assists in quantifying the model performance with regards to classification accuracy and reliability. The trained system then predicts a given input image to be real or AI-generated using the features that it has learned.

The suggested system offers an automated method of identifying AI-generated images that can help enhance the process of reviewing digital content and avoid the dissemination of distorted visual data.

The deep learning model used in this study is based on a Convolutional Neural Network (CNN) architecture. CNNs have found extensive application in image classification applications because they can automatically extract hierarchical features of images. The model comprises of

several convolutional layers, pooling layers and fully connected layers that collaborate to examine image patterns and classify them.

In the training, the model becomes trained to detect minor visual discrepancies between real images and the AI-generated ones. Classification accuracy and prediction reliability are standard evaluation measures that were used in assessing the performance of the model. These metrics give information on the effectiveness with which the model can differentiate between genuine and fake images.

The experimental model was proposed in such a way that it would train and test the model in a controlled systematic way. By using a well-prepared dataset and an effective deep learning architecture, the proposed system demonstrates promising results in detecting AI-generated images.

diffusionbased models, can be used to create very realistic images that are close to real photographs. Although these technologies have tremendous advantages in fields like digital content producing, entertainment, and virtual worlds, they also come with new aspects of misinformation and manipulation of digital media. With the proliferation of AI image generation technologies, it has become more and more challenging to discern between genuine images and those generated through AI. Artificial intelligence (AI) has the potential to be abused to produce bogus visual content, disseminate fake information, or distort perceptions of people on the Internet. This has made the issue of authenticity of digital images an issue of concern in digital forensics and media verification. Conventional approaches to verifying image authenticity are usually based on manual verification, or other basic forensic approaches, including metadata analysis and artifact discovery. These methods are however not effective with high-level AI generated images. Convolutional Neural Networks (CNNs) and other deep learning methods have demonstrated high performance in image analysis tasks because they are able to automatically learn complex visual features. With the fast development of artificial intelligence, it has become possible to make highly realistic synthetic images using current generative models. Although these technologies offer advantages in fields like digital design, media production and virtual simulation, they also bring challenging issues to the levels of authenticity and reliability of visual information. AI-generated images are easily misused to commit misinformation, to commit digital forgery and media manipulation. Therefore, detecting such images has become an important research problem in the field of computer vision and digital forensics. The main aim of the study is to come up with an effective deep learning-based

With the fast development of artificial intelligence, it has become possible to make highly realistic synthetic images using current generative models. Although these technologies offer advantages in fields like digital design, media production and virtual simulation, they also bring challenging issues to the levels of authenticity and reliability of visual information. AI-generated images are easily

be confused with actual photographs, the more researchers have been keen on designing automated detection tools that can be used to guarantee authenticity of digital media.

Wang et al. [1] suggested that AI-generated images can be detected by considering deep learning networks that examine visual artifacts that are fine (synthetic content). Their work covered the use of transformer-based models in order to obtain both local and global image features. The study demonstrated that deep learning methods can effectively detect synthetic images; however, the performance may vary depending on the type of generative model used to create the images.

Li et al. [2] proposed a large-scale dataset that is aimed specifically at identifying AI-generated images. The data consists of actual photos and fake images created with the help of various generative models. Their study emphasized the value of training and detecting models using a variety of datasets to enhance generalization and strength in the face of novel AI systems.

A framework called SCADET was created by Kim et al. [3] to identify AI-generated art. The system proposed employed deep learning to detect anomalies in image texture, color distribution, and structuring patterns. It was demonstrated that deep learning models can successfully find subtle artifacts in synthetic images, but the system needs large volumes of training data to operate at high levels.

Chen et al. [4] performed a systematic review of different approaches towards the detection of AI-generated images. Their article examined various detection methods comprising CNN-based models, transformer-based architectures, and statistical methods of analysis. The paper has found that the deep learning methods are more accurate than the conventional digital forensic methods.

Patel et al. [5] explored the issues related to identifying high-quality AI-generated images. They demonstrated that the better generative models become the harder it is to identify synthetic images. The paper has highlighted the importance of more powerful detection mechanisms with the ability to detect subtle visual artifacts.

Despite the number of detection methods suggested, a lot of currently existing systems still have certain limitations in handling images produced by new AI models or when images are highly compressed or modified. Therefore, there is a need for more efficient deep learning-based detection frameworks. Here, a detection system with Convolutional Neural Networks (CNN) and EfficientNet structures is suggested to enhance the precision of the real and AI-generated images.

III. PROBLEM STATEMENT

The development of generative AI technologies has facilitated the production of very realistic synthetic images that cannot be easily differentiated with real images. This creates issues about misinformation, online forgery, and the reliance on visual media. Current detection systems tend to be not very accurate across various AI generation models.

Therefore, an efficient deep learning-based system is required to reliably detect and classify AI-generated and real images.

IV. OBJECTIVE OF THE STUDY

With the fast development of artificial intelligence, it has become possible to make highly realistic synthetic images using current generative models. Although these technologies offer advantages in fields like digital design, media production and virtual simulation, they also bring challenging issues to the levels of authenticity and reliability of visual information. AI-generated images are easily misused to commit misinformation, to commit digital forgery and media manipulation. Therefore, detecting such images has become an important research problem in the field of computer vision and digital forensics.

The main aim of the study is to come up with an effective deep learning-based model that can help in differentiating between real images and AI-generated images with high accuracy. Through the power of the modern neural networks, the system seeks to automatically extract the patterns and characteristics that exist in the digital images with the aim of establishing their authenticity.

The other goal of this work is to examine the visual features and concealed artifacts that people tend to observe in artificial images produced with the help of artificial intelligence models. Such artifacts can contain anomalies in textures, extreme pixel layouts, unnatural edges, or structural anomalies, which are sometimes hard to perceive by humans but can be detected using machine learning methods.

The study will also enhance the credibility and quality of identifying the manipulated or AI-generated visual material. The system can learn complex patterns that distinguish between real and synthetic images by training the model on a wide range of data of both real and generated images.

Moreover, the research is aimed at creating an automated detection tool that will help to check the authenticity of digital images. This type of system may be applicable in many areas such as media verification, digital journalism, cybersecurity, and monitoring social media, where credibility of visual information needs to be ensured.

V. PROPOSED SYSTEM

The suggested system proposes the implementation of a deep learning-based architecture that attempts to identify AI-generated images based on the visual features of the images. The primary objective of the system is to categorize digital images into two groups, namely, real and AI-generated images. The system involves the use of superior image processing and machine learning algorithms to detect underwater patterns and artifacts that differentiate between a synthetic and an authentic image.

The initial step in the proposed system would be the acquisition of a set of images. The data is comprised of two sets of images: natural images of real sources and AI-generated images created by generative models.

After the collection of the dataset, the images go through preprocessing phase. Another significant procedure in the data preparation in the deep learning model training is image preprocessing. At this phase, we reduce images to a consistent resolution to make them consistent in the dataset. The values of pixels are scaled to allow the model to work effectively with the data. Moreover, the data augmentation methods like rotation, flipping, and scaling can also be used to enhance the variety of the dataset and train the model more resiliently.

Once the data has been preprocessed, a deep learning model is trained with the prepared data. Convolutional Neural Network (CNN) is employed in this study because it is effective in classifying images. The CNNs can automatically detect significant features of images with multiple convolution and pooling operators. These layers assist the model to identify complex features like textures, shapes, edges, and structural anomalies found in synthetic images.

In the training process, CNN model learns to recognize the distinguishing features between real and AI generated images. The model modifies its internal parameters in an optimization procedure to reduce errors in classification. Through this training process, this system is able to create a trusted classification model that is able to identify subtle differences between authentic and synthetic images.

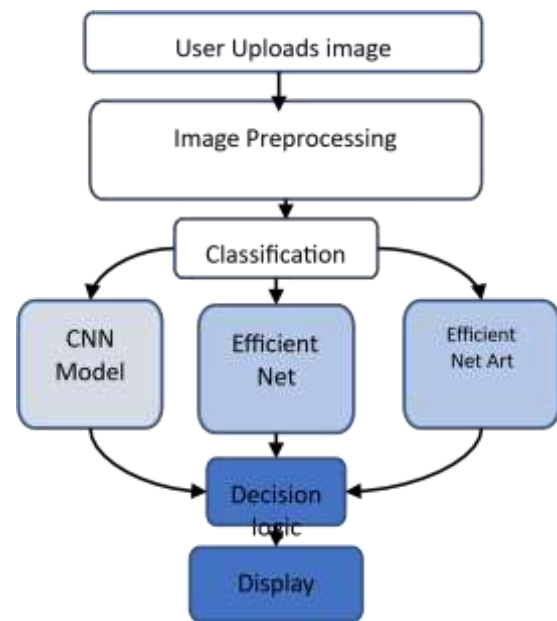
Evaluation of the model is done after successful training with a separate test dataset of images that had not been used in training the model. This assessment assists in quantifying the model performance with regards to classification accuracy and reliability. The trained system then predicts a given input image to be real or AI-generated using the features that it has learned.

The suggested system offers an automated method of identifying AI-generated images that can help enhance the process of reviewing digital content and avoid the dissemination of distorted visual data.

The suggested system proposes the implementation of a deep learning-based architecture that attempts to identify AI-generated images based on the visual features of the images. The primary objective of the system is to categorize digital images into two groups, namely, real and AI-generated images. The system involves the use of superior image processing and machine learning algorithms to detect underwater patterns and artifacts that differentiate between a synthetic and an authentic image. Once the data has been preprocessed, a deep learning model is trained with the prepared data. Convolutional Neural Network (CNN) is employed in this study because it is effective in classifying images. The CNNs can automatically detect significant features of images with multiple convolution and pooling operators. These layers assist the model to identify complex features like textures, shapes, edges, and structural anomalies found in synthetic images. In the training process, CNN model learns to recognize the distinguishing features

between real and AI generated images. The model modifies its internal parameters in an optimization procedure to reduce errors in classification. Through this training process, this system is able to create a trusted classification model that is able to identify subtle differences between authentic and synthetic images.

VI.



SYSTEM ARCHITECTURE

Fig 1: System Architecture

The proposed system will automatically identify the presence or absence of AI-generated image or an actual photograph based on deep-learning models. The architecture is made up of several steps such as image input, preprocessing, deep neural networks classification, decision logic and display of results. The overall workflow of the system is illustrated in the system architecture diagram.

1. User Image Upload

The interface starts with the user posting an image onto the system. The image uploaded could be a real world photograph or an image generated by AI. The main data on which the detection process will take place is this input image.

2. Image Preprocessing

The input image is first processed in a preprocessing stage before classification to enhance the performance of the model. This step comprises of operations that include: Scaling of images to a fixed resolution as needed by the models.

- Normalization of pixels.
- De-noising and de-featureing.

Preprocessing so that the image is converted into a format that can be used in deep learning models and enhances the accuracy of the classification.

3. Image Classification

The image is then preprocessed and then subjected to a series of deep learning models to extract features and classify the images. The system makes use of three models: CNN Model:

The image is extracted to a Convolutional Neural Network (CNN) to extract the spatial features of the image. CNNs are effective in detecting patterns, textures, and structural inconsistencies that may exist in AI-generated images. EfficientNet Model:

EfficientNet is a highly optimized deep learning model that trades model depth, model width and model resolution. It effectively boosts the extraction of features and better classification.

EfficientNet-Art Model:

This model specifically targets identifying artistic or AI-generated patterns that are often present in synthetic images. It assists in increasing the accuracy of detection of images produced by the current AI tools.

4. Decision Logic

The CNN, EfficientNet and EfficientNet-Art models are used to provide the outputs, which are combined through a decision logic mechanism. This step combines the forecasts of all the models and arrives at the final classification decision. Decision logic based on ensembles enhances reliability and risk of misclassification.

5. Result Display

Lastly, the system shows the user the prediction result. The output shows whether uploaded image is categorized as:

- Real Image
- AI Generated Image

This output is displayed to the system interface to be interpreted by the user.

VII. METHODOLOGY

1. Dataset Collection

The initial process of the proposed system is the gathering of a dataset of actual images and images created by AI. Real images are obtained by using publicly available image datasets and AI-generated images are generated by using different generative models including GAN based and diffusion based image generators. The dataset will be divided into two categories: Real and AI-Generated images. This dataset is used to train and evaluate the deep learning models.

2. Data Preprocessing

The images obtained are then subjected to various preprocessing stages before training the model to ensure they become suitable to the deep learning algorithms. The preprocessing procedures are:

- Resizing of the image to a constant size (e.g., 224x224 pixels)
- Normalization of pixel values.

Such preprocessing methods can enhance the performance of the model and avoid overfitting in the training.

3. Feature Extraction Using CNN

Convolutional Neural Networks (CNN) are used to extract features. The CNN models can automatically extract meaningful image features like edges, textures, shapes, and patterns. These are the characteristics that make it easier to identify real and AI-generated images.

In this section, the model is developed with the help of EfficientNet. EfficientNet is another pre-trained deep learning architecture in addition to the custom CNN model. EfficientNet is more effective and uses fewer parameters, easily scaling network depth, width and resolution.

The EfficientNet model is pretrained on the gathered data then fine-tuned to enhance the accuracy of the detection of the images generated by AI.

5. Model Training

In training the models are trained to classify images according to their characteristics. The backpropagation and optimization of the Adam optimizer are used in the training process to update the model weights and reduce errors in the classification.

6. Image Classification

The system can classify new images of input after some training. Upon inputting an image to the system, the image goes through the trained model where features are extracted and analyzed. The last classification layer classifies the input image as:

- Real Image
- AI-Generated Image

7. System Output

The last system takes an input image, feeds the image into the trained deep learning model, and the result of the prediction is whether the image is real or AI-generated.

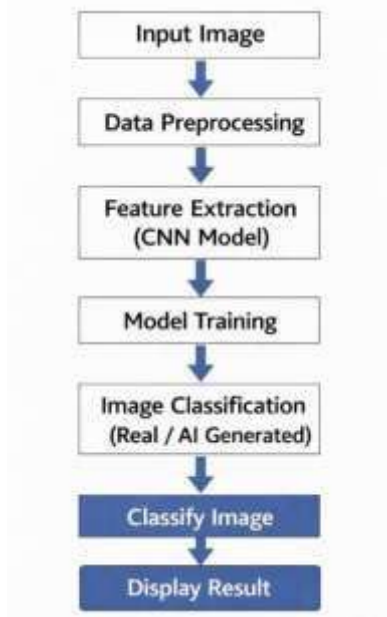


Fig 2: Flowchart

VIII. DATASET DESCRIPTION AND EXPERIMENTAL SETUP

The quality and diversity of the dataset used to train and test a deep learning model greatly determine the performance of the model. A dataset of real and AI-generated images was utilized in this study to train and test the suggested system.

The data includes two major groups, namely, realworld pictures and artificial pictures produced by artificial intelligence-based generative models. Real images were gathered using publicly available image datasets and the internet, whereas AI-generated images were taken using image generation models, which generate synthetic visuals. The fact that both of the categories are included in the model makes it possible to learn distinguishing features of real and artificial generated images.

A series of preprocessing steps were performed on the dataset before training the model, to bring a measure of consistency and enhance the performance of the model. All the images were downsampled to a fixed resolution to allow them to be effectively processed by the neural network. The pixel values were also normalized to fall within a regularized scale that facilitates quicker convergence of the model in the training process. Moreover, data augmentation methods were used, which included rotation, horizontal flipping, and scaling to diversify the dataset and minimize the chances of overfitting.

The data was then separated into three parts: training data, validation data and testing data. The deep learning model was trained using the training dataset and enabled it to learn patterns and features that occur in the images. Validation set was to monitor the model performance during training process and to modify hyperparameters in case of need. Lastly, the

performance of the trained model on unseen data was compared to the testing dataset.

The deep learning model used in this study is based on a Convolutional Neural Network (CNN) architecture. CNNs have found extensive application in image classification applications because they can automatically extract hierarchical features of images. The model comprises of several convolutional layers, pooling layers and fully connected layers that collaborate to examine image patterns and classify them.

In the training, the model becomes trained to detect minor visual discrepancies between real images and the AI-generated ones. Classification accuracy and prediction reliability are standard evaluation measures that were used in assessing the performance of the model. These metrics give information on the effectiveness with which the model can differentiate between genuine and fake images.

The experimental model was proposed in such a way that it would train and test the model in a controlled systematic way. By using a well-prepared dataset and an effective deep learning architecture, the proposed system demonstrates promising results in detecting AI-generated images.

IX. RESULT AND DISCUSSION

The dataset used to train the proposed deep learning model was comprised of real images and AI-generated images. Throughout the training, the model was able to learn how to discover subtle patterns and visual features that help differentiate between synthetic images and real ones. The CNN model was in a position to examine texture patterns, pixel distributions as well as structural inconsistencies that existed in the images.

Experimental findings show that the trained model is able to categorize images without errors. The model showed a high degree of performance in detecting the artifacts that are typically found in AI-generated pictures. Such artifacts can be unnatural textures, bad pixel layouts, and structural anomalies, which are hard to see by the human eye.

Testing dataset which did not take part in training phase was used to test the performance of the system. The results of the evaluation indicate that the model is able to recognize both the real and synthetic images with high degree of reliability. This shows that deep learning methods are effective in addressing image authenticity detection issues.

Yet, it should be mentioned that AI image generation technologies are rapidly developing. The new generative models can produce more realistic images and, therefore, could decrease the efficiency of current detection techniques. Thus, the development of detection methods should be improved to remain on a par with the progress of generative AI.

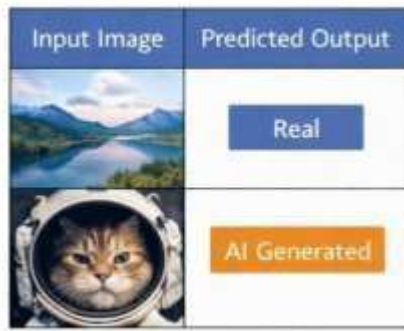


Fig 3: Image Classification

The findings of this study demonstrate the promise of deep learning-based solutions to detecting manipulated visual content and assist digital media verification.

X. CONCLUSION AND FUTURESCOPE

The proposed approach to the detection of AI-generated images based on deep learning has been introduced and discussed in this study. The system makes use of the image preprocessing techniques and a convolutional neural network model to categorize the pictures into real and synthetic. Through visual patterns and structural features analysis of images, the suggested system will be capable of detecting even minor variations between original photographs and AI-generated images.

The results of the experiment support the claim that deep learning methods can be incredibly effective in identifying AI-generated images. The trained CNN model could predict high classification accuracy because it detected the artifacts and inconsistencies that were in the synthetic images that were hidden.

The proposed system offers a viable idea towards enhancing the authenticity and reliability of digital visual content. This type of systems can be significantly involved in the avoidance of misinformation, manipulation of digital data, and the reliability of online media.

To advance the work in the future, the system may be enhanced by training the model to work with larger and more diverse datasets. Also, more complicated deep learning designs and a composite detection method can be considered to improve the model to detect more lifelike images produced by AI. The trained CNN model could predict high classification accuracy because it detected the artifacts and inconsistencies that were in the synthetic images that were hidden.

XI. REFERENCES

[1] Q. Wang, L. Zhang, X. Li, J. Chen, and Y. Liu, "Advanced Detection of AI-Generated Images Through Vision Transformers," in Proc. IEEE Int. Conf. Computer Vision, pp. 1234-1245, 2024.

[2] Y. Li, Z. Wang, H. Liu, M. Zhang, and K. Chen, "A Comprehensive Dataset for Human vs. AI Generated Image Detection," arXiv preprint arXiv:2601.00553, 2026.

[3] J. Kim, S. Park, K. Lee, H. Choi, and D. Kim, "SCADET: A detection framework for AI-generated artwork," Int. J. Multimedia Tools and Applications, vol. 84, no. 12, pp. 789-802, 2025.

[4] X. Chen, Y. Liu, W. Zhao, and P. Yang, "Methods and trends in detecting AI-generated images," Digital Investigation, vol. 45, pp. 301-315, 2026.

[5] R. Patel, A. Gupta, S. Singh, N. Kumar, and V. Rao, "Are High-Quality AI Images More Difficult for Detectors?," in Proc. Int. Conf. Machine Learning (ICML), pp. 45674578, 2025.

[6] J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI Generated Synthetic Images," arXiv preprint arXiv:2303.14126, 2023.

[7] M. Keita, W. Hamidouche, H. Bougueffa Eutamene, A. Taleb-Ahmed, and A. Hadid, "DeeCLIP: A Robust and Generalizable Transformer-Based Framework for Detecting AI-Generated Images," arXiv preprint arXiv:2504.19876, 2025.

[8] F. Ding, J. Zhang, X. He, and J. Xu, "FairAdapter: Detecting AI-Generated Images With Improved Fairness," arXiv preprint arXiv:2411.14755, 2024.