Attribution of Animated Images: the CNN-DCT-Mask model

Zehao Zhu
ECE
University of British Columbia
Vancouver, Canada
zzh2015@ece.ubc.ca

Ali Magzari
ECE
University of British Columbia
Vancouver, Canada
amagzari@student.ubc.ca

Zhao Hong Chen
ECE
University of British Columbia
Vancouver, Canada
chenzha7@student.ubc.ca

Hanxin Feng
ECE
University of British Columbia
Vancouver, Canada
fenghx@student.ubc.ca

***Abstract*—Our purpose is to develop deep learning models to detect whether images are human-created or AI-generated, and to attribute AI-generated images with the names of their generation models. We choose animated-based images as the category of our focus due to the robust ecosystem of talented and engaged artists who relentlessly create new art pieces in this style by hand. The propensity towards sharing images online pioneered by this community provides us with a robust set of hand-drawn and AI-generated images to conduct our testing. The enthusiasm shown by this community in developing open-source models based on the latest published techniques also provides us with methods to generate novel images with which we conduct our tests. This document will explore our techniques with the proposed network CNN-DCT-Mask and findings regarding the attribution of AI-generated images and distinguishing them from those created by human artists.**

I. Introduction

As the field of deep learning advances, many new avenues of exploration become feasible and marketable. One such technology which has made waves recently in communities across the creative space is the ability to generate images based on text prompts. The first iterations of such models often used proprietary techniques to associate contextual text with output images. They often did so by combining traditional natural language descriptions with novel transformer models like GPT in the case of the DALL-E models. Oftentimes these models were trained with large scale datasets of text and image pairs scoured from the internet. As the model development advanced at a breakneck pace, the output images from these models began to become significantly higher in fidelity. Newer models which leveraged diffusion based techniques began to take over the conversation and output images became even more detailed and accurate. Today, many images generated from these models are virtually indistinguishable from those drawn by actual human artists. This proves to be a problem both commercially and ethically when it comes to determining the authenticity of an image. This also creates exciting new avenues of creativity with artistic creations being the product of an exhaustive list of prompts and descriptions.

II. Background

A. History

The concept of computers generating art dates back to the 1960s with AARON, a loosely defined set of computer programs meant to programmatically create art using a rules-based approach written by artist Harold Cohen [1]. This concept wouldn't see much additional development until the late 2010s with the popularization of deep learning techniques.

B. Generative Adversarial Networks

Generative adversarial networks or GANs are one of the first model architectures used to generate art. These networks consist of 2 main parts, a generator and a discriminator [2]. The generator takes in text-based inputs or prompts and generates data. The discriminator is used to classify whether or not the data which was generated is genuine (from the training set) or artificial (generated by the generator). This architecture then forms a feedback loop to update the generator to produce higher-quality images while also updating the discriminator to improve its classification accuracy [2].

C. Stable Diffusion

Most recently, Stable Diffusion was released as a collaboration between industry and academia. Stable diffusion relies on the latent diffusion model which is a generative model that is optimized for speed through the compression of the image space into the latent space [3]. The process of generating images starts with the variational autoencoder which compresses the image from the pixel space into the latent space.

The noise prediction engine (U-Net) picks up the latent output as well as the text prompts and predicts the noise from the latent output. The difference between the noise and the latent output is the new latent image which is subsequently decoded back into the pixel space and the result is the newly generated image [3].

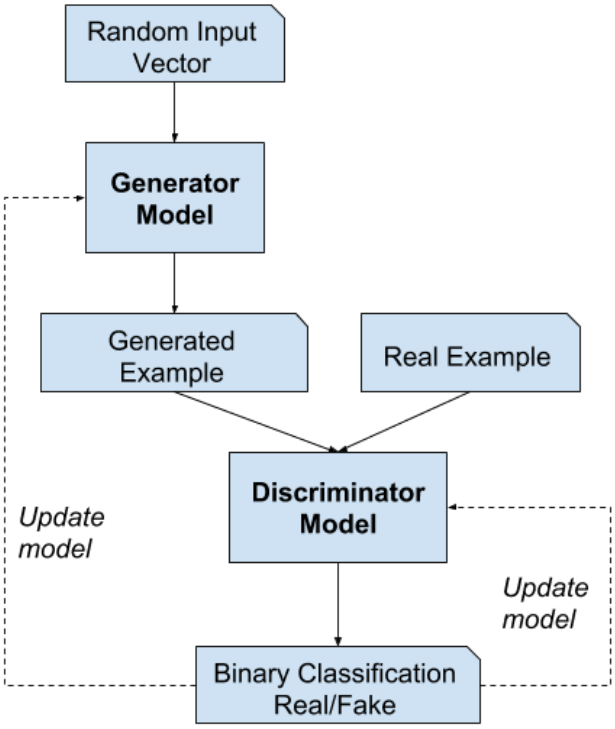

Fig. 1. Architecture diagram for a GAN

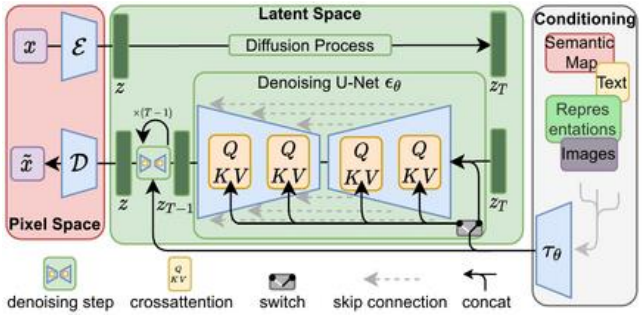

Fig. 2. Architecture diagram for stable diffusion

D. Motivations

The rapid advancement and growth of this field bring forth a number of concerns regarding the adoption of these models and also highlights the importance of being able to detect images created by generative models. Concerns have been raised about intellectual property and to whom specific images belong. Recently, the United States Copyright Office has ruled that AI-generated images cannot be protected by copyright law [4]. Academic integrity especially in the artistic fields is another area of concern for these generated images since it is quite difficult to differentiate between generated and real images. Once mature, this technology could even influence the creative job market by replacing human artists with simple input prompts and parameters. These factors make it imperative that images which are generated by modern models are able to be detected and identified.

III. METHODS

A. Convolutional Block Attention Module (CBAM)

CBAM stands for Convolutional Block Attention Module, and is an attention mechanism proposed by Woo et. Al [5]. Due to it being a lightweight and general module, CBAM can easily be incorporated in any CNN architecture. It is applied at any intermediate feature map, where attention maps are computed along the channel and spatial dimensions. The attention maps are then multiplied by the input feature to produce a refine feature (Fig. 3).

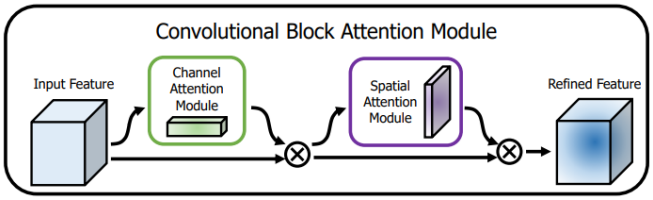

Fig. 3: CBAM Architecture

More precisely, CBAM consists of two parts: A Channel Attention Module (CAM) that and a Spatial Attention Module (SAM), as displayed in Fig. 4. CAM first tries to aggregate spatial information by utilizing both average and max-pooling, hoping that the latter infers finer channel-wise attention. Two spatial context descriptors are generated, then forwarded to a multi-layer perceptron, after which the two outputs summed in an element-wise fashion. Finally, the *Sigmoid* function is applied, and yields a channel attention map. On the other hand, SAM leverages the inter-spatial relationship of features by applying the average and max-pooling operations along the channel dimension. The resulting features are concatenated, and convolved to finally result in a spatial attention map.

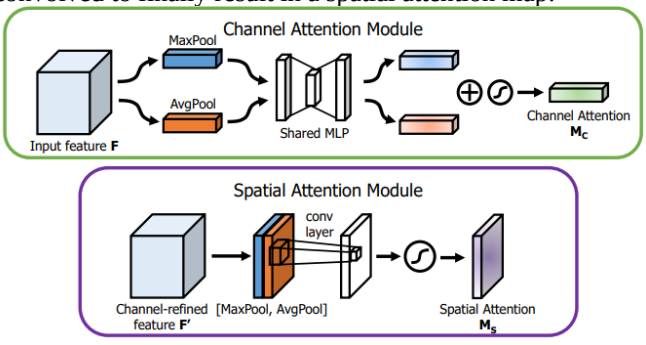

Fig. 4: CAM and SAM Architectures

B. CNN-DCT-MASK

Our proposed model, illustrated in Fig. 5, is a convolutional network (CNN) that leverages the frequency representation of the image and the concept of attention, as it incorporates the Discrete Cosine transform (DCT) and CBAM, respectively. The DCT is applied on an input image to express its pixel-signal in terms of weighted cosine functions oscillating at different frequencies. This is done in the hope to extract artifacts or a unique signature that an image generated by a particular deep learning generative model may exhibit in the frequency domain. Frank et al. [6] showed that the mean spectrum of natural images and images generated by different Generative Adversarial Networks (GANs) exhibit specific patterns and artifacts. Based on this finding, the initial step of the proposed model is to implement a DCT of the input image, after which it is forwarded through a CBAM module. These operations are followed by a short series of convolution and average-pooling. The resulting feature map is flattened and forwarded onto a linear transformation to be reduced into a vector of 5 elements, 5 being the number of classes in our classification problem (Human-created, Classic-anim-diffusion, Mo-di-diffusion, NovelAI, and Waifu-diffusion).

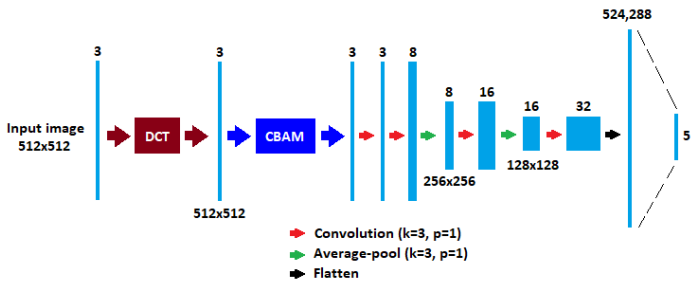

Fig. 5: CNN-DCT-MASK

C. Xception

Xception is a network that builds on top of the Inception hypothesis presented by Szegedy et al [7]. The theoretical principal driving the Inception architecture is that cross-channel correlations and spatial correlations are “sufficiently decoupled that it is preferable not to map them jointly”. François Chollet pushes this principle even further by posing the question if cross-channel correlations and spatial correlations can be mapped completely separately [8]. He first reformulates the Inception module by a large 1x1 convolution followed by 3 spatial convolutions that are applied on non-overlapping partitions of the output channels. He then generalizes this concept to a 1x1 convolution followed by a high number of towers, each separately mapping the spatial correlations of every output channel. This is the main idea behind Chollet referring to Xception as an extreme Inception module. Two differences are added beyond this generalization. First, the order consisting of implementing a 1x1 cross-channel convolution followed by a number of spatial convolutions is switched. Xception replaces this set of operations by Depthwise Separable Convolutions. A Depthwise Separable Convolution consists of first breaking the input volumes into sub-volumes by implementing Depthwise Convolutions (channel-wise spatial convolutions), then a 1x1 convolution, also known as Pointwise Convolution. The second difference is that unlike in Inception, Depthwise Seperable Convolutions do not inject non-linearity (ReLU activation function). Holding these changes as the heart of its architecture, Xception is composed of three flows: Entry, Middle (repeated 8 times,) and Exit flow, as illustrated in the figure below:

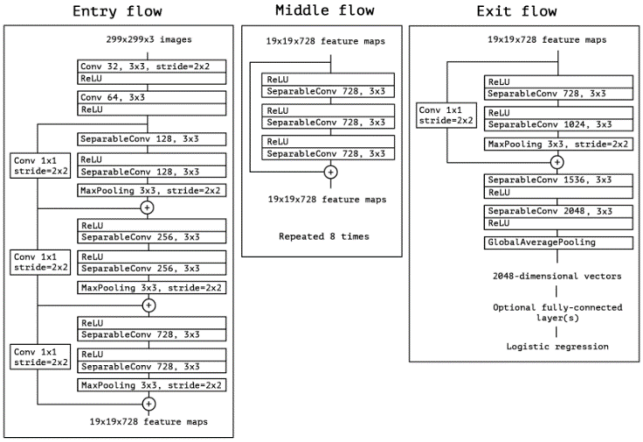

Fig. 6: Xception Architecture

D. Network Variations

Using the option of either inputting the image at the pixel or frequency level, and the option to use CBAM, 6 network variations were constructed, as detailed in the table below:

TABLE I: NETWORK VARIATIONS

Model	PIXEL/DCT	CBAM
CNN-PIXEL	PIXEL	NO
CNN-DCT	DCT	NO
CNN-DCT-MASK	DCT	YES
XCEPTION-PIXEL	PIXEL	NO
XCEPTION-DCT	DCT	NO
XCEPTION-DCT-MASK	DCT	YES

IV. Data Generation

The model requires a large dataset for training. Our task is an attribution task which means each image in the dataset has to be labeled not only AI-generated or human-created but also which model is used for the generation if it is AI-generated. The collection of AI-generated images from online resources does not have attribution of generating models for all images. The distribution of dataset from online resources is not guaranteed to be balanced. Another main problem is that the AI-generated images online may use prompts with a non-uniform distribution. Therefore, we decided to use four state-of-art models to generate our AI-generated images dataset. For human-created images, we collect them from a website called Safebooru and artworks of Disney.

The balanced training and testing datasets are significant for reaching higher accuracy of the test dataset [9]. It also makes the calculation of accuracy easier when using a balanced testing dataset. Thus, for each of five classes that our model attributes,

the dataset for training and testing should maintain the same number of samples. Randomly selected 500 images for each class are used in training, validation, and testing splitted by 8:1:1. The dataset contains 2500 images in total, 2000 images for training, 250 images for validation, and 250 images for testing, in which the number of each class is even. The training dataset, validation dataset, and testing dataset are mutually exclusive to avoid the negative effect of training bias on the testing accuracy.

A. AI-generated images

We select four state-of-art models to generate the datasets. These four models are the classic-anim-diffusion model, mo-di-diffusion model, waifu-diffusion model, and NovelAI. All of these four models are pre-trained and fine-tuned for different styles of images. Classic-anim-diffusion model generates classic Disney-style images using the specified token "classic disney style" in prompts. This model can generate images not only with human characters but also with pure nature, animal characters, or common non-human objects in Disney movies.

The mo-di-diffusion model generates modern Disney-style images using the specified token "modern disney style" in prompts. This model works similarly to the classic-anim-diffusion model but generates images with a style often seen in Disney 3-D movies.

Waifu-diffusion model and NovelAI are fine-tuned for Japanese anime-style artworks. They generate a large range of anime images from the late 20th century to current artworks. They can also generate images with non-human characters with high quality.

Then, we generate two mutually exclusive sets of AI-generated images: random AI-generated images and manually selected images.

B. Random AI-generated images

These images are generated using randomly generated prompts. The random prompts generator implemented by us contains the hundreds of popular tags collected from Safebooru. We generated 2000 images for each class and 8000 random images in total. The resolution of these images is 512 by 512 by default. Larger resolutions will require almost the same ratio of GPU memory. Resolutions lower than 512 will generate images with anomaly and bad qualities because the model suggests an over 512 resolution is required for generating.

C. Manually selected AI-generated images

Then we generated random images using the waifu-diffusion model and NoveAI and applied two-round selections on these images. The originally generated images are viewed by one person first and then the second one to filter images with bad qualities such as abnormal anatomy, blurred areas, and signatures. The Disney images using the classi-anim-diffusion model and mo-di-diffusion model are generated with certain character names that appeared in Disney artworks. Images with characters that have never been shown in Disney artworks can be detected by fans as AI-generated images easily. They are selected with a similar process. All four classes have an exclusive rate of around 80% during each selection. The dataset contains around 1000 images for each of waifu-diffusion and NovelAI and around 600 images for each of classic-anim-diffusion and mo-di-diffusion. The selected images appear in natural and common art styles.

D. Human-created images

The human-created images are collected from Safebooru and Disney artworks. They are randomly collected without manual selection. After the collection, we applied a 512 by 512 cropping on them and kept the main characters or objects. The dataset contains around 3000 human-created images in total, 2000 for the Japanese anime style and 1000 for the Disney style.

Fig. 7: Samples of random AI-generated images and manually selected AI-generated images

Fig. 7 shows some samples in random AI-generated images and manually selected AI-generated images. The manually selected images have higher probabilities that humans are not able to detect whether they are generated by AI or not.

V. PERFORMANCE EVALUATION

TABLE 2: THE COMPARISON OF DIFFERENT NETWORKS TRAINED ON DIFFERENT DATASETS

	Training Set		Accuracy on the Following Test Set	
	Random	Manual	Random	Manual
CNN-PIXEL	√		0.68	0.43
CNN-DCT	√		0.64	0.53
CNN-DCT-MASK	√		**0.86**	**0.84**
XCEPTION-PIXEL	√		0.84	0.65
XCEPTION-DCT	√		0.87	0.77
XCEPTION-DCT-MASK	√		**0.90**	0.74
CNN-PIXEL		√	0.40	0.66
CNN-DCT		√	0.55	0.78
CNN-DCT-MASK		√	0.76	0.89
XCEPTION-PIXEL		√	0.74	0.96
XCEPTION-DCT		√	0.78	**0.98**
XCEPTION-DCT-MASK		√	**0.82**	**0.89**

We have two base models: CNN and Xception. For each of the two base models, they use either the pixel domain (the original images) or the discrete cosine transform (DCT) domain (the transformed DCT images) as input. The two models using pixel-domain inputs are CNN-Pixel and Xception-Pixel. The two models using DCT-domain inputs are CNN-DCT and Xception-DCT. After we applied the attention masks on these two base models, they are trained using DCT-domain inputs. The two new models are CNN-DCT-Mask and Xception-DCT-mask. Therefore, we have six models to compare.

As shown in TABLE 2, we compared the six models trained using random AI-generated images with the six models trained using manually selected AI-generated images. Both kinds of them are trained using the same human-created images. All tests contain human-created images that are not in the training set. The testing accuracy of the testing dataset containing random AI-generated images is the baseline for performance evaluations of these models. The testing accuracy with manually selected AI-generated images is the advanced benchmark. Our task is to attribute the images that humans cannot do and the manually selected images have much higher probabilities that humans cannot even detect whether they are drawn by humans or AI. Thus, we have testing datasets with random AI-generated images as a baseline for evaluation, and testing datasets with manually selected AI-generated images as an advanced benchmark for evaluation.

TABLE 2 shows that the DCT domain works better as inputs. The models with attention masks achieve higher test accuracy than those without attention masks. However, training on manually selected images results in the low accuracy of the baseline, the test dataset with random AI-generated images. Therefore, a better choice is to train on random AI-generated images and use manually selected AI-generated images as benchmarks. Even though Xception-DCT-Mask has the highest test accuracy on random AI-generated images, it has a 7% lower accuracy on manually selected AI-generated images than CNN-DCT-Mask does. CNN-DCT-Mask has a balanced test accuracy of around 85% for both testing datasets. CNN-DCT-Mask is more stable, well-performed, and also easy to train.

TABLE 3: THE DETECTION ACCURACY OF DIFFERENT MODELS

	Training Set		Accuracy on the Following Test Set			
	Random	Manual	Random	Manual	Stable diffusion v1.4	Stable diffusion v1.5
CNN-PIXEL	√		0.86	0.81	0.94	0.94
CNN-DCT	√		0.82	0.79	0.9	0.92
CNN-DCT-MASK	√		**0.95**	**0.96**	0.95	0.95
XCEPTION-PIXEL	√		0.95	0.78	0.73	0.77
XCEPTION-DCT	√		**0.95**	**0.96**	0.97	0.96
XCEPTION-DCT-MASK	√		**0.96**	**0.95**	0.94	0.96
CNN-PIXEL		√	0.82	0.83	0.92	0.92
CNN-DCT		√	0.87	0.92	0.76	0.82
CNN-DCT-MASK		√	0.86	0.99	0.994	0.99
XCEPTION-PIXEL		√	0.88	0.97	0.33	0.42
XCEPTION-DCT		√	**0.94**	**0.99**	0.962	0.96
XCEPTION-DCT-MASK		√	**0.97**	**0.99**	0.95	0.95

As shown in TABLE 3, the models with DCT-domain inputs have excellent detection accuracy. Since the detection is a binary classification to detect whether the images are AI-generated or human-created, the accuracy is higher than attribution. The extra stable diffusion tests are included to test the generalization of our model.

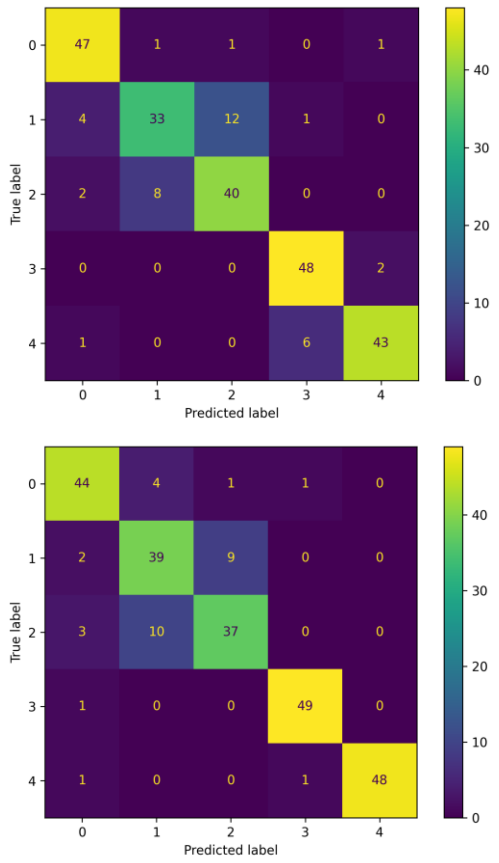

Fig. 8: The confusion matrices of DCT-CNN-Mask. The upper one is testing on the dataset with random AI-generated images and the lower one is testing on the dataset with manually selected AI-generated images.

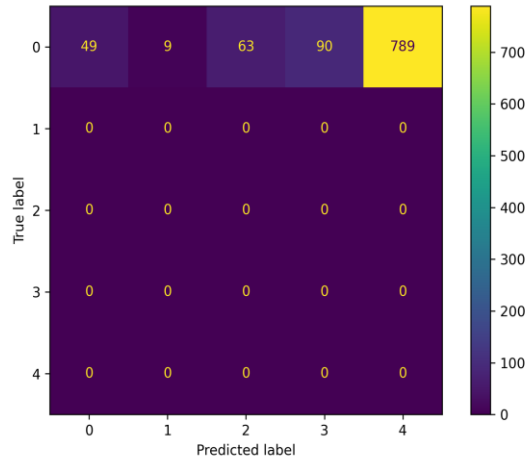

Fig. 9: The confusion matrix of testing on stable diffusion v1.4 images.

Fig. 8 presents the confusion matrices of our best model, DCT-CNN-Mask. The upper one is testing on the dataset with random AI-generated images and the lower one is testing on the dataset with manually selected AI-generated images. Class 0 is a human-created image class, and classes 1 to 4 are classic-anim-diffusion, mo-di-diffusion, NovelAI, and waifu-diffusion respectively. Both confusion matrices have similar distributions of true positives. As the data shown for class 1 and class 2, which are classic-anim-diffusion and mo-di-diffusion, class 1 and class 2 are the only two classes with relatively high rates of misclassification.

VI. Fingerprints

It is intuitive to use DCT-domain inputs for training, validation, and testing. AI models leave unique artifacts/fingerprints in the DCT domain [6]. We thus confirm that our collected images also have these fingerprints.

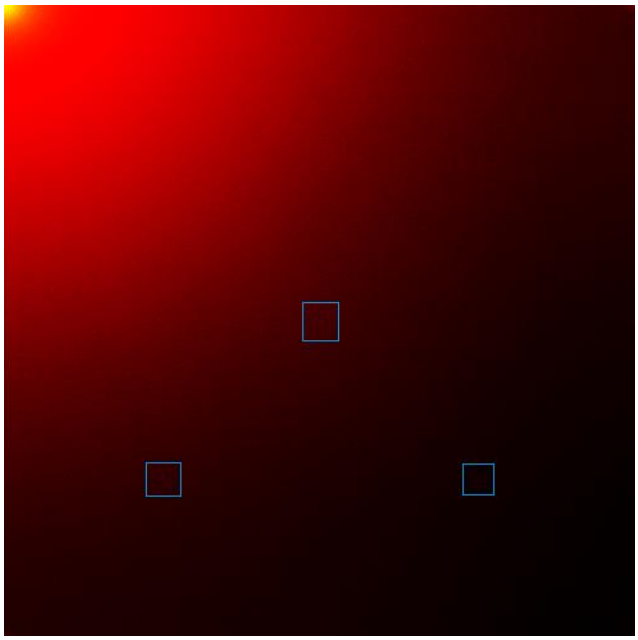

Fig. 10: Fingerprints of the classic-anim-diffusion model.

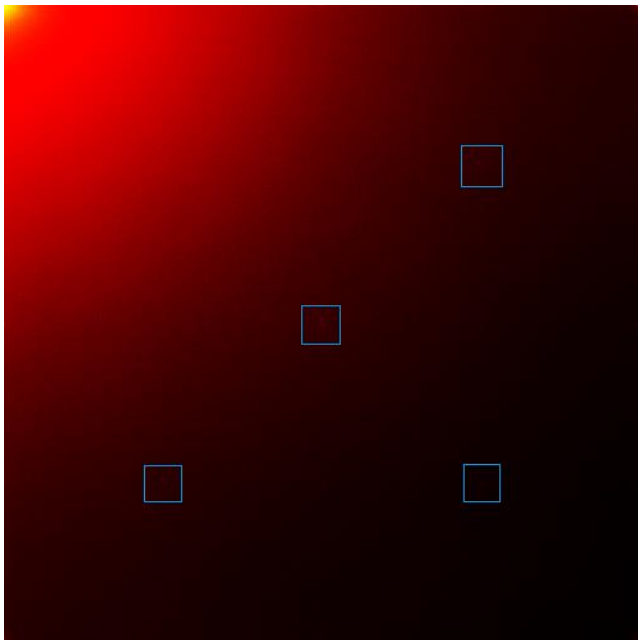

Fig. 11: Fingerprints of the mo-di-diffusion model.

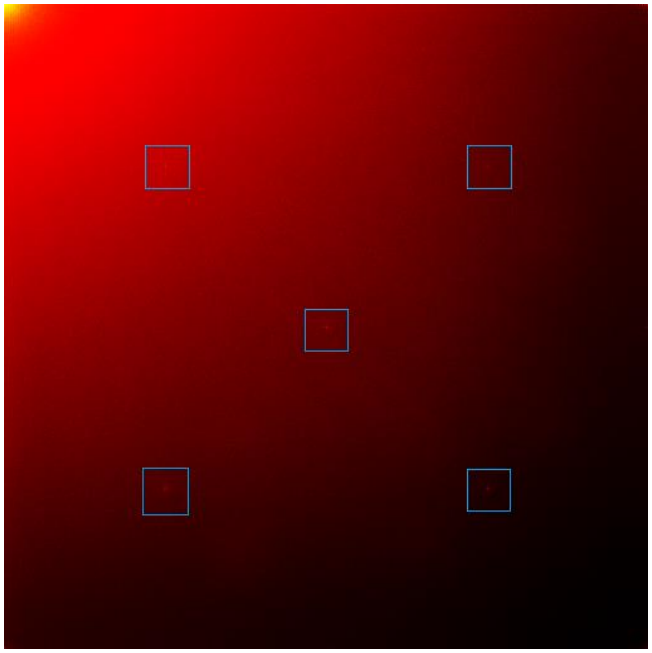

Fig. 12: Fingerprints of the NovelAI model.

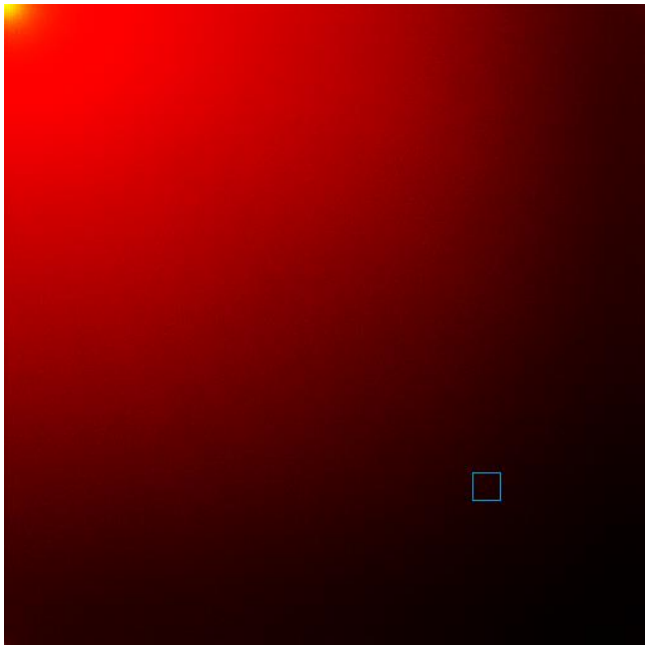

Fig. 13: Fingerprints of the waifu-diffusion model.

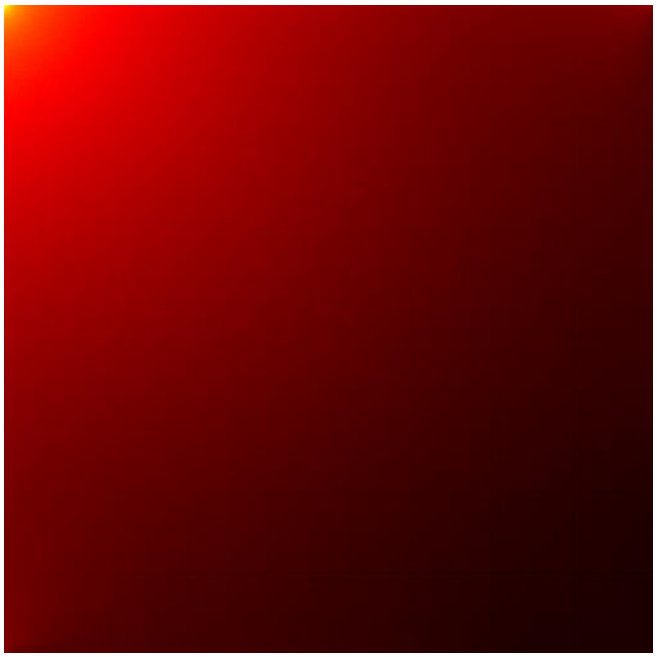

Fig. 14: No fingerprints for human-created images

Fingerprints are hard to detect because of noises. Therefore, for visualization, the calculation of the average on the DCT domain is applied. Fig. 10 to Fig. 14 are visualizations (refined with contrast values) of the fingerprints for classic-anim-diffusion, mo-di-diffusion, NovelAI, and waifu-diffusion correspondingly. Figure 8 shows that there are no fingerprints for human-created images. Different models have different characteristics of fingerprints, while similar models such as classic-anim-diffusion and mo-di-diffusion have similar characteristics of fingerprints.

VII. PROBLEMS OF FINGERPRINTS

However, many factors affect the patterns of fingerprints. Resizing the AI-generated images changes the distribution of fingerprints.

Fig. 15: Fingerprints after resizing

Fig. 15 shows the abnormal changes while resizing the NovelAI images. The number of fingerprint clusters, the distribution of clusters are changed. Enlarging the resolution also created the grid-style artifacts in the DCT domain.

As a matter of fact, we use the random prompts generator so it is hard to detect whether there are changes using specific prompts. In other words, the effects of prompts on fingerprints

or the DCT domain are still unknown which is a good research area in the future.

Moreover, the resizing of AI-generated images changes the pattern of fingerprints and even changes the number of clusters of fingerprints. This makes the prediction more difficult and less accurate after resizing. TABLE 4 shows that the accuracy decreases as the resizing frequency increases. However, resizing once keeps the fingerprints as opposed to the human-created images in the DCT domain and the detection accuracy is still high.

TABLE 4 THE TABLE OF TESTING ACCURACY OF DIFFERENT RESIZING WAYS

Resize Times	Resize way	Crop size	Accuracy type	Accuracy on the Following Test Set				
				Total	Classic	MO-DI	NovelAI	Waifu
0	X	X	Attribution	0.86	0.78	0.74	0.98	0.86
			Detection	0.95	0.96	0.94	0.98	0.98
1	512-1024	512	Attribution	0.70	0.38	0.55	0.95	0.89
			Detection	0.94	0.88	0.93	0.98	0.98
1	768-512	X	Attribution	**0.70**	0.91	0.24	0.82	0.82
			Detection	**0.97**	0.99	0.98	0.96	0.94
2	512-1024-512	X	Attribution	0.42	0.52	0.52	0.88	0.24
			Detection	0.78	0.57	0.56	0.99	0.98

VIII. SUBJECTIVE TEST

A. Define the Research Question

The first phase of the subjective test is to define the assessment criteria or metrics, which will be used to evaluate the performance of both humans and the AI model in subjective testing.

The goal of this experiment is to see whether our algorithm outperforms people in differentiating between artwork created by humans and artificial intelligence (AI).

B. Choose Appropriate Stimuli

The selection of the stimuli for the subjective testing is the next stage. To provide a variety of detection challenges, these stimuli should be changed in terms of difficulty or complexity and should be indicative of the kinds of images or objects that the AI model is intended to detect.

The images shown in the subjective test have one or more features:

- Contains different content, such as boys, girls, scenery, and animals
- From different classes, such as AI-generated classes, and artists' work
- Should be randomly selected from the dataset
- Same image size
- Different forms, 2D or 3D

As a result, we select 180 images from 9 different classes: waifu, novelai, classic-anim-diffusion, mo-di-diffusion, waifu_random, novelai_random, classic_random, modi_random, and human-created.

C. Determine Evaluation Criteria

Define the assessment standards or metrics that will be applied to judge how well humans and the AI model performed in the ostensible testing. These criteria, which might include metrics like accuracy, precision, recall, or subjective assessments of confidence or certainty in detection replies, should be precisely stated and in line with the study topic.

In the experiment, each image should be given a result based on the participants' perspective. Participants will classify all images into three groups: artist work, AI-generated, and unsure.

D. Select Human Evaluators

Establish the qualities of the human judges who will take part in the subjective exams. Think about things like their knowledge, experience, and familiarity with the job or field. To get a variety of viewpoints, it could be advantageous to involve assessors with different degrees of experience or knowledge.

Twenty volunteers from various groups—some of whom are familiar with animation visuals and others who are not—are invited to participate in the experiment. And participants also show differences in gender, age, and occupation.

E. Design Test Protocol

Create a thorough methodology for carrying out the subjective tests, including guidelines for the judges, the order of the stimuli, and the assessment standards to be applied. To reduce biases and confounding variables, take into account variables like randomization of stimuli, counterbalancing, and control conditions.

Participants will get information regarding the experiment, including the dataset's size and the nature of the images. Each participant's dataset is the same but with a different random sequence.

F. Conduct Subjective Tests

The next step is to conduct subjective tests. Give the human assessors the subjective tests in accordance with the established methodology. Gather the evaluations from the assessors, including their assessments of detection, ratings, and any other pertinent comments.

TABLE 5: Results of subjective test

Participant id	*Accuracy (%)*
1	63.33
2	77.22
3	36.67
4	46.67
5	65.00
6	48.89
7	70.56
8	32.78
9	48.33
10	72.22
11	52.78
12	92.78
13	62.22
14	57.22
15	51.11
16	66.67
17	93.33
18	82.22
19	63.33
20	40.00
Average	61.17

TABLE 5 above contains test results and the average value for the test group including 20 participants.

G. Analyze Results

Analyze the subjective test results, taking into account both human and AI model performance, using the assessment criteria. Compare and contrast the AI model's and humans' detection abilities, then interpret the results in light of the research question and testing goal.

In the experiment, the accuracy of most of the participants in distinguishing AI-generated images centered between 50% and 75%. And the average accuracy is 61%.

H. Consider Limitations

Although this project has already got satisfactory results, there is a requirement to recognize and address any restrictions or potential biases in the subjective tests at this point, such as sample size restrictions, evaluator biases, or restrictions on the stimuli or rating standards. Think about how these restrictions could affect how the results should be interpreted.

In the experiment, how to balance the trade-off on the amount between animation-familiar participants and others when selecting human evaluators is a further issue we can work on. At the same time, how to decide the dataset scale is also an issue we should consider.

IX. CONCLUSION

We use the CNN-DCT-MASK model to attribute ai-generated images. Our experimental results show that our approach reaches around 85% on the baseline tests and the advanced tests. At the same time, it performs much better than humans based on the subjective test. Therefore, CNN-DCT-MASK can be an effective choice for distinguishing AI-generated images and artists' work and attributing models of generation in a real-world scenario.

X. REFERENCES

[1] *Garcia, C. (2016, August 23). Harold Cohen and AARON—A 40-Year Collaboration. CHM. https://computerhistory.org/blog/harold-cohen-and-aaron-a-40-year-collaboration/*

[2] Jason Brownlee. (2019, June 16). A Gentle Introduction to Generative Adversarial Networks (GANs). Machine Learning Mastery. https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

[3] Stable Diffusion with Diffusers. (n.d.). Huggingface.co. https://huggingface.co/blog/stable_diffusion

[4] Recker, J. (2022, March 24). U.S. Copyright Office Rules A.I. Art Can't Be Copyrighted. Smithsonian Magazine. https://www.smithsonianmag.com/smart-news/us-copyright-office-rules-ai-art-cant-be-copyrighted-180979808/

[5] S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, "CBAM: Convolutional Block Attention Module," ECCV, 2018.

[6] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa and T. Holz, "Leveraging Frequency Analysis for Deep Fake Image Recognition," ICML, 2020.

[7] C. Szegedy, W. Liu, y. Jia, P. Sermanet, S. A. D. Reed, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going Deeper with Convolutions," 2014.

[8] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," 2016.

[9] Wei Q, Dunbrack RL Jr. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS One. 2013 Jul 9;8(7):e67863. doi: 10.1371/journal.pone.0067863. PMID: 23874456; PMCID: PMC3706434.