Enhancing AI-Generated Image Detection: A Comparative Study of CNNs, Transformers, and Contrastive Learning

DL4DS Boston University

Viktoria Zruttova, Junhui Cho, Cordell Cheng

February 26, 2025

Abstract

AI has reached a point where it can generate highly realistic faces, scenes, and objects. This study addresses the problem of distinguishing AI-generated visuals from authentic photographs using a unique dataset, "AI vs. Human-Generated Images," from a Kaggle competition. Unlike conventional datasets, this dataset provides paired images where each real image has a corresponding AI-generated counterpart, allowing for direct comparative analysis. We leverage this structured pairing within a deep learning framework, incorporating convolutional neural networks (CNNs) and transformer-based architectures to develop robust classifiers. In addition, we explore contrastive learning to enhance feature discrimination, hypothesizing that it improves generalization by enforcing a more distinct separation between real and AI-generated images.

Introduction

AI-generated images – including deepfakes and other content created by generative adversarial networks (GANs) or diffusion models – have advanced to a point of being almost indistinguishable from real images. These technologies enable the creation of hyper-realistic faces, scenes, and objects, raising concerns about the spread of misinformation and the erosion of trust in digital media. Early research in this domain focused on deepfakes of human faces (e.g., face swaps or manipulations), but the challenge now extends to natural images generated by AI (such as artworks or scenery), which has only recently become a focus of studies [2]. Detecting AI-generated content reliably is difficult because modern fakes closely mimic the appearance of genuine images, often containing only subtle inconsistencies or artifacts [1]. Some detection approaches have exploited such artifacts – for example, detecting unnatural blending, color inconsistencies, or frequency-domain irregularities [16]. However, as generation methods improve and remove obvious artifacts, detectors that rely on fixed cues may become fragile [8].

Another major challenge is the generalization of detectors to new or previously unseen types of AI-generated images. Many detection models perform well when evaluated on fake images similar to those they were trained on, but their accuracy degrades significantly on novel manipulation techniques [16, 18]. In practice, a detector must handle an open-set scenario, where the specific characteristics of fakes at test time may differ from the training data. Limited generalizability can hinder real-world deployment of deepfake detectors, as they may fail to recognize new attack methods. For instance, a model trained on faces manipulated by one GAN might not detect faces generated by a different, more advanced GAN [11]. This issue has been highlighted in studies showing that cross-dataset detection performance can drop sharply compared to in-dataset results. Developing techniques to make detectors more robust against unseen manipulations is therefore an active area of research.

In this context, we consider two prominent types of deep learning architectures for image classification: CNNs and vision transformers. CNNs have been the cornerstone of image recognition for years, excelling at learning hierarchical feature representations via convolution and pooling operations. Transformers, on the other hand, employ self-attention mechanisms to capture long-range dependencies in the image. Vision transformers have shown promise in image classification tasks, but they often require large training datasets and are computationally intensive, whereas CNNs come with strong inductive biases for locality that make them effective even with limited data [9]. In the realm of deepfake detection, CNN-based models (e.g. Xception, EfficientNets) initially dominated and achieved high accuracy on benchmark datasets, but researchers are now exploring transformer-based models (such as ViT or Swin Transformer) for potentially improved performance. Comparing these two architecture families in the specific context of AI-generated image detection is important for understanding their relative strengths. For example, CNNs might better capture fine-grained pixel artifacts, while transformers could detect inconsistencies in global image context. Recent work has indeed suggested that each may have unique advantages for deepfake detection.

Finally, contrastive learning has emerged as a promising technique to improve representation learning for classification tasks. In contrastive learning, models learn by pulling semantically similar examples closer and pushing dissimilar examples apart in the feature space. We hypothesize that contrastive learning can be especially beneficial for distinguishing real vs fake images, as it can encourage the model to develop a representation that maximally separates authentic images from AI-generated ones. Prior studies have begun to apply contrastive learning to deepfake detection to address generalization issues [16, 14, 10, 13].

“AI vs. Human-Generated Images“ is a dataset introduced specifically for authenticity detection. The dataset consists of authentic images from Shutterstock paired with AI-generated counterparts created using cutting-edge generative models such as DeepMedia.

In other words, for each real image, there is a closely matching synthetic image depicting the same or similar content. This structured pairing enables a direct comparison between real and AI-generated content, providing a robust foundation for model training and evaluation.

The insights from this work have significance for digital content verification, helping pave the way for tools that can automatically flag AI-generated images in news media, social networks, or other platforms – a key step toward mitigating the spread of misleading visual content.

Related Work

CNNs and Transformers for AI-Generated Image Detection

Deep learning has shown great promise in tackling the AI-image detection problem. In particular, convolutional neural networks (CNNs) have long been the backbone of image classification and have been applied successfully to discriminate between real vs. generated images. CNNs excel at learning local visual features and textures, which is advantageous when AI-generated images contain subtle pixel-level artifacts [6]. Indeed, recent work by Bird and Lotfi achieved about 93% classification accuracy distinguishing real photographs from diffusion-generated fakes using a CNN on the CIFAKE dataset [3]. Their experiments revealed that the CNN's attention was not on the main subjects of images but rather on background irregularities, suggesting these models identify imperceptible noise patterns left by generation processes [4].

Complementing CNNs, transformer-based vision models (Vision Transformers, ViTs) have emerged as powerful image classifiers by capturing global context through self-attention mechanisms. Transformers consider the entire image patch relationships, potentially enabling the detection of more holistic anomalies in AI-generated images that might evade localized feature detectors [5]. Hossain et al. explored both CNN and ViT architectures for AI-generated image detection, finding an optimized CNN slightly outperformed the transformer, reaching 96.3% accuracy on a similar real-vs-fake image task [7]. This suggests that, despite transformers' success in general vision tasks, CNNs with their inductive bias for texture may still hold an edge in detecting the current generation of fakes. In practice, a hybrid or ensemble approach could leverage both local and global feature modeling to improve detection robustness [6].

CNNs and Transformers-Based for Deepfake Detection

CNNs have been the foundation of most early deepfake image detectors. Researchers have applied off-the-shelf architectures (pretrained on ImageNet) and fine-tuned them to classify images as real or fake. For instance, the Xception network was used by Rößler et al. in the FaceForensics++ benchmark and achieved high accuracy in detecting manipulated facial images [15, 11]. In general, CNN detectors can achieve excellent in-dataset performance; one recent study reported accuracy above 97% (with AUC near 99%) on certain deepfake benchmarks using a CNN model [15]. Common CNN-based approaches often look for subtle artifacts left by the generation process – for example, unnatural textures, missing reflections, or eye and facial aberrations in deepfake faces. Some methods explicitly analyze frequency-domain patterns, operating on the observation that GAN-generated images may have spectral discrepancies from natural images [16]. Despite their success on known data, a well-documented drawback of these CNN classifiers is poor generalization. When evaluated on a different dataset or a new type of fake, their performance can drop drastically. This is because CNNs may overfit to particular artifacts or data characteristics present in the training set. For example, a model trained to detect FaceSwap manipulations might latch onto compression artifacts or blending errors specific to that method; if confronted with a NeuralTextures-generated fake (with different artifact signatures), it might fail to detect it. Thus, improving the general robustness of CNN-based detectors remains a key issue.

Transformers process images in a fundamentally different way, using global self-attention to relate patches of an image, which could be advantageous for detecting inconsistencies that span across an image (such as misaligned lighting or context that CNNs looking at local patches might miss). Vrizlynn Thing recently compared several CNN and transformer architectures on multiple deepfake image datasets. That study found that both kinds of models can attain high accuracy on deepfake detection given sufficient training data, with top performances exceeding 90% accuracy and 99% AUC on benchmarks like FF++ and Google's DeepFake Detection dataset [15]. Notably, a Swin Transformer (a type of vision transformer) was shown to perform on par with a strong CNN (ResNet) on the challenging DFDC dataset (Facebook's Deep Fake Detection Challenge). This suggests that transformers are a viable alternative to CNNs for this task. Nonetheless, transformers often require careful training and regularization when data is limited. A general observation is that CNNs bring strong local feature biases that are helpful for detecting pixel-level artifacts, whereas transformers might better capture global anomalies (e.g., a face that doesn't match its surroundings). Combining these strengths is an open research question.

Contrastive Learning for Fake Image Detection

In recent studies, contrastive learning has been applied to address the aforementioned generalization problem. The core idea is to learn an embedding where authentic vs. synthetic images are well-separated. Xu et al. employed a supervised contrastive learning (SupCon) loss to train a deepfake detector, forcing representations of real images to cluster together and away from representations of fake images [16]. By doing so, the model learned to emphasize features that consistently distinguish real from fake rather than features specific to one fake type. In a true open-set evaluation (where the model was tested on an entirely novel fake type), their contrastive-learning-based model achieved about 78.7% accuracy, significantly outperforming a standard CNN, and further improved to 83.99% when combined with an Xception model in a fusion approach. This demonstrates that contrastive training can yield more generalizable detectors. Other works have explored combinations of unsupervised and supervised contrastive learning. One approach first learns features without labels by maximizing agreement between two augmented views of the same image (unsupervised contrastive pretraining) and then fine-tunes with supervised contrastive or classification loss [17].

Proposed Work

Building on the insights from prior work, we propose a systematic comparison between Convolutional Neural Networks (CNNs) and transformer-based models for AI-generated image detection, augmented with contrastive learning to enhance model generalization. We aim to evaluate both local and global feature extraction capabilities by leveraging CNNs' strength in texture learning and transformers' ability to capture long-range dependencies. Additionally, we investigate the impact of contrastive learning on both architectures to improve feature separation between real and fake images.

The primary objectives of our proposed work are as follows:

1. **Dataset Exploration and EDA:** We will begin by exploring the dataset through basic statistics and exploratory data analysis (EDA) to understand its structure, distribution, and potential biases. This step will involve visualizing class distributions, identifying any data imbalances, and investigating correlations between different image characteristics. Basic statistics such as mean, variance, and class distribution will be calculated to provide insight into the dataset's quality and suitability for training.

2. **Preprocessing:** Before feeding the images into the models, we will apply preprocessing steps to improve the data quality and model performance. These steps include:

   - **Resizing:** Images will be resized to a consistent shape, ensuring compatibility with the model input requirements. Typically, we will resize images to 224×224 or 256×256 pixels, depending on the model architecture. The resizing process will be mathematically represented as:

     $$\text{Resized Image} = f(\text{Original Image}, W, H),$$

     where W and H are the target width and height of the image.

   - **Normalization:** To standardize the pixel values and improve convergence during training, images will be normalized to have zero mean and unit variance. This can be expressed mathematically as:

     $$\hat{x} = \frac{x - \mu}{\sigma},$$

where x is the pixel value, μ is the mean of pixel values in the dataset, and σ is the standard deviation of the pixel values in the dataset. This scales the pixel values to have a zero mean and a unit variance, improving the model's ability to converge faster.

- **Augmentation:** We will apply data augmentation techniques to improve model robustness and prevent overfitting. Common augmentations will include random transformations, such as:
  - **Rotation:** Randomly rotating the image by a specified angle, typically between $-30°$ and $30°$, to simulate different orientations.
  - **Flipping:** Horizontal or vertical flipping to introduce mirror versions of the images, which helps the model learn invariant features.
  - **Zooming:** Randomly zooming in or out on the image to simulate different scales.
  - **Brightness/Contrast Adjustment:** Randomly changing the brightness or contrast of the image to account for varying lighting conditions in real-world images.
  - **Color Jittering:** Randomly altering the color balance of the image, including adjustments to hue, saturation, and exposure.
  - **Cropping:** Random cropping of the image to simulate partial views of objects, helping the model to focus on key features within the image.

3. **Model Comparison:** We will compare the performance of CNN-based and transformer-based models, including common architectures such as ResNet50, EfficientNet, and Vision Transformers (ViT), in detecting AI-generated images. The comparison will consider the strengths of CNNs in extracting local pixel-level features and the global contextual understanding offered by transformers through self-attention mechanisms.

4. **Incorporating Contrastive Learning:** We will enhance both the CNN and transformer models with contrastive learning. Specifically, we will use a supervised contrastive loss function to enforce a better feature separation between real and fake images in the learned embedding space. The loss function will be formulated as follows:
$$\mathcal{L}_{\text{SupCon}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(z_i, z_{i+})/\tau)}{\sum_{a=1}^{N}\exp(\text{sim}(z_i, z_a)/\tau)},$$
where z_i represents the embedding of the i-th image, z_{i+} is its positive pair, $\text{sim}(\cdot,\cdot)$ is the similarity measure (e.g., cosine similarity), and τ is a temperature parameter that controls the sharpness of the probability distribution.

5. **Generalization Evaluation:** We will evaluate the models' generalization capabilities by testing them on a diverse set of AI-generated images and real images that may differ from the training set. In particular, we hypothesize that contrastive learning will enhance generalization by encouraging models to focus on the most consistent and distinguishing features across various image types.
6. **Performance Metrics:** Model performance will be measured using standard classification metrics such as accuracy, Area Under the Curve (AUC), and F1-score, as well as additional evaluation in terms of generalization and robustness. Specifically, we will analyze the models' ability to maintain high performance across datasets with different characteristics (e.g., images generated by various AI techniques).

In summary, this work seeks to bridge the gap between CNN and transformer architectures in the context of AI-generated image detection while leveraging contrastive learning to improve the models' robustness and generalization. Our findings will provide valuable insights into the capabilities of these models and contribute to the development of more effective techniques for AI image forensics.

Datasets

We will use a curated dataset of AI-generated and Authentic Images obtained from Kaggle [12].

Authentic Images: Sourced from the Shutterstock platform, the dataset includes a diverse array of genuine images spanning various categories. Notably, approximately one-third of these images prominently feature human subjects, ensuring a balanced representation.

AI-Generated Images: Each authentic image is paired with an AI-generated counterpart. These synthetic images are created using a state-of-the-art generative model from DeepMedia, providing a rich set of examples for training and evaluation.

Evaluation

Our model's performance will be assessed using the following metrics:

- **Accuracy**: Measures the overall classification correctness, defined as:

  $$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

  where TP, TN, FP, and FN stand for True Positives, True Negatives, False Positives, and False Negatives, respectively.

- **Precision & Recall**: Evaluate the balance between false positives and false negatives:
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **F1-score**: The harmonic mean of precision and recall, providing a single metric for performance evaluation:
$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Area Under Precision-Recall Curve (AUC-PR)**: This metric is particularly useful when the dataset is imbalanced, focusing on the detection of the minority class (AI-generated images):
$$AUC - PR = \int_0^1 \text{Precision}(r)\, \mathrm{d}r$$

- **Matthews Correlation Coefficient (MCC)**: A balanced metric that accounts for all four confusion matrix categories and is particularly useful when dealing with imbalanced datasets:
$$MCC = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

- **Area Under ROC Curve (AUC-ROC)**: Evaluates the overall discriminative ability of the model by considering the trade-off between the true positive rate and false positive rate:
$$AUC - ROC = \int_0^1 \text{TPR}(f)\, \mathrm{d}f$$

- **Log Loss (Cross-Entropy Loss)**: Used when the model outputs probabilities rather than hard classifications, penalizing false classifications more heavily:
$$\text{Log Loss} = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right)$$
where y_i is the true label and p_i is the predicted probability.

- **Additional Metrics**: More evaluation criteria may be considered as necessary based on the model's requirements and deployment context.

Timeline

1. **Week 1:** Literature Review & Dataset Exploration
2. **Weeks 2:** Data Preprocessing & Augmentation
3. **Weeks 3-4:** Model Selection + Implementation (Traditional Supervised Learning)
4. **Weeks 5-6:** Model Selection + Implementation (Contrastive Learning)
5. **Weeks 7-8:** Hyperparameter Tuning
6. **Week 9:** Model Evaluation vs. Scratch
7. **Week 10:** Comparative Analysis
8. **Weeks 11:** Finalizing Results & Report Writing

Conclusion

As AI-generated imagery becomes increasingly sophisticated, distinguishing between human-created and AI-generated images is a critical challenge in areas such as media verification, digital forensics, and content authenticity. This project aims to address this challenge by leveraging deep learning techniques to build a robust classification model using CNNs and vision transformers combined with contrastive learning. Through rigorous experimentation with different architectures, hyperparameter tuning, and explainability analysis, we seek to develop a high-accuracy model capable of detecting AI-generated images.

Beyond achieving strong classification performance, this project will provide valuable insights into the distinguishing characteristics of AI-generated images, contributing to the broader discourse on AI transparency and misinformation prevention.

References

[1] Samah S. Baraheem and Tam V. Nguyen. Ai vs. ai: Can ai detect ai-generated images? *Journal of Imaging*, 9, 10 2023.

[2] Lorenzo Baraldi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara. Contrasting deepfakes diffusion via contrastive learning and global-local similarities. 7 2024.

[3] Jordan J. Bird and Ahmad Lotfi. Cifake: Image classification and explainable identification of ai-generated synthetic images. *IEEE Access*, 12:15642–15650, 2024.

[4] Druthik Sai Chinta, Sindhusha Kamineni, Ratna Pujitha Chatragadda, and Sujatha Kamepalli. Analyzing image classification on ai-generated art vs human created art

using deep learning models. In *2024 3rd International Conference on Electrical, Electronics, Information and Communication Technologies, ICEEICT 2024*. Institute of Electrical and Electronics Engineers Inc., 2024.

[5] Stefano Filipazzi, Christopher D. Hacon, and Roberto Svaldi. Boundedness of elliptic calabi-yau threefolds. 12 2021.

[6] FlyPix AI. Image recognition models: Cnns and the future of ai vision, 2025. Accessed: 2025-02-27.

[7] Md Zahid Hossain, Farhad Uz Zaman, and Md Rakibul Islam. Advancing ai-generated image detection: Enhanced accuracy through cnn and vision transformer models with explainable ai insights. In *2023 26th International Conference on Computer and Information Technology, ICCIT 2023*. Institute of Electrical and Electronics Engineers Inc., 2023.

[8] Fernando Martin-Rodriguez, Rocio Garcia-Mojon, and Monica Fernandez-Barciela. Detection of ai-created images using pixel-wise feature extraction and convolutional neural networks. *Sensors*, 23, 11 2023.

[9] José Maurício, Inês Domingues, and Jorge Bernardino. Comparing vision transformers and convolutional neural networks for image classification: A literature review, 5 2023.

[10] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. 12 2017.

[11] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. 1 2019.

[12] Alessandra Sala, Harshika, Manuela Jeyaraj, Margarita Pitsiani, Niamh Belton, and Toma Ijatomi. Detect ai vs. human-generated images. `https://kaggle.com/competitions/detect-ai-vs-human-generated-images`, 2025.

[13] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, 12 2019.

[14] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin L, and Rongrong Ji. Dual contrastive learning for general face forgery detection. 12 2021.

[15] Vrizlynn L L Thing. Deepfake detection with deep learning: Convolutional neural networks versus transformers. Technical report.

[16] Ying Xu, Kiran Raja, and Marius Pedersen. Supervised contrastive learning for generalizable and explainable deepfakes detection. Technical report.

[17] Jun Shuai Zheng, Yi Chao Zhou, Xi Yuan Hu, and Zhen Min Tang. Deepfake detection with combined unsupervised-supervised contrastive learning. In *Proceedings - International Conference on Image Processing, ICIP*, pages 787–793. IEEE Computer Society, 2024.

[18] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. Technical report.