AI Art Detection: Distinguishing Human-Created from AI-Generated Artworks

Matthew Dim   Sepideh Fatemi   Shivangi Sarkar   Mostafa Zakeri
Virginia Tech

Abstract

The detection of AI-generated artwork remains a difficult task for humans, often resulting in cases of both deception and false accusations. This paper leverages the existing AI-ArtBench dataset to explore the efficacy of various methods to classify organic and AI-generated artworks. We also study the accuracy of human intuition when asked to distinguish between AI-created works and human-created works. Our approaches include fast Fourier transforms (FFT), vision transformers (ViT), and ResNets. Through our findings, we determined that humans still struggle to distinguish between AI artwork and organic artwork. Additionally, many of our various classifier approaches achieved robust accuracy in classifying AI and non-AI art.

1. Introduction

With recent advances of artificial intelligence (AI), there is a growing difficulty in the ability of individuals in the art community to distinguish between human-created and AI-generated artworks. This has led to many scenarios of conflict in which artists are accused of using artificial intelligence (AI) to produce their artwork. However, other individuals are using generative AI models to deceive individuals into believing the work is authentic. These kinds of scenarios are occurring more often as AI model outputs continue to improve in quality, making it more difficult to distinguish between AI and human art. As a result, there is a growing need for tools that can successfully make this distinction, preventing misunderstandings and deception.

While AI detection tools exist, there is a lack of detection tools specializing in the domain of artwork. We aim to address this gap by exploring different approaches to implementing AI art classifiers. Additionally, we investigate what aspects of an image drive the decisions made by our classification models as well as our human survey participants. Through our efforts, we aim to support artists and institutions in maintaining authenticity and trust within the art community.

In this paper, we experiment with various models and methods to determine whether or not a piece of artwork is AI-generated and assess the accuracy of those methods. Additionally, we examine results from a questionnaire survey in which human participants were shown a combination of AI and human artworks and asked about the authenticity of those artworks.

To summarize, our contributions are:

- Applications of FFT to various classification algorithms to detect AI/human artwork.
- A fine-tuned ViT and ResNet capable of making distinctions between AI and human art.
- Insights on what regions of an image contribute to a model's classification decisions.
- Insights on human capability to detect AI art provided from a questionnaire survey completed by human participants.

2. Related Works

Recent advancements in detecting AI-generated content have explored a variety of methodologies. The Diffusion Reconstruction Error (DIRE) method [6] focuses on analyzing the diffusion process in AI models to identify AI-generated art. This approach demonstrates effectiveness in detecting generation artifacts introduced during the creation process. Similarly, the DE-FAKE method [4] utilizes multimodal embeddings to identify AI-generated content across diverse media types, including images. By incorporating multiple data modalities, DE-FAKE improves detection accuracy.

Wang et al. [5] proposed a technique that identifies CNN-based generated images by examining the unique fingerprints left by different GAN architectures. While successful for GAN-generated content, this method faces challenges when applied to newer diffusion-based models. Frank et al. [1] introduced a frequency-domain analysis approach to distinguish real images from AI-generated ones. By examining frequency characteristics, this method captures subtle artifacts that may not be evident in the spatial domain. Furthermore, Jeong et al. [2] analyzed fundamental trends in

AI content detection, evaluating both research-based methods, such as DIRE and DE-FAKE, and commercial tools like Hive and Optic.

While these methods exhibit promising results, challenges remain in accurately detecting highly realistic AI-generated art, particularly from advanced diffusion-based models. Our work builds upon these existing techniques, placing a specific emphasis on identifying the unique characteristics present in artworks.

3. Experiments

3.1. Dataset

We relied on the AI-ArtBench dataset available on Kaggle. This dataset consists of over 180,000 images, with one half comprising human-created artwork from the ArtBench dataset [3], and the remaining images being diffusion-generated artwork. The diffusion-generated portion of the dataset was produced from a latent diffusion model and standard diffusion model. The images capture 10 different art styles: Art Nouveau, Baroque, Expressionism, Impressionism, Post Impressionism, Realism, Renaissance, Romanticism, Surrealism, and Ukiyo-e. The dataset is conveniently separated into test and train sets with the train set including 155,015 images and the test set containing 30,000 images. Both the train and test sets include representation of all three image types (human-produced, latent diffusion-generated, and standard diffusion-generated), along with all 10 art styles.

3.2. FFT Implementation

First, we preprocessed the images by converting them to grayscale, resizing them, and applying the FFT to extract frequency-based features. We calculated the magnitude spectrum of the FFT to highlight distinguishing patterns and flattened these features into 1D vectors for analysis.

Next, we reduced the dimensionality of the features using principal component analysis (PCA) and truncated singular value decomposition (SVD), retaining 50 key components to balance computational efficiency with feature preservation. Using these reduced features, we trained and evaluated several machine learning classifiers, including random forest (RF), Logistic Regression (LR), support vector machine (SVM), K-nearest neighbor (KNN), and XGBoost. We assessed model performance based on accuracy, receiver operating characteristic (ROC) curve or area under curve (AUC) scores, and confusion matrices, comparing the results across both PCA and SVD. Visualization techniques such as ROC curves and bar plots were employed to analyze and compare the effectiveness of the classifiers and dimensionality reduction methods.

3.3. ViT Implementation

For our ViT approach, we began with a pretrained vit-base-patch16-224-in21k model and fine-tuned it on the train set from our Kaggle dataset. Prior to finetuning, the labels (AI vs real) were associated with the inputs. We selected AdamW as our stochastic optimization method and trained our model on cross entropy loss. After fine-tuning our model on labeling images of artwork as either AI or human-produced, we ran our model on the test portion of the dataset and observed its performance.

Following our real vs. AI detection, we used the same dataset to train a separate instance of the pretrained ViT model on art style labels, employing the same optimization and loss methods. We tested our fine-tuned art style classifier on the test portion of the AI-Artbench dataset and assessed its performance in comparison to the AI-detection classifier.

3.4. ResNet Implementation

Our ResNet50 model was derived from a pre-trained model from torch libraries labries ResNet50.IMAGENET1K_V2. This model was trained to classify over 1000 sets of images. In order to generate binary classifcation for our model, we froze the first $k-2$ layers where k is the number of layers in the model. Using the last two layers we trained our model to classify real vs. AI using the Adam optimizer. We then apply GradTorch to detect the feature space of AI vs. real art detection for understanding our model and what it learned.

3.5. Human Survey

To better understand the human ability in distinguishing AI-generated artworks from human-created ones, we conducted a user study with 30 participants. Each participant was shown a set of 20 images, consisting of 10 AI-generated artworks and 10 human-created artworks. All images were selected randomly, covering all 10 art styles present in the AI-ArtBench dataset. Participants were tasked with classifying each image as either *AI-generated* or *Human-created*. Here is the link to our survey.

4. Results

4.1. FFT Results

In this section, we aim to evaluate the effectiveness of frequency-domain features combined with machine learning models for distinguishing real artworks from AI-generated ones. The results demonstrate the strengths and limitations of various classifiers and dimensionality reduction methods.

As shown in Figure 1, we present two sample images: one where the label is correctly predicted and another where the prediction is incorrect. Their corresponding frequency-domain representations highlight the patterns that influence model performance, emphasizing the complexity of differentiating these images in certain cases.

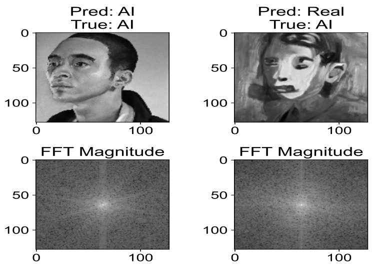


Figure 1. Two samples with real and predicted labels in real and frequency domain.

When comparing the accuracy of the models across PCA and SVD approaches, Figure 2 demonstrates that XGBoost consistently achieved the highest accuracy, followed by Random Forest and KNN. While SVD slightly improved the performance of XGBoost and Random Forest compared to PCA, KNN performed better with PCA. Interestingly, SVD had a significant positive impact on the performance of SVM, despite it being the overall lowest-performing model. These findings suggest that the choice of dimensionality reduction method can influence classifier performance, particularly for models like SVM that are sensitive to feature representation.

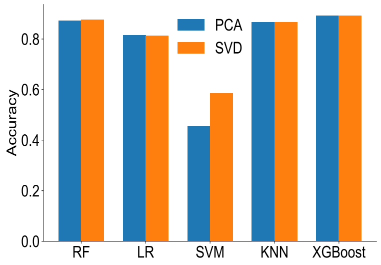


Figure 2. Accuracy of testing dataset for each classifier using PCA and SVD.

The ROC curves presented in Figure 3 further confirm the superiority of XGBoost and Random Forest, with AUC scores of 0.95 and 0.94, respectively. KNN followed with a score of 0.92, while Logistic Regression and SVM achieved 0.88 and 0.39, respectively. These results highlight the strong performance of tree-based models, particularly XGBoost, in distinguishing between real and AI-generated images. Overall, the combination of FFT-based features, dimensionality reduction, and robust classifiers proves to be effective for this classification task.

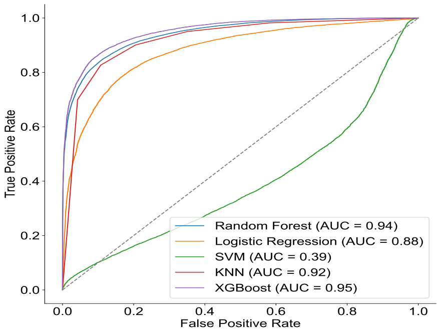


Figure 3. ROC curve for different classifiers using PCA.

4.2. ViT Results

Figure 4 displays the results from testing our binary ViT classifier to have an outstanding accuracy of 99% with a precision score of 1.0 and recall of 0.99. We were unable to determine any errors associated with the model's surprisingly high performance. However, we believe it may be attributed to the test set's similarity to the train set in terms of art style. Additionally, the remarkable size of our train set may have also contributed to such an impressive accuracy.

The accuracy of our art style classifier which was 86% had a lower accuracy in comparison to our binary classifier. As shown in Figure 5, our model performed exceptionally well with detecting Ukiyo-e style artworks, with high accuracies for detecting Baroque, Surrealism, and Renaissance styles as well. There were lower accuracies of 72% for Post Impressionism and Impressionism.

Label	**Precision**	**Recall**	**F1-score**
AI	1.00	0.99	1.00
Real	1.00	1.00	1.00

Figure 4. Performance metrics for binary classification for AI and real images with our finetuned ViT.

Label	**Precision**	**Recall**	**F1-score**
Ukiyo-e	1.00	0.99	1.00
Art Nouveau	0.89	0.89	0.89
Post-Impressionism	0.72	0.74	0.73
Impressionism	0.72	0.79	0.75
Expressionism	0.89	0.82	0.85
Baroque	0.92	0.88	0.90
Surrealism	0.95	0.92	0.94
Realism	0.79	0.79	0.79
Romanticism	0.82	0.84	0.83
Renaissance	0.90	0.93	0.91

Figure 5. Art style classification metrics for different art styles with our finetuned ViT.

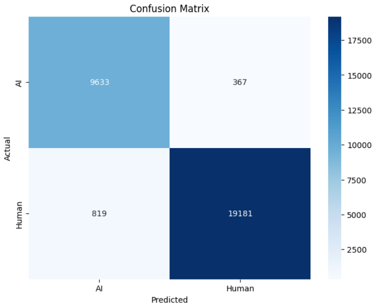


Figure 6. ResNet-50 Confusion Matrix

4.3. ResNet Results

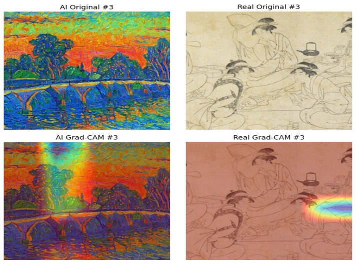
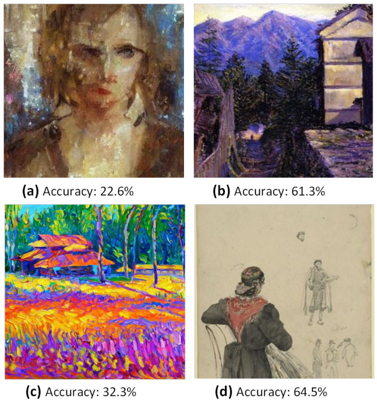


Figure 7. ResNet-50 Real vs. AI GradCam for weight predictions

The ResNet-50 classifier was able to achieve a classification accuracy of 96% over a test dataset of 30,000 images. The precision score obtained was 0.96 and recall of 0.96 making it a competitive model given the short training time due to freezing layers. From Figure 6 we see that even when our model predicted incorrectly, it tended to lean on the safer side of favoring AI generated images. This means that it was more likely to classify AI imagery and in the context of AI vs. real this would decrease the likelihood of a mis-classification of important human imagery as AI results. Additionally from Figure 7 we see that the model tended to favor smoother regions from the AI imagery versus the fine-details highlighted from specific regions in the real art. This highlights the differences in our models to focus on AI imagery as opposed to our ViT model.

4.4. Human Survey Results

On average, participants achieved an accuracy of 66.4%. The scores ranged from 7/20 (35%) to 16/20 (80%), reflecting a varying degree of success among participants. The accuracy for identifying human-created artworks was notably higher at **74.2%**, whereas the accuracy for AI-generated images dropped to **58.7%**. This discrepancy suggests that participants found it more challenging to identify AI-generated content, often misclassifying it as human-created.



(a) Accuracy: 22.6% **(b)** Accuracy: 61.3%

(c) Accuracy: 32.3% **(d)** Accuracy: 64.5%

Figure 8. Examples of AI-generated and human-created artworks with corresponding human classification accuracies. (a) and (c) are AI-generated, (b) and (d) are Human-created arts.

Figure 8 illustrates examples of AI-generated and human-created artworks alongside their respective classification accuracies from the user study. AI-generated images, particularly those with smooth or abstract textures, were often misclassified, achieving accuracies as low as **22.6%** and **32.3%**. In contrast, human-created images with more identifiable details and imperfections were classified with higher accuracy, such as **61.3%** and **64.5%**. These results emphasize the challenges humans face in detecting AI-generated content, especially when the generative outputs closely mimic human artistry.

The findings from this user study emphasize the need for automated tools to assist in detecting AI-generated content. While humans rely on visual cues like texture, color, and artistic detail, our ViT and ResNet classifiers provide consistent and robust performance, addressing the challenges posed by increasingly realistic AI-generated artworks.

5. Conclusion

Our research is a case-study of relevant methods for AI image detection using light weight models. We compare methods such as the Fast Fourier Transform (FFT), Vision Transformer (ViT) and ResNet-50 for application in image and style classification. We found that our methods were all effective in classification in a fast and understandable model as compared to diffusion. More work needs to be done to combine the models and test on AI generated data from other sources such as DALLE. Additionally, more work should be done reconstructing relevant diffusion methods to downscale the cost of model construction and find other light-weight methods for detection.

References

[1] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *Proceedings*

of the 37th International Conference on Machine Learning, volume 119 of *PMLR*, Online, 2020. 1

[2] Anna Yoo Jeong Ha, Josephine Passananti, Ronik Bhaskar, Shawn Shan, Reid Southen, Haitao Zheng, and Ben Y. Zhao. Organic or Diffused: Can We Distinguish Human Art from AI-generated Images? *arXiv preprint arXiv:2402.03214*, 2024. 1

[3] Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. The artbench dataset: Benchmarking generative models with artworks, 2022. 2

[4] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models. pages 3418–3432, 2023. 1

[5] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8692–8701, Seattle, WA, USA, 2020. 1

[6] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li. DIRE for Diffusion-Generated Image Detection. pages 22388–22398, 2023. 1