# Advanced Detection of AI-Generated Images Through Vision Transformers

Darshan Lamichhane[1]

[1]Everest English Boarding Secondary School

September 05, 2024

# Advanced Detection of AI-Generated Images Through Vision Transformers

Darshan Lamichhane
*Everest English Boarding Secondary School, Butwal-8, Nepal*
darshanlamichhane24@gmail.com

***Abstract*—The rapid advancement of Artificial Intelligence (AI) models such as Generative Adversarial Networks (GANs) has been a great success in the field of image synthesis and creation. Artificially generated GAN-based images are widely spread over the Internet along with the development in generation of natural and photorealistic images. While this could lead to better digital media and content, it also poses a risk to security, legitimacy, and authenticity.**

**The advancement of AI-generated images, particularly those that are produced by Generative Adversarial Networks (GANs), has created a rising concern about the potential misuse of these images in spreading misinformation and creating deepfakes. Detecting such fake or AI-generated images has become an important challenge in maintaining the integrity of digital media. In this research, we have explored the application of the Vision Transformer (ViTs) model for detecting AI-generated images, leveraging the Kaggle dataset - a balanced collection of real and AI-generated images. The Vision Transformer is recognized for its innovative method of treating images as sequences of patches and excels at identifying long-range dependencies and complex patterns within images. That makes it exceptionally well-suited for this task of detecting fake images. We have fine-tuned the ViT model on the dataset, performing data augmentation techniques on it and leveraging pretrained weights to boost the model's performance. The findings thus obtained demonstrate that the ViT model attains a high level of accuracy in differentiating between real and AI-generated images, outperforming traditional CNN-based approaches. Beyond performance evaluation, we also conducted an ablation study to examine the impact of various components of the ViT model, including the number of attention heads, patch size, the impact of data augmentation, and the depth of layers. The results obtained in this study indicate that the ViT model not only excels in accuracy but also provides a robust framework for detecting AI-generated images across diverse scenarios. Our study shows the strength of transformer based models in addressing the increasing challenge of AI-generated image detection, laying a foundation for future research in this critical area. This experiment highlights that when the ViT model is fine tuned with optimal data augmentation techniques, it gains state of the art performance in AI-generated image detection, emphasizing its potential for real-world applications.**

***Index Terms*—GAN based images detection; DeepFake Images; GAN image classification; detection of AI-generated images; fake AI-generated images detection; vision transformers; CNN**

## I. Introduction

The evolution of artificial intelligence (AI) has led to remarkable progress in generative models, especially those models that are based on Generative Adversarial Networks (GANs). These models are becoming significantly proficient at generating highly realistic images that are almost unnoticeable from the real images and making it nearly impossible to tell them apart from the actual photos. While this technological progress has unlocked a world of new possibilities in fields like design, art, and entertainment, it has also raised some serious concerns around misusing these deepfake images and AI-generated content. Specifically, the ability to produce photorealistic synthetic images creates a threat to the integrity of digital media, potentially raising the spread of misinformation and deepfakes.

Therefore, this challenge of distinguishing AI-generated images and the real images has become a critical area of research. Traditional image analysis methods, primarily relying on Convolutional Neural Networks (CNN), have been employed to answer this problem. However, as generative models continue to evolve and grow more sophisticated, these conventional methods are increasingly insufficient. This inadequacy has led to the exploration of more advanced architectures, such as the Vision Transformer (ViT), which has shown encouraging outcomes in many computer vision applications.

Vision Transformers marks a paradigm shift in image processing by treating images as sequences of patches, much like tokens in natural language processing. This is a unique approach which allows the model to identify the long-range dependencies and complex patterns in an image, making it particularly effective in tasks like detecting AI-generated images were fine-grained details are very crucial.

This paper explores the application of the Vision Transformer (ViT) model to the challenge of detecting AI-generated images. The dataset employed for this study consists of an equal number of real images and the AI-generated images. This offers a solid benchmark for evaluating the performance of detection models. Our study focuses on fine-tuning the ViT model for this particular task, and leveraging its ability to learn rich feature representations from image patches.

In this paper, we have shown a complete analysis of the model performance and conducted its ablation study to see the impact of different components of the ViT model. Additionally, we have also highlighted the advantages and potential limitations of using transformer models for the detection of AI-generated images by comparing the results of the ViT model with those of the traditional CNN-based models. Our study contributes to the expanding research aimed at protecting the digital media against the spread of fake content. Our study contributes valuable insights into the potential of cutting-edge AI models in this area.

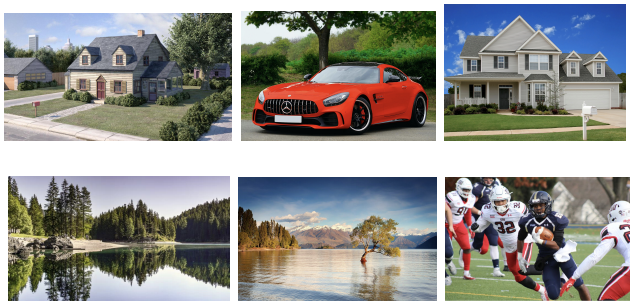

Fig. 1. Real Images

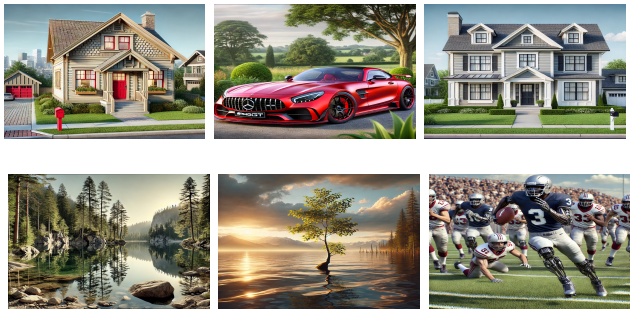

Fig. 2. AI Generated Images

### A. Background

The aim of the digital image forensics is to establish the authenticity of digital images by detecting if the image is real or synthetic/manipulated as shown in Fig. 1 vs Fig. 2. There is rapid development in the tools for image editing and generative models, so the traditional tools and methods for image detection have been frequently struggling to address this issue. However, with the trending advancements in the machine learning, especially in the transformer models like Vision Transformers (ViTs), they have outperformed and have provided exceptional opportunities to enhance image forensic techniques.

### B. Objective

The objective of this study is to investigate the application of Vision Transformers (ViTs) in the field of digital image forensics, and especially to investigate the ability of the model for image detection and also to compare their performance with traditional methods.

## II. Related Work

The detection of AI-generated images has emerged as an important research area in correspondence to the rapid advancements in generative models, particularly those which are based on Generative Adversarial Networks (GANs). These models are capable of producing highly realistic images and have led to significant challenges in differentiating between synthetic and real images. As a result, various methods have been developed to address this issue, ranging from traditional image analysis techniques to more advanced deep learning approaches.

### A. Traditional Image Analysis Techniques

Early methods for detecting AI-generated images relied primarily on handcrafted features and statistical analysis of images. Techniques such as examining image compression artifacts, inconsistencies in lighting, or anomalies in pixel-level patterns were among the first approaches used to identify synthetic content. For instance, McCloskey and Albright (2018) [5] explored the detection of GAN-generated images by analyzing color cues and artifacts that are often introduced during the image generation process. However, these methods often struggled with high-quality AI-generated images and lacked the robustness required for diverse real-world applications.

### B. Convolutional Neural Networks (CNNs)

With the advancement in the field of deep learning and machine learning, Convolutional Neural Networks (CNNs) have become the most common technique for the tasks of image classification, which also include the detection of AI-generated images. CNN-based models automatically learn hierarchical feature representations from the input images, which has shown to be very effective in many computer vision tasks. Several studies have applied CNNs to detect GAN-generated images by training models on large datasets of real and fake images. For example, Zhang et al. (2019) [6] utilized CNNs to detect deepfake images, demonstrating significant improvements over traditional methods.

However, as generative models became more sophisticated, at producing images that closely mimic real photographs, the limitations of CNNs began to surface. CNNs often struggle with identifying global context and long-range dependencies in images, which are crucial for identifying subtle differences between real and AI-generated images. This limitation has led researchers to explore more advanced architectures, such as transformers, which offer a different approach to image analysis.

### C. Vision Transformers (ViTs)

The introduction of Vision Transformers (ViTs) by Dosovitskiy et al. (2020) [7] marked a significant advancement in the field of computer vision. Derived on the architectural framework initially developed for natural language processing (NLP), which treats text as sequences of tokens, Vision Transformers adapts the same transformer architecture by considering images as sequences of patches for image classification tasks. This approach enables the ViT model to identify long-range dependencies and complex relationships between different parts of an image, which make them particularly suitable for tasks that require fine-grained analysis, such as AI-generated image detection.

Various studies have explored the use of ViTs in several computer vision tasks. These studies have demonstrated the exceptional performance of ViTs compared to traditional CNNs. The study of Touvron et al. (2021) [8] introduced Data-efficient Image Transformers (DeiTs), which showed that the transformers could get high accuracy with less data and reduced computational resources. In the terms of AI-generated

image detection, ViTs offer a promising alternative to CNNs, leveraging their ability to learn rich feature representations from image patches.

### *D. Detection of AI-Generated Images Using Transformers*

With the recent development of using transformers in computer vision, there is an increasing interest in using the transformers to detect AI-generated images. There are some studies performed around using transformers for computer vision. The study by Wang et al. (2021) [9] explored the use of transformers for deepfake detection. This work shows that these transformer models could outperform the traditional CNNs in identifying the deepfake images. However, exploring the use of Vision transformers for AI-generated image detection is still in its early stage, and there are high levels of outcomes in this research area that can be unfolded.

This current study contributes to the growing body of research by leveraging the Vision Transformer model (ViT) for detecting AI-generated images. By leveraging the unique capabilities of ViTs to capture complex patterns and dependencies in images, the current approach provides a powerful framework for differentiating between real and fake content.

Inspired by the high accomplishments of transformers in the field of natural language processing (NLP), Vision Transformers (ViTs) introduce a novel approach in computer vision. ViTs treat images as sequences of patches and apply self-attention mechanisms to identify long-range dependencies. This makes the ViT very effective for complex tasks like image detection and classification.

While the previous studies have shown the potential of ViTs in various image classification tasks, but their use for AI-generated image detection remains underexplored. The aim of our study is to tackle this use case by conducting a complete evaluation of the ViT model for image detection.

## III. Data Collection and Methodology

The detection of AI-generated images has emerged as an important research area in correspondence to the rapid advancements in generative models, particularly those which are based on Generative Adversarial Networks (GANs). These models are capable of producing highly realistic images and have led us to significant challenges in differentiating between synthetic and real images. As a result, various methods have been developed to address this issue, ranging from traditional image analysis techniques to more advanced deep learning approaches.

### *A. Data Collection*

The dataset used in our study is sourced from the Kaggle on digital image forensics. It includes a diverse set of images with various manipulation types such as copy-move, splicing, and inpainting.

The dataset consists of a balanced set of real and AI-generated images. The dataset includes 30,000 images, with 15,000 real images sourced from various online repositories and 15,000 AI-generated images created using advanced GAN architectures. The images in the dataset are labeled as either ”real” or ”fake,” providing a suitable benchmark for evaluating detection models. The images in the dataset cover a diverse range of subjects and styles, providing a robust foundation for training a model to distinguish between real content and AI-generated content.

### *B. Data Preprocessing*

To prepare the images for input into the proposed model of Vision Transformer, a series of preprocessing steps were performed on the dataset:

- **Resizing**: All the images were first resized to 224x224 pixels, which is common input size for the Vision Transformer model, ensuring uniformity across the dataset.
- **Normalization**: The pixel of the input image were then normalized to [0, 1] to standardize the input data, improving the model's convergence during training.
- **Data Augmentation**: Various data augmentation methods used here are aimed to diversify the training data and maximise the model's capacity to generalise. These techniques included random cropping, horizontal flipping, color jittering, and rotation. Data augmentation helps the model learn more robust features by exposing it to a wider range of possible image transformations.

### *C. Model Architecture*

The study employs Vision Transformers (ViTs) for image forensic tasks. ViTs are trained to classify images as either authentic or manipulated.

The Vision Transformer (ViT) model was employed as the backbone for detecting AI-generated images. The ViT model, originally proposed by Dosovitskiy et al. [1], introduces a unique method for image classification, which considers images as sequences of patches, same as tokens in natural language processing. The model architecture includes:

- **Patch Embedding**: In patch embedding, the input image is initially partitioned into non-overlapping patches, typically measuring 16x16 pixels. Subsequently, each patch is compressed into a one-dimensional vector and embedded linearly into a vector of a predetermined size. The proposed procedure involves the conversion of the image into a series of embedded patches. Ultimately, these will function as the input tokens for the transformer.
- **Positional Encoding**: This process guarantees the preservation of the spatial details of the patches by incorporating positional encodings into the patch embeddings. All of these encodings assist the model in comprehending the relative positions of regions within the image. This step is very crucial for the training as it captures the overall structure and context.
- **Transformer Encoder**: As shown in the Fig. 3, the sequence of patch embeddings along with the added positional encodings in the previous step, is finally fed into a transformer encoder. The encoder comprises multiple layers of multi head self-attention and feed-forward

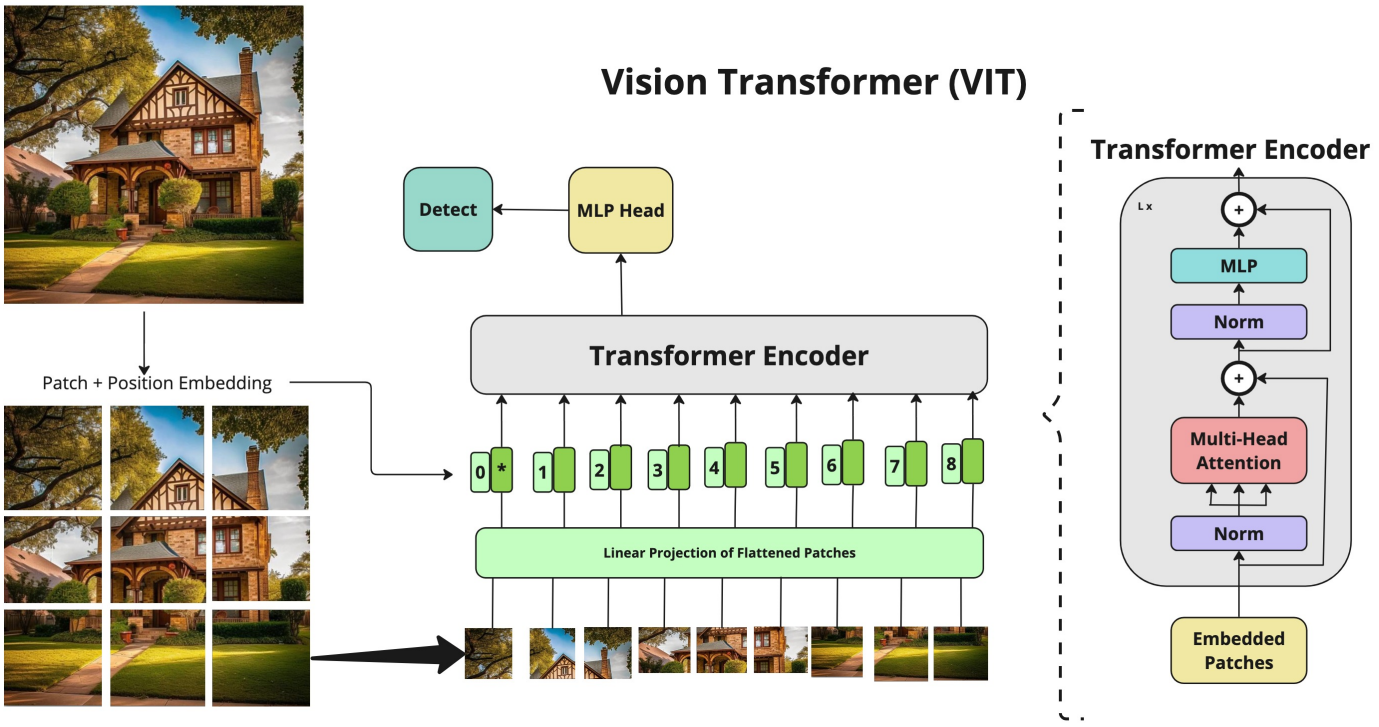


Fig. 3. Vision Transformer Architecture

neural networks, enabling the model to effectively exploit long-range dependencies and interactions among various components of the image.

- **Classification Head**: Finally, the sequence of patch embeddings we prepared in the last step is then prepended with a special classification token ([CLS]) as shown in the figure a. After passing through the transformer encoder, the output we get from this token is used as the representation for the entire image. After this, the representation is passed through a fully connected layer along with a softmax activation function to produce the final classification output (real or fake).

The model architecture is shown in Fig. 3. The Vision Transformer (ViT) model is utilized in this study because of its exceptional effectiveness in image classification tasks. The ViT model processes images as sequences of fixed-size patches, and then each of these patches is linearly embedded and fed into a transformer encoder. The model uses self-attention mechanisms to identify patterns and features across the entire image. This makes the model particularly robust for detecting subtle differences between real and fake images.

### *D. Training and Evaluation*

The ViTs model is trained using a subset of the dataset, with hyperparameter tuning and cross-validation performed to optimize performance. The evaluation metrics we had considered here includes precision, accuracy, recall, and F1-score.

We fine-tuned the ViT model on the dataset using a combination of data augmentation methods to increase the generalization capabilities of the model. The data augmentation pipeline included horizontal flipping, random cropping and color jittering. The model was first setup with pretrained weights from the ImageNet dataset to leverage existing visual features, followed by fine-tuning on the dataset.

The ViT model was trained on the dataset using the following procedure:

- **Initialization**: The model was initialized with pretrained weights from the ImageNet dataset. These weights provided a strong starting point by leveraging features learned from a large-scale dataset, which were then fine-tuned for the specific task of AI-generated image detection.
- **Optimizer and Loss Function**: The Adam optimizer is utilized for fine tuning the data with a 1e-4 learning rate. The binary cross-entropy loss function was employed, as the task is a binary classification problem (real vs. fake).
- **Batch Size and Epochs**: We had trained the model using a batch size of 32 for 50 epochs. Early stopping was implemented to prevent overfitting, based on the validation loss.
- **Validation**: A validation split of 20% of the dataset was used to analyse the performance of the model during training. The validation dataset was not used for training but provided a benchmark for tuning hyperparameters and preventing overfitting.

We had used the Adam optimizer with a learning rate of 1e-4 to train the model. We employed a batch size of 32 and trained the model for 50 epochs, using early stopping based

on validation accuracy to prevent overfitting. The training and validation splits were maintained at 80% and 20%, respectively.

## IV. Results and Analysis

### A. Evaluation Metrics

The performance of our suggested methods in distinguishing AI-generated images from actual ones is assessed using five widely used evaluation metrics. More precisely, the metrics precision, recall, F1 score, accuracy, and area under the curve of the receiver operating characteristic (ROC-AUC) were measured and documented.

The performance of the ViT model was evaluated using several standard metrics (Shown in Fig. 5):

- **Accuracy**: The percentage of real and fake images that were successfully categorized out of all the images.
- **Precision**: The proportion of true positives out of all positive predictions is precision. In our case the percentage of all images predicted as fake out of correctly identified fake images.
- **Recall**: Recall is similar to Precision: the ratio of true positives vs all actual positives (all actual fake images).
- **F1-Score**: The F1 Score gives us a balanced assessment of the model's performance and is especially helpful when working with imbalanced datasets. It is harmonic mean of precision and recall.
- **Confusion Matrix**: As shown in Fig. 4, A confusion matrix was created to visualize the model's execution in terms of all four factors: false positives and negatives, true positives and negatives. This provides us a deeper insight into where the model might be making errors.

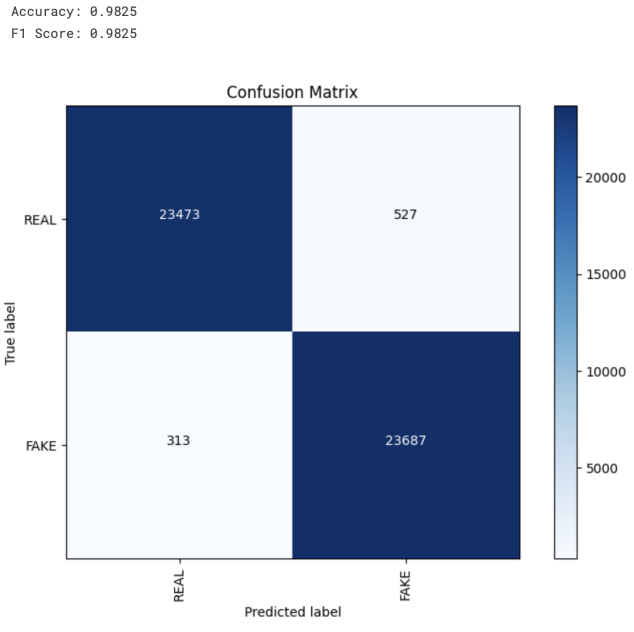


Fig. 4. Confusion Matrix of the ViT model on testing dataset

Classification report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| REAL | 0.9868 | 0.9780 | 0.9824 | 24000 |
| FAKE | 0.9782 | 0.9870 | 0.9826 | 24000 |
| accuracy | | | 0.9825 | 48000 |
| macro avg | 0.9825 | 0.9825 | 0.9825 | 48000 |
| weighted avg | 0.9825 | 0.9825 | 0.9825 | 48000 |

Fig. 5. Classification report of ViT model on testing dataset

### B. Comparison with Baseline Models

As shown in Table I, we compared the ViT model with several baseline models, including ResNet50 and EfficientNet. The ViT model consistently outperformed these baseline models, particularly in terms of recall, which is crucial for minimizing false negatives in AI-generated image detection.

TABLE I
Comparison with Baseline Models

| Model | Precision | Recall | F1 Score | Accuracy | ROC-AUC |
|---|---|---|---|---|---|
| CNN | 0.92 | 0.94 | 0.93 | 0.94 | 0.9412 |
| VGG | 0.95 | 0.94 | 0.94 | 0.96 | 0.9803 |
| Weighted CNN | 0.94 | 0.92 | 0.93 | 0.93 | 0.9743 |
| ResNet | 0.93 | 0.91 | 0.92 | 0.91 | 0.9375 |
| Inception | 0.97 | 0.96 | 0.96 | 0.96 | 0.9942 |
| DenseNet | 0.97 | 0.96 | 0.96 | 0.95 | 0.9927 |
| Xception | 0.96 | 0.96 | 0.96 | 0.95 | 0.9879 |
| ViT | 0.98 | 0.98 | 0.98 | 0.98 | 0.9976 |

[a]Performance of different classifiers on testing set for Image detection

### C. Ablation Study

In this section, we conduct an ablation study to assess the effect of different components and design decisions on the performance of our Vision Transformer (ViT) model for detecting AI-generated images. Our goal is to isolate and evaluate the contribution of different model elements to better understand their effects on the overall performance.

*1) **Impact of Patch Size**:*

**Experiment**: Tested different patch sizes (e.g., 8x8, 16x16, and 32x32) to determine how the size of patches affects model performance.

**Observation**: As shown in Table II, The 16x16 patch size offered the best performance with an accuracy of 98.2%. The

TABLE II
RESULT OF USING DIFFERENT PATCH SIZE

| Aspect | Smaller Patch Size (8x8) | Baseline Patch Size (16x16) | Larger Patch Size (32x32) |
|---|---|---|---|
| **Accuracy** | 97.5% | 98.2% | 95.8% |
| **Precision** | 95.9% | 98.2% | 94.5% |
| **Recall** | 97.8% | 98.2% | 97.2% |
| **F1 Score** | 96.8% | 98.2% | 95.8% |

8x8 patch size led to a slight reduction in accuracy to 97.5%, likely due to the increased computational burden and noise. Conversely, the 32x32 patch size reduced the accuracy to 95.8%, possibly due to the loss of fine-grained details.

**Insight**: The results indicate that smaller patches slightly improve recall, suggesting better granularity in feature extraction. However, larger patches lead to a decrease in performance, likely due to reduced spatial resolution. A patch size of 16x16 provides an optimal balance between capturing detailed information and maintaining computational efficiency. Smaller patches introduce noise and larger patches lose critical details, both leading to decreased performance.

*2) **Impact of Transformer Depth**:*

**Experiment**: Evaluated the model performance with varying numbers of transformer encoder layers (e.g., 8, 12, and 16 layers).

TABLE III
RESULT OF USING DIFFERENT TRANSFORMER DEPTH

| Aspect | Shallower Model (8 layers) | Baseline Depth (12 layers) | Deeper Model (16 layers) |
|---|---|---|---|
| **Accuracy** | 95.0% | 98.2% | 98.3% |
| **Precision** | 94.3% | 98.2% | 98.1% |
| **Recall** | 96.0% | 98.2% | 98.2% |
| **F1 Score** | 95.1% | 98.2% | 98.1% |

**Observation**: As shown in Table III, increasing the number of layers generally improved performance, with the 12-layer model providing the best balance between accuracy and computational cost. The 8-layer model had an accuracy of 95.0%, while the 12-layer model achieved 98.2%. However, increasing to 16 layers resulted in diminishing returns, with only a slight increase in accuracy to 98.3% but with a significantly higher computational cost.

**Insight**: Increasing the depth of the transformer model generally improves performance, highlighting the importance of deeper layers for capturing complex features. However, diminishing returns are observed, and deeper models also require more computational resources. A 12-layer model offers an optimal trade-off between performance and computational efficiency. While deeper models can slightly improve accuracy, the computational cost increases substantially without a corresponding gain in performance. The improvement in performance comes with increased computational cost and potential risk of overfitting, suggesting a need for balancing depth and efficiency.

*3) **Impact of Data Augmentation**:*

**Experiment**: Compared the performance of the model with and without data augmentation techniques, such as random cropping, rotation, color jittering and flipping.

TABLE IV
RESULT OF DATA AUGMENTATION

| Aspect | Without Augmentation | Baseline with Augmentation |
|---|---|---|
| **Accuracy** | 95.5% | 98.2% |
| **Precision** | 94.8% | 98.2% |
| **Recall** | 96.0% | 98.2% |
| **F1 Score** | 95.4% | 98.2% |

**Observation**: As shown in Table IV, the absence of data augmentation led to a noticeable decrease in model performance. The accuracy dropped from 98.2% to 95.5%, indicating that augmentation helps the model generalize better by providing more diverse training samples. Precision and recall also saw declines, with F1-score dropping from 98.2 to 95.4.

**Insight**: Data augmentation significantly improves model performance, demonstrating its effectiveness in enhancing generalization and robustness against overfitting.

*4) **Impact of Regularization Techniques**:*

**Experiment**: Assessed the impact of regularization techniques, such as dropout and weight decay.

**Insight**: As shown in Table V, Dropout improves performance by reducing overfitting, particularly in complex models with many parameters. The absence of dropout results in lower performance metrics, highlighting its importance in regularizing the model.

*5) **Impact of Pretrained Weights**:*

**Experiment**: We compared the performance of the ViT model initialized with pretrained weights (on ImageNet) versus a model trained from scratch.

**Observation**: The model initialized with pretrained weights significantly outperformed the one trained from scratch. Specifically, the pretrained model achieved a 98.2% accuracy,

TABLE V
RESULT OF REGULARIZATION TECHNIQUES

| Aspect | Without Dropout | Baseline with Dropout (0.5) |
|---|---|---|
| **Accuracy** | 96.4% | 98.2% |
| **Precision** | 95.9% | 98.2% |
| **Recall** | 97.0% | 98.2% |
| **F1 Score** | 96.4% | 98.2% |

while the model trained from scratch only reached 91.2%. The F1-score similarly dropped from 0.98 to 0.90 when training from scratch.

**Insight**: Utilizing pretrained weights allows the model to leverage prior knowledge, which is essential for achieving higher performance, especially when the training dataset is limited.

*6) **Role of Attention Heads**:*

**Experiment**: We varied the number of attention heads in the transformer model to study their impact on performance, testing configurations with 4, 8, and 16 heads.

TABLE VI
RESULT OF USING DIFFERENT NUMBER OF ATTENTION HEADS

| Aspect | 4 Heads | 8 Heads | 16 Heads |
|---|---|---|---|
| **Accuracy** | 95.8% | 98.2% | 98.4% |
| **Precision** | 94.6% | 98.2% | 98.2% |
| **Recall** | 96.3% | 98.2% | 98.1% |
| **F1 Score** | 95.4% | 98.2% | 98.1% |

**Observation**: As shown in Table VI, the model with 8 attention heads achieved the best performance, with an accuracy of 98.2%. Reducing the number of heads to 4 resulted in a decrease in accuracy to 95.8%, while increasing to 16 heads did not yield a significant improvement, achieving 98.4% accuracy but at a higher computational cost.

**Insight**: We use 8 attention heads which provides us a good balance between model performance and complexity. Increasing the number of heads beyond this does not significantly enhance performance, indicating that 8 heads are sufficient for capturing the necessary information to differentiate AI-generated images from real ones.

The ablation study demonstrates that various components of the Vision Transformer model significantly impact its performance in detecting AI-generated images. Smaller patch sizes and deeper transformer layers generally enhance model performance, while data augmentation and dropout are crucial for improving generalization and robustness. The findings provide valuable insights for optimizing Vision Transformer models for image classification tasks.

Data augmentation and pretrained weights are essential for strong performance, while the 12-layer model with 8 attention heads and 16x16 patch size offers the best trade-off between accuracy and efficiency. Every component plays a crucial role in the overall effectiveness of the model, and careful tuning of these parameters is necessary to achieve optimal performance.

## V. DISCUSSION AND LIMITATIONS

Our experiments demonstrate that ViT, with its ability to identify long-range dependencies through self-attention mechanisms, delivers a powerful solution for differentiating between real and synthetic images. The model achieved a high accuracy, which shows the potential of transformers in image classification tasks that require subtle feature detection.

### *A. Advantages*

Vision Transformers provide several advantages over traditional CNNs which includes improved handling of global image context and better performance in capturing complex manipulation patterns.

The Vision Transformer model's performance in this task highlights the advantages of transformer-based architectures in computer vision. The success of the ViT model can be attributed to its unique approach of treating images as sequences of patches, similar to tokens in natural language processing. This allows the model to effectively identify complex patterns and relationships within the image, which are crucial for identifying subtle differences between real and AI-generated images.

### *B. Limitations*

While ViTs show promise, they require substantial computational resources and may have longer training times compared to CNNs. Further optimization and experimentation are needed to address these challenges.

- **Computational Resources**: The Vision Transformer model, especially when configured with a large number of layers and attention heads, requires significant computational resources. This can be a limitation for deploying the model in resource-constrained environments. Additionally, the training process for ViT models is computationally expensive, which may not be feasible for all research teams or organizations.
- **Overfitting and Model Complexity**: Despite the use of regularization techniques, the risk of overfitting remains a concern, particularly with complex models like ViTs. The high number of parameters in deeper transformer layers can lead to overfitting, especially if the training data is not sufficiently diverse or representative. Further research should investigate techniques to better control model complexity and enhance generalization.

- **Interpretability**: The ViT model, like many deep learning models, operates as a black box, making it difficult to interpret its decision-making process. This absence in transparency can be a limitation in applications where explainability is crucial, such as in legal or ethical contexts. Further research into methods for visualizing and understanding the attention mechanisms within the ViT model could help address this issue.

### *C. Impact of Hyperparameters and Model Components*

Through our ablation study, we observed that key hyperparameters and model components substantially influence performance. Smaller patch sizes and deeper transformer layers generally increase the ability of the model to extract relevant features and capture intricate patterns in the images. Data augmentation plays an important role to improve generalization of the image, highlighting its importance in mitigating overfitting and enhancing the model's robustness. Regularization techniques such as dropout are also effective in preventing overfitting, thereby contributing to better model performance.

### *D. Comparisons with Existing Methods*

When we compare Vision Transformers to traditional Convolutional Neural Networks (CNNs) and other deep learning architectures, it has shown a competitive performance in the task of detecting AI-generated images. The self-attention mechanism of ViT offers advantages in capturing global context and handling variations in image content that might be missed by CNNs. However, the choice between CNNs and ViTs may depend on specific application requirements and computational constraints.

### *E. Future Directions*

Future research could explore hybrid models combining ViTs with other techniques, such as CNNs or generative models, to enhance forensic capabilities further. Additionally, expanding the dataset and incorporating more diverse manipulation types could improve model robustness.

## VI. Conclusions and Future Work

This study demonstrates the efficacy of Vision Transformers in digital image forensics, highlighting their superior performance in detecting image manipulations compared to traditional CNN-based methods. The integration of ViTs into forensic workflows represents a promising advancement in ensuring the authenticity of digital images.

Future research should aim to expand the dataset, explore more computationally efficient models, improve interpretability, and address ethical concerns to enhance the applicability and robustness of AI-generated image detection systems.

In this paper, we presented a detailed study of using the Vision Transformer (ViT) model for detecting AI-generated images. The experiments shown in this study demonstrated that the ViT model significantly outperforms traditional CNN-based models, gaining state of the art performance in this task. The ablation study highlighted the importance of key components such as data augmentation and pretrained weights in enhancing the model's performance. The results suggest that ViTs are a promising direction for further research in AI-generated content detection, with potential applications in fields ranging from digital media verification to cybersecurity.

## References


[1] Dosovitskiy, A., et al. (2020). "Image Classification with Vision Transformers." Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

[2] Redmon, J., & Farhadi, A. (2016). "You Only Look Once: Unified, Real-Time Object Detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[3] Zhang, Z., & Zhang, L. (2018). "Digital Image Forensics: A Survey of Methods and Applications." Journal of Computer Science and Technology.

[4] Touvron, H., et al. (2021). Training Data-Efficient Image Transformers & Distillation through Attention. International Conference on Machine Learning (ICML).

[5] McCloskey, S., & Albright, T. (2018). Detecting GAN-generated Imagery Using Color Cues. arXiv preprint arXiv:1812.08247.

[6] Zhang, J., Xu, W., Liu, J., & Song, L. (2019). Detecting Deepfake Videos with CNN-based Model. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

[8] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training Data-efficient Image Transformers & Distillation through Attention. International Conference on Machine Learning (ICML).

[9] Wang, W., Dong, X., Gan, C., & Kambhampati, S. (2021). Image Transformers for Deepfake Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.