# Detecting AI-Generated Art -A Self-supervised Approach

**Jiaqi Li**
Department of Computer Engineering
New York University
Brooklyn, NY 11201
jl15442@nyu.edu

**Jinrui Zhang**
Department of Computer Engineering
New York University
Brooklyn, NY 11201
jz6578@nyu.edu

## Abstract

Recently, AI-Generated Content (AIGC) is getting more and more attention. From images to videos, AIGC not only has an advantage in quantity, but also gradually catches up with humans in quality. Therefore, the need to distinguish AIGC from human works is also increasing. For various reasons such as copyright and platform policies, we urgently need to know whether the content is generated by AI. Among the fields where AIGC is applied, AI-generated art is one of the most widespread and also one of the most controversial. In view of this, we try to use self-supervised learning methods to develop a deep learning model to distinguish AI-generated art from human art. The code of our work is available at https://github.com/JQLLICIA/AML-PJ

## 1 Introduction

As one of the earliest AIGC fields to be widely used, August 2022 saw models such as stable diffusion begin to produce a large amount of AI generated art, even three months before the release of chatgpt3.5. Since then, the technology of AI-generated images has continued to advance, gradually becoming comparable to human art (see Figure 1), and the art community has become more and more controversial, especially against its application in commercial works [1].

In view of the powerful imitation ability of AI, many platforms have introduced policies requiring users to label works generated by AI [2]. However, these policies either require strong user awareness, which proves to be unreliable [3]; or they require a method, or model, to automatically identify AI-generated content.

Self-supervised learning is a method within unsupervised machine learning. Its key feature is that it doesn't require manually labeled data. Instead, it employs a variety of methods, called pretext tasks, to generate supervisory signals from the data itself to train models. The data features learned through the pretext task are used in different downstream tasks, such as data classification or reconstruction. Self-supervised learning is particularly effective when dealing with large volumes of unlabeled data. Given the difficulty of labeling AI generated art, self-supervised learning is a natural candidate.

Our goal is to train a self-supervised learning model on a dataset of AI-generated art and human-generated art, and transfer it to downstream tasks to try to distinguish them. To achieve this goal, we constructed a data set containing 900 art pictures in 3 different art styles to ensure the internal balance of the data. We trained our model on this dataset and simulated different situations with the proportion of labeled data in the dataset. Ultimately, our self-supervised learning model achieved competitive results with fully supervised learning.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Figure 1: Samples of AI-generated art (left) and human art (right).

## 2 Background and Related Work

Our work comes from two areas: self-supervised learning and research on AI-generated image detection.

### 2.1 Self-supervised learning in computer vision

Self-supervised learning in computer vision has gained significant attention in recent years. It offers a promising approach to learning visual representations from unlabeled data. Several methods have been proposed in the literature to achieve this. He et al. introduced Momentum Contrast for Unsupervised Visual Representation Learning [5], which focuses on learning representations by contrasting positive pairs with negative samples. Similarly, Gidaris et al. proposed Unsupervised Representation Learning by Predicting Image Rotations [11], where the network learns by predicting the rotation of an input image. These approaches aim to leverage the inherent structure within the data to learn meaningful representations.

Furthermore, Chen presented A Simple Framework for Contrastive Learning of Visual Representations [12], which emphasizes the use of contrastive learning to train visual representations. This method encourages similar representations for augmented views of the same image and dissimilar representations for different images. Additionally, Noroozi explored Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles [13], where the network learns by solving jigsaw puzzles created from input images. This encourages the model to understand the spatial relationships between different parts of the image.

In the domain of computer vision, Ledig et al. proposed Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network [14], which focuses on enhancing the resolution of images using a generative adversarial network. This work demonstrates the application of self-supervised learning in improving the visual quality of images. Moreover, Zeng et al. introduced Learning Pyramid-Context Encoder Network for High-Quality Image Inpainting [15], which leverages deep feature learning and adversarial training for semantic image inpainting. While not directly related to self-supervised learning, these works showcase the broader applications of unsupervised and self-supervised techniques in computer vision tasks.

In summary, self-supervised learning in computer vision encompasses various techniques such as contrastive learning, rotation prediction, and jigsaw puzzle solving, aiming to learn rich visual representations from unlabeled data. These methods have shown promising results and have the potential to advance the field of computer vision by reducing the reliance on labeled datasets.

The loss function used in our work is triplet loss, proposed by F Schroff et al. [5] in 2015. It defines an anchor, a positive sample similar to anchor, and a negative sample dissimilar to anchor. Then the pretext task is that the model's output for the anchor and the positive sample should be close, and the output for the anchor and the negative sample should be far away. The corresponding loss function is then defined by the two distances.

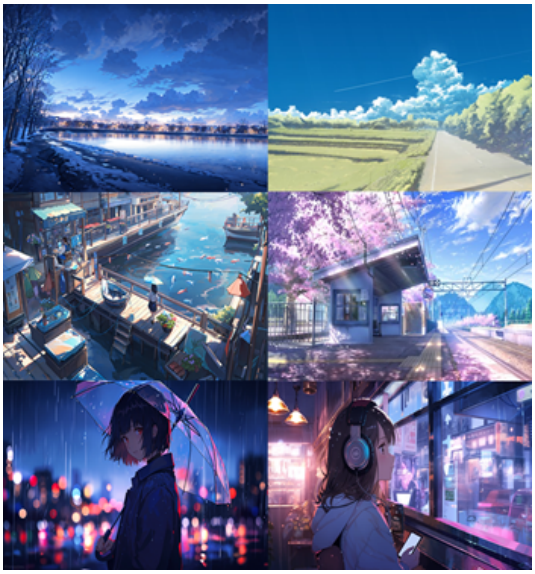

Figure 2: Samples of our dataset. The themes from top to bottom are landscapes, buildings and characters, with AI generation on the left and human works on the right.

### 2.2 AI-generated image detection

Since the explosive growth of AI-generated art, corresponding detection technology has also been continuously developed. There are many commercial [6, 7] as well as academic art detectors [8, 9]. However, relevant data sets have been lacking. Large-scale AI-generated image data sets often focus on AI-generated real-life images rather than AI-generated art, and there is an essential difference between the two: AI-generated art cannot be judged by common sense, so it is more difficult to obtain labels. In the data sets about AI-generated art, almost no internal data balance is achieved according to art styles, so that the model may judge AI and human art based on art styles rather than creative characteristics.

The work of Anna Yoo Jeong Ha et al. [10] specifically addresses this issue. With the help of the art community, they collected a total of 280 human-created artworks in 7 categories, and correspondingly generated 350 AI-generated artworks. Unfortunately, their high-quality datasets are not publicly available, so in our work we had to create our own datasets.

## 3 Methodology

### 3.1 Dataset Construction

In order to eliminate the inherent imbalance of the data and enable the model to distinguish based on the characteristics of AI and human art rather than other factors such as artistic style and tone, we followed the work of Anna Yoo Jeong Ha et al. [10] and created our own data set. Specifically, we selected 150 AI-generated artworks in each of three themes from the illustration website pixiv, namely landscapes, buildings, and characters. In order to ensure the quality of AI-generated works, they are all works within the past year. Then, we selected 450 human-generated works from the same time period with the same theme and similar style. Finally, we obtained a data set containing 900 art pictures, including 450 human artworks and AI-generated artworks in 3 categories, consistent in theme and style, which greatly eliminated the imbalance of the data. Figure 2 shows part of it.

### 3.2 Pretext training

Our approach consists of three steps: partitioning the dataset, employing self-supervised learning on an unlabeled dataset via a pretext task, and fine-tuning on a labeled dataset. For comparison, we

Table 1: Different partitions of the dataset

| Model | % Training data | % Unlabeled data | % Test data |
|---|---|---|---|
| With self-supervised learning | 15% | 70% | 15% |
| Without self-supervised learning | 15% | - | 15% |
| Without self-supervised learning, Large training set | 75% | - | 15% |

also experimented with a supervised learning variant that omits the second step above and instead performs a classification task directly on the dataset.

#### 3.2.1 Dataset separation

We first randomly split 70% of the dataset (630 images), remove their labels, and use them for self-supervised learning. The remaining 30% is used for downstream tasks, that is, the final classification task, of which 15% (135 images) is the training set and 15% (135 images) is the test set. In this way, we only used 15% of the original data set for training. To show the comparison, we later split 15% (135 images) of the original data set as the training set, and 15% (135 images) as the test set, as a supervised learning model under the same amount of labeled data. Finally, we split 75% (675 images) of the original data set as the training set and 15% (135 images) as the test set, as an example of supervised learning when there is plenty of labeled data (See Table 1).

#### 3.2.2 Loss function

After partitioning the training set, test set, and self-supervised set, we define the corresponding loss functions. For the supervised learning process that utilizes the training and test sets, we employ cross-entropy loss; for the self-supervised aspect, we apply the Triplet loss as described in [5]:

$$L = \sum_{i=1}^{N} \left[ \|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+ \tag{1}$$

where $x_i^a$ is the anchor, $x_i^p$ is the positive sample and $x_i^n$ is the negative sample, and $f(x_i^a)$, $f(x_i^p)$, $f(x_i^n)$ are the model's outputs. $\alpha$ is a hyper perameter, called the margin, that controls the difference in distance between the anchor and the positive samples, and the anchor and the negative samples. When the difference in distance between the positive and negative samples is less than $\alpha$, the loss function is positive; otherwise, it is zero.

#### 3.2.3 Pretext task

Based on the above loss function, the pretext task will be to minimize the loss function, ensuring that the model's distance between the anchor and the negative sample is not less than the distance between the model and the positive sample plus the margin. In our dataset, we create anchors, corresponding positive samples, and negative samples through data augmentation. We apply two random transformations to the original images, such as flipping, cropping, color jitter and Gaussian blurring, designating one as the anchor and the other as the positive sample. Meanwhile, a transformation of a different image is selected as the negative sample. Figure 3 presents a set of examples for the anchor, positive sample, and negative sample.

And the model $f$ in our work is an unpretrained ResNet50 network that outputs a vector of length 128, as the output for the anchor, positive sample, and negative sample i.e. $f(x_i^a)$, $f(x_i^p)$ and $f(x_i^n)$.

### 3.3 Downstream classification

Once training on the pretext task is complete, the self-supervised model classifies AI-generated and human-generated art as a downstream task, undergoing fine-tuning on the training and test sets to obtain the final classification results. In contrast, the supervised model proceeds directly to this step to obtain results.


Figure 3: Samples of the anchor, positive sample and negative sample.

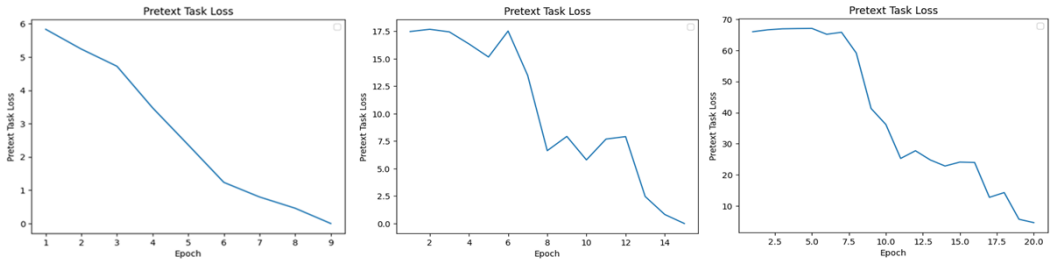


Figure 4: Loss curves of different margins. The margins from left to right are 4, 16, and 64, respectively. Notably, with a margin of 64, convergence was not achieved within 20 epochs.

# 4 Experiments

## 4.1 Pretext task

Experiments show that for a relatively small margin, the pretext task always converges to zero. Figure 4 illustrates the transformation curves of triplet loss under different margins, using the Adam optimizer at a learning rate of 1e-5.

## 4.2 Downstream classification

In the downstream classification task, our self-supervised learning model achieved better results compared to the unsupervised learning version. After setting the learning rate to 1e-5 and using the Adam optimizer for 50 epochs, the results obtained are shown in Figure 5. Additionally, for comparison, we also conducted training using the supervised learning version with a large training set. The final accuracies of each model are presented in Table 2.

It can be observed that after 20 epochs, the accuracy of the self-supervised learning model surpasses that of the non-self-supervised learning model and maintains the lead. Additionally, from the supervised learning model that uses a large training set, it is evident that after 50 epochs, the model begins to overfit, leading to an increase in test set loss. This is why we limited the training to only 50 epochs.

Table 2: Final accuracies of different models

| Model | Final accuracy |
|---|---|
| With self-supervised learning | 68.89% |
| Without self-supervised learning | 65.19% |
| Without self-supervised learning, Large training set | 71.85% |

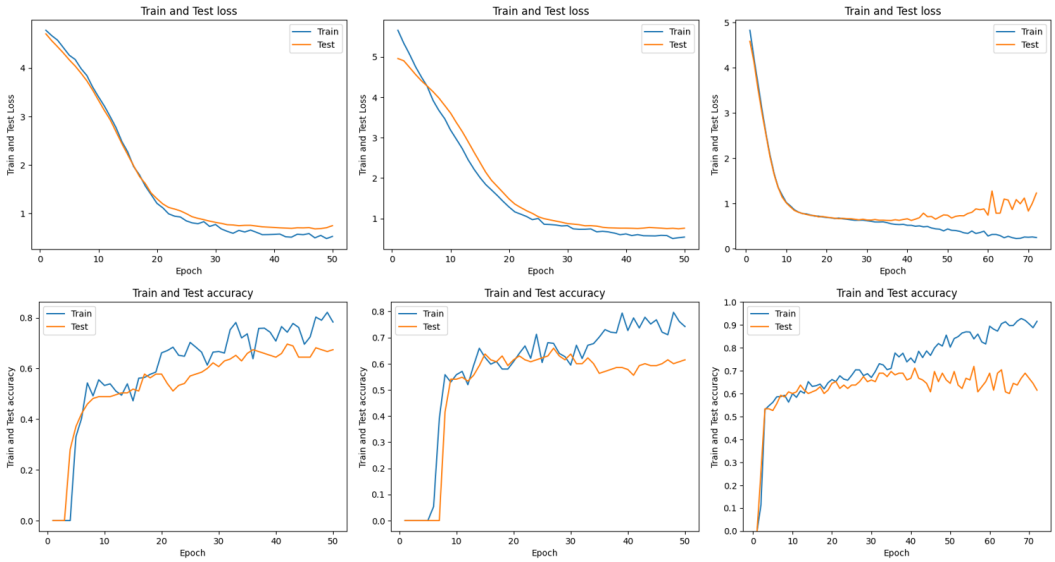


Figure 5: Loss and accuracy curves of different models. The top row shows the loss curves, while the bottom row displays the accuracy curves. From left to right, they represent with self-supervised learning, without self-supervised learning, and large training set without self-supervised learning.

## 5 Conclusions and future works

As AIGC continues to advance, distinguishing AI-generated content becomes increasingly crucial. Our work has demonstrated that self-supervised learning approaches can better differentiate between AI-generated and human-generated art when labeled data is scarce.

However, there are still issues and room for improvement in our work, which we intend to further explore and address in future work. 1. Insufficient data volume. Although self-supervised learning does not require a large amount of labeled data, our dataset was still too small, leading to overfitting on the test set. In the future, we plan to expand our dataset while ensuring class balance. 2. Employing more advanced self-supervised learning methods. In recent years, more pretext tasks and loss functions for self-supervised learning, such as MOCO [4], have been proposed. Using these methods might further improve the performance of our self-supervised model.

## References

[1] Wang, B. & Yeung, J. (2023, September 28). Chinese artists boycott big social media platform over AI-generated images. *CNN*. https://www.cnn.com/2023/09/28/tech/chinese-artists-boycott-ai-generator-intl-hnk/index.html

[2] *pixiv's policy on AI-generated work.* (n.d.). Pixiv. https://www.pixiv.net/info.php?id=8710&lang=en

[3] Sato, M. (2023, June 9). How AI art killed an indie book cover contest. *The Verge*. https://www.theverge.com/2023/6/9/23752354/ai-spfbo-cover-art-contest-midjourney-clarkesworld

[4] He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9729-9738).

[5] Schroff, F., Kalenichenko, D. & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 815-823).

[6] *AI or Not | AI Detector to Check for AI in Images & Audio.* (n.d.). https://www.aiornot.com/

[7] *AI generated Content Detection - Illuminarty - Home.* (n.d.). https://illuminarty.ai/en/

[8] Jeong, Y., Kim, D., Kim, P., Ro, Y. & Choi, J. (2021). Self-supervised gan detector. *arXiv preprint arXiv:2111.06575.*

[9] Sha, Z., Li, Z., Yu, N. & Zhang, Y. (2023, November). De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security* (pp. 3418-3432).

[10] Ha, A. Y. J., Passananti, J., Bhaskar, R., Shan, S., Southen, R., Zheng, H. & Zhao, B. Y. (2024). Organic or Diffused: Can We Distinguish Human Art from AI-generated Images?. *arXiv preprint arXiv:2402.03214.*

[11] Gidaris, S., Singh, P. & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728.*

[12] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PMLR.

[13] Noroozi, M. & Favaro, P. (2016, September). Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision* (pp. 69-84). Cham: Springer International Publishing.

[14] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4681-4690).

[15] Zeng, Y., Fu, J., Chao, H. & Guo, B. (2019). Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1486-1494).