# From Real to Rendered: Transfer Learning Approach for Detection of AI-Generated Images

Samihan Narendra Apte
7061559

Niranjan Madan
7062325

Karan Rajshekar
7062715

## Abstract

*With the growing prevalence of AI-generated images (AIGIs), their detection has become a crucial task in maintaining the integrity of visual media. While methods like watermarking have been proposed, the lack of standardization highlights the need for improved traditional detection approaches. This study explores the potential of transfer learning with ResNet50, enhanced by custom layers and loss functions, to boost AIGI detection accuracy beyond basic transfer learning. We compare our method against current techniques, including a self-supervised pipeline for ResNet50, to evaluate the overall effectiveness of our transfer learning approach. The CIFAKE dataset, a variant of CIFAR-10, serves as our training and evaluation foundation. Our novel contribution lies in integrating a fractal dimension analysis layer and a custom loss function, both designed to heighten the model's sensitivity to AIGI characteristics. Experimental results show that our method outperforms baseline models, achieving better performance than basic models. Our findings suggest that tailored architectural modifications and loss functions can significantly enhance transfer learning performance in this domain.*

## 1. Introduction

The rise in generative models presents new challenges in digital media authenticity. AI-generated images (AIGIs), produced by sophisticated models such as Generative Adversarial Networks (GANs) and diffusion models, have reached a level of realism that can easily deceive the human eye [1][2].

The potential for misuse of AIGIs, from spreading misinformation and creating deepfakes to facilitating identity theft, the malicious applications of this technology pose serious ethical and security challenges. Though there have advancements in detecting AIGIs using methods like watermarking or retrieval-based approaches, the lack of standardization along with strengths and weaknesses of each approach has resulted in no one model which can definitively be used for detection. That's why it is a necessity to explore into post-hoc classifier to differentiate between real and generated images.

We investigate the efficacy of transfer learning using a pre-trained ResNet50 model for AIGI detection. We conduct a comparative analysis, pitting our modified transfer learning approach against both a model trained from scratch and one utilizing self-supervised learning techniques. By leveraging pre-existing knowledge embedded in the ResNet50 architecture, we aim to enhance the model's ability to discern subtle differences between real and AI-generated images. Furthermore, motivated by post-hoc detection methods like analyzing the traces left by GANs while generating an image [3], we explore the integration of a fractal dimension analysis layer to identify micro-patterns in real images and a custom loss function to specifically target the anomalies often present in AIGIs.

## 2. Related Work

Studies have shown that transfer learning can be highly effective in image classification tasks, including deepfake detection and more general forms of AIGI detection. Previous studies have demonstrated the potential of CNNs with transfer learning in various classification tasks [4, 5], including AIGI detection. More recently, the benefits of using vision transformers for similar tasks have been highlighted [6], but such methods require vast amount of data to train the model.

The idea of fractal dimension analysis is based on observations related to real images where fractals i.e. micro-patterns are repeated throughout the image, no matter what the receptive field of our observation is [7]. The concept itself is quite old, but its implementation in image classification has been rare. Our hypothesis is that it is difficult for a generative model to produce these fractals and our aim to explore further whether we could exploit this natural tendency to detect AIGIs. The application of fractal dimension analysis in image processing, although not new, has seen limited use in AIGI detection. This study seeks to bridge this gap by integrating a fractal layer into our model. Additionally, the use of custom loss functions in deep learning, such as the combination of cross-entropy loss and feature similarity loss, has been shown to improve model robustness and accuracy.

## 3. Method

### 3.1. Dataset

We use CIFAKE dataset [8] for training all models to maintain consistency in this comparative study. The dataset consists of 120,000 images, equally split between real and AI-generated images with 60,000 real images and 60,000 fake images. The fake images in the dataset have been created using latent diffusion. As it is based on the CIFAR-10 dataset, it has 10 classes across which the data has been segregated. Each image is an RGB image with a resolution of 32x32 pixels. Out of total images, 80,000 images have been used for training, 20,000 as a validation set, and the remaining 20,000 were reserved as a test set. The training set is used to train and fit the models, the validation set is used for hyperparameter tuning and early stopping, and the test set is used to evaluate the final performance of the models.

### 3.2. Model Architecture

The study consists of five models, out of which four models are flavours of the same architecture with an addition at each step to measure the iterative performance fluctuation that might result from each of the addition. The fifth model uses the same backbone of ResNet50, but with self-supervised training rather than fully supervised training for the first four models.

Our modified model builds upon the ResNet50 architecture, leveraging pre-trained weights from ImageNet to provide a strong starting point. The reason for choosing ResNet50 as our backbone was because of its good overall performance and lower compute cost relative to current state-of-the-art models.

The architecture is modified with the addition of a fractal dimension analysis layer, which captures the self-similarity properties of feature maps, a characteristic often disrupted in AIGIs. The final model structures are as follows,

- **Baseline ResNet50**: All layers are trained on the CIFAKE Dataset without any pre-trained weights.
- **ResNet50 with pre-trained weights**: The model is initialized with weights pre-trained on ImageNet, followed by fine-tuning on the CIFAKE Dataset. All layers except the final classification layer are frozen. The final layer is fine-tuned on CIFAKE Dataset, the model weights in earlier layers are not updated during training.
- **Fractal Dimension Layer**: Calculates the fractal dimension of the CNN feature maps, appended to the existing feature tensor.
- **Custom Loss Function**: Combines cross-entropy loss with a feature similarity loss and a confidence penalty

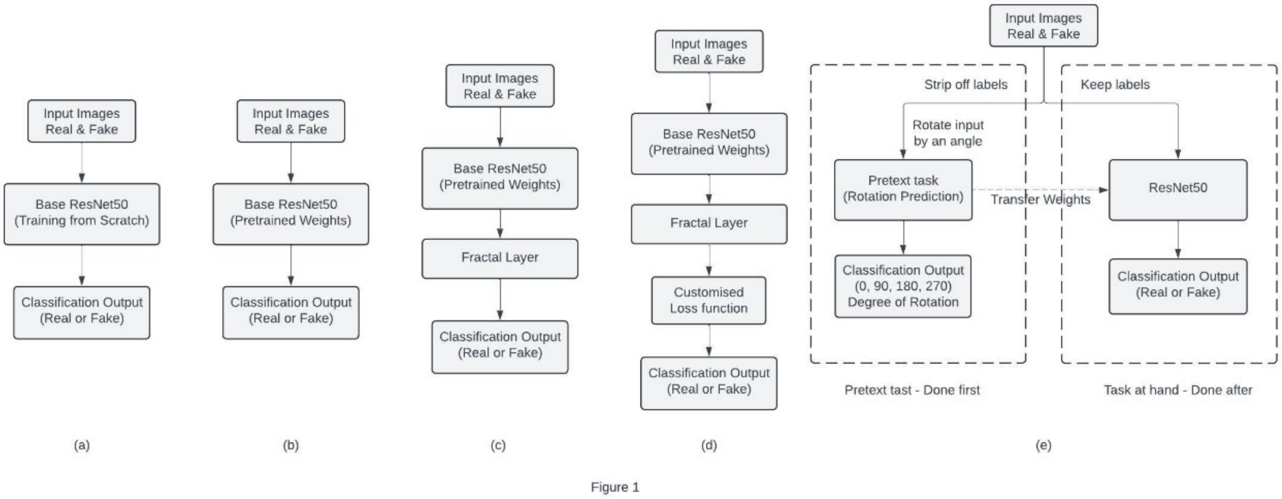

Figure 1. Overview of the structure of 5 models that are compared. (a) ResNet50 trained from scratch. (b) ResNet50 transfer learning with pre-trained weights where the output is modified for binary classification. (c) ResNet50 transfer learning with fractal dimension analysis layer added after pre-trained layers. (d) ResNet50 transfer learning with fractal dimension analysis layer added after pre-trained layers and trained on custom loss. (e) Self-supervised pipeline for the ResNet50 model as a backbone modified for binary classification.

to improve detection performance which takes advantage of the fractal dimension layer.

- **Self-supervised learning**: The self-supervised pipeline is employed on ResNet50 where we use the same CIFAKE dataset for both, pretext task and downstream task.

### 3.3. Fractal Dimension Analysis

The fractal dimension layer is based on the concept of fractals which can be found in most of the natural objects. The method we employ to estimate a fractal dimension of a shape in an image is called box-counting method [9], where the image is first initialized with variable size of boxes that span the entire image. The number of boxes that contain a part of the fractal object are counted and the same process is repeated with increasingly smaller size of boxes. We then map these on a log-log scale with boxes counted (N) and the size of the boxes (r). he slope of the resulting line on the log-log scale gives an estimate of the fractal dimension. This estimate is given by (1),

$$D = \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log(1/\epsilon)} \tag{1}$$

Where D is fractal dimension, $N(\epsilon)$ is the number of boxes of side length $\epsilon$ required to cover the image.

In practice, we can't take a limit to zero, we are estimating this using linear regression on the log-log scale which modifies the above formula to (2),

$$D \approx -\frac{\Delta \log N(\epsilon)}{\Delta \log \epsilon} \tag{2}$$

### 3.4. Custom Loss Function

The loss function we used for the final model is a combination of three types of losses:

**Cross Entropy Loss**: This is the standard loss which is used for classification tasks for measuring the dissimilarity between predicted probability distribution and true distribution.

**Feature Similarity Loss**: This is based on contrastive learning where the aim is to minimize the feature similarity between the two classes, real and fake. To find this, we first calculate a centroid of the feature maps based on real and fake images and find the cosine distance between the said feature maps.

$$L_{FS} = \cos(\mathbf{c_r}, \mathbf{c_f}) = \frac{\mathbf{c_r} \cdot \mathbf{c_f}}{\|\mathbf{c_r}\| \|\mathbf{c_f}\|} \tag{3}$$

where $\mathbf{c_r}$ and $\mathbf{c_f}$ are the centroids of real and fake features, respectively.

**Confidence Penalty**: To prevent the model from being overconfident in its predictions, we employ a confidence penalty that penalizes overconfident predictions by calculating the negative entropy of the softmax probabilities.

$$L_{CP} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} p_{ij} \log(p_{ij}) \tag{4}$$

where $N$ is the number of samples, $C$ is the number of classes, and $p_{ij}$ is the softmax probability of class $j$ for sample $i$.

**Total Loss**: The total loss is a weighted sum of these components:

$$L_{TOTAL} = \alpha L_{CE} + \beta L_{FS} + \gamma L_{CP} \tag{5}$$

where $\alpha$, $\beta$, and $\gamma$ are learnable parameters that determine the relative importance of each loss type.

### 3.5. Self-supervised Learning

To make effective use of the available resources, we use the same CIFAKE dataset for both, pretext task and downstream task. For pretext task, we first import the pre-trained ResNet50 model based on ImageNet dataset to get a headstart, and train all the layers of the model by stripping off the labels from the CIFAKE images for the task of predicting rotation of the image in four classes - 0, 90 ,180, 270 degrees.

For the downstream task of detecting real or fake image, we use the same weights from the pretext task, but modify the output layer to accommodate binary classification instead of four-class classification.

### 3.6. Training and Fine-tuning

The model is first trained on the CIFAKE dataset, which contains real and AI-generated images. The pretrained ResNet50 model undergoes fine-tuning, adjusting the weights to the specific task of AIGI detection. The fractal dimension layer is integrated during this phase, allowing the model to learn additional features relevant to detecting synthetic images. The primary loss function used is cross-entropy loss wherever besides the fourth model where we add the custom loss function. The models are trained on Adam optimizer with learning rate $10^{-4}$. The convergence was improved by applying learning rate scheduling.

## 4. Experimental Results and Analyses

### 4.1. Data Augmentation

To improve the generalization of our models and reduce the risk of overfitting, we applied several data augmentation

| Model | Test Loss | Test Accuracy | Test F1 Score |
|---|---|---|---|
| ResNet50 | 0.5571 | 71.49% | 0.7314 |
| ResNet50 + Pre-trained Weights | 0.4081 | 81.23% | 0.8464 |
| ResNet50 + Pre-trained Weights + Fractal Layer | 0.4185 | 83.81% | 0.8345 |
| ResNet50 + Pre-trained Weights + Fractal Layer + Custom Loss | 0.1316 | 84.75% | 0.8469 |
| Self-Supervised Method | 0.2029 | 95.91% | 0.9591 |

Table 1. Results

techniques during the training process. These augmentations include random horizontal flipping, and normalization based on the ImageNet dataset statistics. These augmentation techniques collectively enhance the model's ability to generalize well to unseen data.

### 4.2. Evaluation Metrics

We employ the following metrics to evaluate model performance:

- **Accuracy**: Measures the overall effectiveness of the model in correctly classifying both real and AI-generated images.

- **F1-Score**: Provides a balanced measure of precision and recall.

### 4.3. Results

The results indicate that the transfer learning approach significantly increases the model's performance over the baseline model. The use of pre-trained weights and freezing the layers results in a substantial increase in accuracy with less training time, demonstrating the power of transfer learning.

Our experiments show that the integration of the fractal dimension layer itself doesn't improve much when compared to basic transfer learning, but while using custom loss, the model performs slightly better which indicates that fractal dimension analysis does help in differentiating between real and fake images. The best-performing model with an accuracy of 95.91% was achieved through the self-supervised method. With high accuracy, F1 score, and a considerable reduction in test loss, the self-supervised method demonstrates it's effectiveness in AIGI detection.

## 5. Conclusion

This study presents a novel approach to detecting AI-generated images by combining transfer learning with custom layers and loss functions. The results show that the integration of a fractal dimension analysis layer, coupled with a designed loss function, can improve the detection capabilities of a ResNet50 model. The results suggest that these modifications allow the model to pick up on subtle differences that are often challenging to detect.

An important factor was the introduction of the fractal dimension layer, which helped the model to understand the texture and structural details of the image. This proved useful in identifying the intricate nuances of real images compared to AIGIs. In addition, the custom loss function balanced the process during training, so that the model didn't overfit and could generalize better to new unseen data. An important next step would be to test the model over a wide range of datasets, with higher resolution and divers AI- Content. Additionally, implementing the custom layers with self-supervised learning might bring about further improvement. Self-supervised learning learns from vast amounts of unlabeled data, which could be valuable in real-life scenarios where labeled data is scarce.

In summary, this is the significance of this study in improving deep learning models: specialized analytical layers with in-depth considerations and loss functions that are tailored to solve the advancing challenge of detecting AI-generated images. Since the development of AI technologies is quickly evolving, such advancement will be highly critical for designing systems with both effective and adaptive detection. Future work will explore the combination of self-supervised learning methods with these custom layers to further improve detection accuracy and model generalization.

# 6. References

[1] Safa Hassan Ali K A, and Chinchu Krishna S. Generating Text to Realistic Image using Generative Adversarial Network. In IEEE Access (2021).

[2] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In arXiv preprint arXiv:2112.10741v3, 2022.

[3] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. ON THE DETECTION OF SYNTHETIC IMAGES GENERATED BY DIFFUSION MODELS. In arXiv preprint arXiv:2211.00680v1, 2022.

[4] Hussain, M., Bird, J.J., Faria, D.R. (2019). A Study on CNN Transfer Learning for Image Classification. In Advances in Computational Intelligence Systems. UKCI, 2018.

[5] Manali Shaha, Meenakshi Pawar. Transfer learning for image classification. In Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018) IEEE Conference Record 42487; IEEE Xplore ISBN:978-1-5386-0965-1.

[6] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, Ross Girshick. Benchmarking Detection Transfer Learning with Vision Transformers. In arXiv preprint arXiv:2111.11429, 2021.

[7] J. M. Keller, R. M. Crownover and R. Y. Chen, ”Characteristics of Natural Scenes Related to the Fractal Dimension,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-9, no. 5, pp. 621-627, Sept. 1987.

[8] Bird, Jordan J., and Ahmad Lotfi. CIFAKE: Image classification and explainable identification of ai-generated synthetic images. In IEEE Access (2024).

[9] Qian Huang, Jacob R. Lorch, and Richard C. Dubes. Can the fractal dimension of images be measured? In Pattern Recognition, Vol. 27, No. 3, pp. 339-349, 1994.