ICIC Express Letters
Part B: Applications
Volume **15**, Number **3**, March **2024**

ICIC International ©2024 ISSN 2185-2766
pp. **295–301**

DETECTION OF AI-GENERATED ANIME IMAGES USING DEEP LEARNING

Surya Widi Kusuma[1], Friska Natalia[1,*], Chang Seong Ko[2] and Sud Sudirman[3]

[1]Faculty of Engineering and Informatics
Universitas Multimedia Nusantara
Scientia Boulevard, Gading Serpong, Tangerang, Banten 15811, Indonesia
surya.kusuma@student.umn.ac.id; *Corresponding author: friska.natalia@umn.ac.id

[2]Department of Industrial and Management Engineering
Kyungsung University
309 Suyeong-ro, Nam-gu, Busan 48434, Korea
csko@ks.ac.kr

[3]School of Computer Science and Mathematics
Liverpool John Moores University
Liverpool, L3 3AF, United Kingdom
s.sudirman@ljmu.ac.uk

Received July 2023; accepted September 2023

Abstract. *Advances in AI allow it to be used to generate many kinds of art in the form of images, music, and even stories. AI-generated arts pose a threat to the livelihood of many artists whose income is reduced due to the decrease in demand. In this paper, we present the result of our study into the different techniques for detecting AI-generated anime images and separating them from human-artist-created images. Using transfer learning, we trained MobileNetV2 and MobileNetV3 models using 750 anime images from a dataset containing 1000 anime images generated using NovelAI and sourced from Danbooru2021 website. We tested the trained models on the other 250 images and our experiment, implemented in Python programming language and using the Keras library, reveals that both models perform well, with accuracy ranging from 96.8% to 97.2%. More importantly, our experiment also shows that both models can retrieve all AI-generated images in the test dataset (100% Precision score) but at the same time incorrectly classify a small number of human-artist-generated images as AI-generated images (Recall score of 94.3% and 95.0%). We argue that, with more work using larger-sized datasets, this approach has the potential to be used in real-world applications to filter out AI-generated anime images from online art marketplaces.*

Keywords: Deep generative models, AI-generated images, Anime images, Image classification, MobileNet, NovelAI

1. **Introduction.** Deep Generative Models (DGMs) have been increasingly used to generate many forms of art content such as images (both realistic and artistic), music, and even stories. The quality of this AI-generated art is comparable to that produced by human artists. The most famous model that has received a lot of attention recently is the OpenAI's ChatGPT [1] which uses an autoregressive language model called Generative Pre-Trained Transformer (GPT) that can be trained to understand the context and relationships between words in a sentence and produce coherent and contextually relevant text generation. The model has been shown to be able to write a literature review [2] or even pass the notoriously difficult bar examination [3]. DGMs can also generate music. Jukebox [4], for example, can imitate many artists and synthesize different music styles.

DOI: 10.24507/icicelb.15.03.295

It is a generative model for music that can create high-fidelity and diverse songs with coherence up to multiple minutes. It uses a combination of multi-scale Vector Quantized Variational Autoencoders and autoregressive Transformers to compress raw audio and model it into discrete codes. The model can also be conditioned on artist, genre, and unaligned lyrics to steer the musical and vocal style. DGMs can also be used to generate images. There are many DGM techniques for image generation such as Generative Adversarial Networks (GAN) [5] and Diffusion Models (DM) [6], with the latter having been shown to produce better results than the former [7]. An example product in this category is Imagen [8,9] by Google, a photorealistic text-to-image diffusion model that leverages large frozen language models trained only on text data as effective text encoders for image generation. The technique uses dynamic thresholding, a new diffusion sampling technique, and a new U-Net architecture variant that is simpler, converges faster, and is more memory efficient.

There are several other image generation products that have been released including Stable Diffusion [10] by StabilityAI, DALL-E2 [11] by OpenAI, and Picasso [12] by NVidia. The Stable Diffusion model from StabilityAI is released as open source so that it can be used freely by anyone and has been used in several web services including Stable Diffusion Online [13] and NovelAI [14]. This factor further encourages people to experiment with using the models to generate all kinds of images. However, there are also some ethical concerns with the use of these models. The lack of built-in internal moderation in the Stable Diffusion model for example allows it to be used to generate images of public figures or celebrities committing acts of violence [15] or pornography [16].

There is another different ethical concern with using AI to power art creation. Many people consider it a threat to the livelihood of many artists because it causes a decrease in demand for human artists and their works, thereby reducing their income and job security [17]. Many image generation models are used to produce images that mimic copyright-protected or proprietary images. These problems occur because these models can be trained using images taken from the Internet without considering the copyright and ownership of the images. The NovelAI model, for example, is trained using anime-style images drawn by many human artists; therefore, it is very likely that the images generated by this model will be similar or almost identical to the images created by the artists. This is a form of art plagiarism. If the images generated from this model are widely spread on platforms for uploading artwork, it will cause copyright issues. The Getty Image platform, one of the biggest marketplace visual arts, has started blocking AI-generate content over fears of legal challenges [18]. However, the detection is still done manually by humans [14] hence time-consuming. Therefore, there is a need for a tool that can automatically detect AI-generated images and distinguish them from genuine artistic human creations.

This paper describes the result of our study into the different techniques for detecting AI-generated images, more specifically anime images generated using the NovelAI image generation model. Our approach uses transfer learning of pre-trained MobileNetV2 and MobileNetV3 models using 750 anime images from a dataset containing 1000 anime images generated using NovelAI and sourced from Danbooru2021 website. Our experiment using the remaining 250 images produces accuracy of 96.8% and 97.2%, recall scores of 94.3% and 95.0%, and precision scores of 100% for MobileNetV2 and MobileNetV3 models, respectively. We present our proposed deep learning-based method using MobileNet to achieve the same. The paper is organized as follows: in Section 2 we present the result of our literature review, in Section 3 we describe the methodology and data that we used, in Section 4 we present and analyze experimental results before we conclude the paper in the last section.

2. Problem Statement and Literature Review.

Before we take a look at the techniques used for detecting AI-generated images, we need to understand first how these images were produced. Since the DM-based technique is considered superior to the GAN-based technique, we will concentrate our analysis on the former. DM is a generative model that converts samples from a standard Gaussian distribution into samples from an empirical data distribution [6]. This model works by adding Gaussian noise to the training data and then learning to recover the original data by reversing the process through an iterative denoising process. Compared to other generative models, DM models are easier to train and can create new data samples with better quality but take longer to generate the images due to its high inference computation cost [7]. The models operate by a process called "diffusion" that starts with an image containing only noise and iteratively and gradually improves upon the image to the point where the noise component is minimal. One such model is called Stable Diffusion (or Latent Diffusion [19]) model. This model learns the semantic and conceptual composition of an image and can generate an image from text, called *prompt*.

The process in the Stable Diffusion model starts with an encoding of the image from pixel space to a more compact space, called latent space. The diffusion takes place in this latent space and it starts by adding Gaussian noise to the latent data. This is then passed through a U-Net that has a role in predicting the latent data before the noise is added. The result of this prediction is then passed through a decoder which transforms the image back from its latent space to pixel space. Since the latent space is more compact than the pixel space, the encoding is often referred to as the image compression stage. This is when the model removes high-frequency details before it learns the semantic and conceptual composition of the image.

The Stable Diffusion models which are used in several web services such as Stable Diffusion Online [13] and NovelAI [14] were trained using the LAION-Aesthetics dataset [20] that contains 120 million of the 5 billion image and text description pairs in the larger LAION-5B dataset [21]. The NovelAI model was finetuned using 5.3 million anime-style image and text description pairs [22]. As a result, the NovelAI model can understand the characteristics of anime-style images and can generate new anime-style images. The input prompt for NovelAI can be slightly different from Stable Diffusion as it also accepts tag-based input, which is in the form of several short sentences separated by commas such as girl, wavy pink hair, maid outfit, purple eyes, short hair, smile, open mouth, ruffled blouse, red blouse, pleated skirt, blonde hair, green scarf, waving at viewer. The generated images are random but will have the characteristics included in the prompts. At this point in time, it is not possible to generate exactly the same character due to the model's inability to retain previous states that generate previous images.

Deep learning techniques have been proposed as a solution to a wide range of object detection/recognition and image classification problems in the past, including sign language recognition [23], car damage detection [24], and bird nest detection [25]. There are also several studies covering the detection of AI-generated images and videos. In [26], Liu et al. presented a study on detecting fake faces generated by GANs using global texture enhancement. The study found that the texture of fake faces is substantially different from real ones, and global texture statistics are more robust to image editing and transferable to fake faces from different GANs and datasets. Based on these findings, the researchers proposed a new architecture called Gram-Net, which leverages global texture features to improve the robustness and generalization ability in fake face detection. The model is trained with ImageNet initial weights and on 10k real and 10k fake images with a test set of the same size. The experiment, conducted in cross-GAN, in-domain, and cross-dataset settings showed that Gram-Net outperformed Co-detect and ResNet models significantly. Other studies, such as in [26], proposed a new attention-based data augmentation framework to improve the accuracy of fake face detection. The method encourages the face detector to obtain more representative forgery features by analyzing certain facial regions deeper through tracking and occluding sensitive facial regions. Similar other studies, such as in [27,28], concentrate on detecting *deepfakes*, which are the result of digital manipulation using deep learning to replace one person's likeness in an image or video with that of another. The majority of these studies focus on detecting AI-generated or AI-manipulated real-looking images. Our review of the literature found that work on detecting AI-generated art, and more specifically Japanese anime art, is still lacking. This is the main motivation behind our study to focus on the detection of AI-generated Japanese anime as opposed to the more general images.

3. **Material and Method.** Our approach to the detection of AI-generated anime from genuine artistic human-created anime is by developing a MobileNet image classification model that is trained using both AI-generated and human-artist-created anime images. The overall methodology that we adopt is depicted in Figure 1.

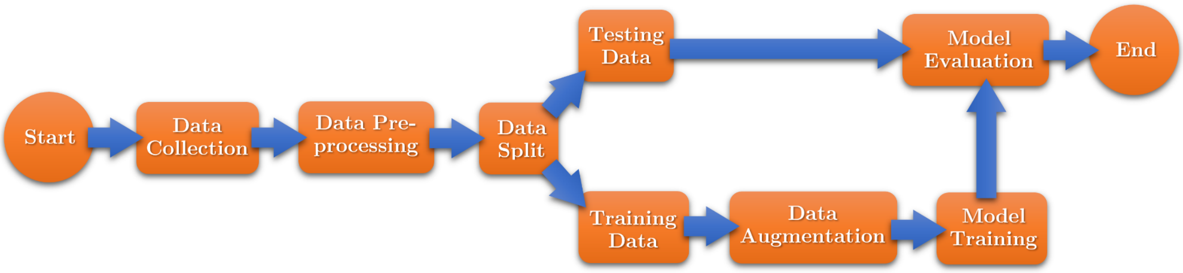

FIGURE 1. The overall methodology to detect AI-generated and human-artist-created anime images

The AI-generated images we used are generated by the NovelAI Image Generation model whereas the human-artist-created images were sourced from the Danbooru2021 dataset [29], a large-scale online crowdsourced and tagged anime illustration dataset. In total, we use 500 images from each category totaling 1000 human-artist-created and AI-generated anime images. In the interest of brevity, from this point onward we will refer to human-artist-created images and AI-generated images as their respective class names which are *human* and *AI*, respectively.

The images have different sizes, so it is necessary to equalize their dimension through cropping and resizing. The cropping stage selects important areas in the image with a variable-sized rectangular window but with a fixed length-to-width ratio. In this case, the ratio used is 1 : 1. The resizing stage is performed by either scaling up or scaling down the cropped images to 256×256. The dataset is then split into training and testing sets. The data is randomly split with a ratio of 75% for training and 25% for testing, so we ended up with 750 training images and 250 images for testing. The test set consists of 117 human images and 133 AI images. Prior to using the 750 images for training the model, we applied data augmentation. It is a process of applying transformations to the available data to create new data [24]. This process is done to increase the amount of data and increase the diversity or variety of data. In our study, we applied a number of image transformations such as flip, rotate, and zoom.

To evaluate the performance of the trained models to detect AI images from the pool of training images, we use four performance metrics namely accuracy, precision, recall, and F1-score. Accuracy is the ratio of the number of correct classification results to the total number of images in the test set. This metric measures the overall accuracy of the model. Precision is calculated as the ratio of all correctly identified AI images to all images identified by the model as AI images. This metric measures the accuracy of all the images that have been identified by the model as AI images. Recall is calculated as the ratio of correctly identified AI images to all AI images in the test dataset. This metric measures the ability of the model to detect all AI images in a dataset. Lastly, F1-score metric is the harmonic mean of precision and recall. The formulas to calculate accuracy (A), precision (P), recall (R), and F1-score (F_1) based on the experiment's True Negative (TN), True Positive (TP), False Positive (FP), and False Negative (FN) are given in Equations (1), (2), (3), and (4), respectively.

$$A = \frac{TN + TP}{TN + FP + TP + FN} \tag{1}$$

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 = 2 \times \frac{P \cdot R}{P + R} \tag{4}$$

4. **Experimental Results, Analysis, and Discussion.** We tested the approach using two versions of MobileNet architectures, namely MobileNetV2 and MobileNetV3. We adopted transfer learning approach by using pre-trained models (models with ImageNet initial weights) as opposed to models with randomly initialized weights. The experiment is implemented in Python using the Keras library. We used a learning rate of 0.001, the Root Mean Square Propagation as the optimizer function, and Binary Cross Entropy as the loss function. We trained both models using an identical number of epochs which is 100. The classification results of the trained models on the test dataset are shown as confusion matrices in Figure 2.

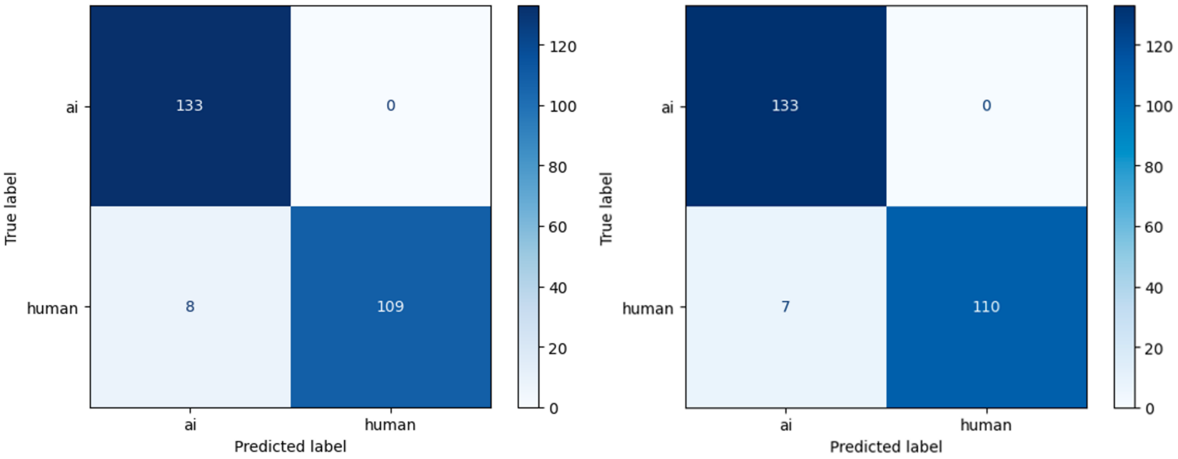

FIGURE 2. Confusion matrix of the experiment results using MobileNetV2 (left) and MobileNetV3 (right)

We found that both models manage to predict all 133 AI images correctly indicating a perfect precision for both. The models' performances slightly differ when it comes to detecting human images as MobileNetV2 managed to detect 109 compared to 110 with MobileNetV3. Both models failed to retrieve all 117 human images. This also means that the MobileNetV2 incorrectly detects 8 human images as AI images as opposed to 7 when using MobileNetV3. The accuracy, precision, recall, and F1-score of both models are summarized in Table 1.

A close inspection of the seven human images that were misclassified as AI images (shown in Figure 3) by both models reveals two rather interesting factors that might cause the misclassification. First, the human images appear to be less colorful than other human images and secondly, the images have more similar anime styles to the AI images

Table 1. The summary of the MobileNetV2 and MobileNetV3 models' performance

	Accuracy	**Precision**	**Recall**	**F1-score**
MobileNetV2	96.8%	100%	94.3%	97.1%
MobileNetV3	97.2%	100%	95.0%	97.4%

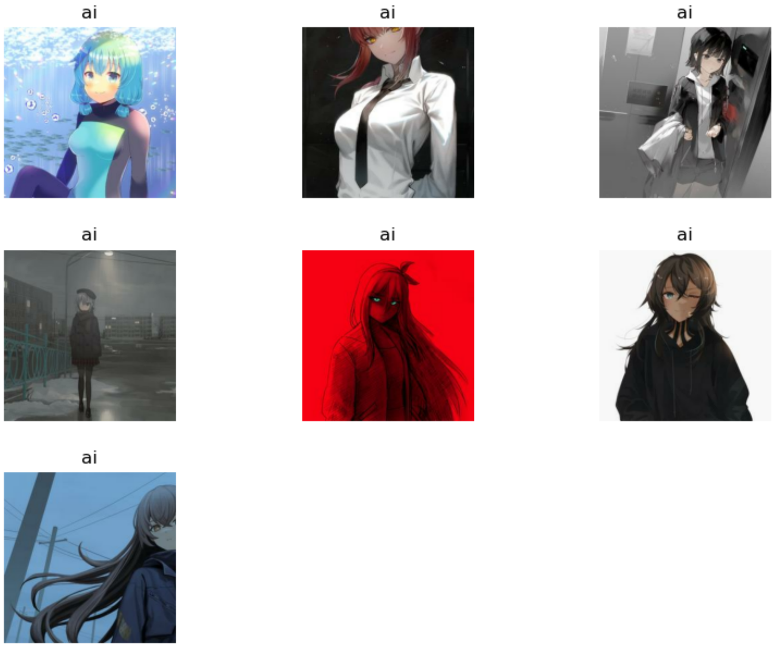

Figure 3. The seven human images that were misclassified as AI images by both models

than to other human images. We believe these two factors are the main reasons why the seven images were mistakenly identified as AI images by both models.

5. **Conclusion.** We have presented in this paper the result of our study into the different techniques for detecting AI-generated images and separating them from human-artist-created images. The driving motivation behind this study is to advance the way AI images can be quickly and automatically identified so that the threat to the livelihood of many artists whose income is reduced due to the decrease in demand, can be mitigated. It is also because research on detecting AI-generated Japanese anime art is still lacking. Using transfer learning, we trained two models based on two versions of MobileNet architectures (MobileNetV2 and MobileNetV3) on a dataset (generated using the NovelAI model and sourced from Danbooru2021 website) containing 750 images and tested the trained models on 250 images. Our experiment reveals that both models perform similarly, with MobileNetV3 being marginally superior. Both models correctly detect all 133 AI images but misclassify 7 to 8 human images. The accuracy, precision, and F1-score are 0.97, 1.00, and 0.97, respectively whereas the recall is 0.94 and 0.95, for MobileNetV2 and MobileNetV3, respectively. We plan, in the future, to further improve this approach by considering more deep learning models as well as increasing the size of the dataset. That way, we can improve the confidence that the methodology can be reliably implemented in real-world settings to detect AI-generated anime images.

REFERENCES

[1] OpenAI, *Introducing ChatGPT*, https://openai.com/blog/chatgpt, Accessed on May 5, 2023.

[2] Ö. Aydın and E. Karaarslan, OpenAI ChatGPT generated literature review: Digital twin in healthcare, *SSRN Electronic Journal*, DOI: 10.2139/ssrn.4308687, 2022.

[3] D. M. Katz, M. J. Bommarito, S. Gao and P. Arredondo, GPT-4 passes the bar exam, *SSRN Electronic Journal*, DOI: 10.2139/ssrn.4389233, 2023.

[4] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford and I. Sutskever, Jukebox: A generative model for music, *arXiv Preprint*, arXiv: 2005.00341, 2020.

[5] M. Mahyoub, S. H. Abdulhussain, F. Natalia, S. Sudirman and B. M. Mahmmod, Abstract pattern image generation using generative adversarial networks, *2023 15th International Conference on Developments in eSystems Engineering (DeSE)*, pp.172-177, 2023.

[6] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet and M. Norouzi, Palette: Image-to-image diffusion models, *ACM SIGGRAPH 2022 Conference Proceedings*, pp.1-10, 2022.

[7] P. Dhariwal and A. Nichol, Diffusion models beat GANs on image synthesis, *Adv. Neural Inf. Process. Syst.*, vol.34, pp.8780-8794, 2021.

[8] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet and M. Norouzi, Photorealistic text-to-image diffusion models with deep language understanding, *arXiv Preprint*, arXiv: 2205.11487, 2022.

[9] Google, *Imagen: Unprecedented Photorealism × Deep Level of Language Understanding*, https://imagen.research.google/, Accessed on May 5, 2023.

[10] StabilityAI, *Stable Diffusion XL*, https://stability.ai/stable-diffusion, Accessed on May 5, 2023.

[11] OpenAI, *DALL-E2*, https://openai.com/product/dall-e-2, Accessed on May 5, 2023.

[12] NVidia, *Picasso*, https://www.nvidia.com/en-us/gpu-cloud/picasso/, Accessed on May 5, 2023.

[13] StabilityAI, *Stable Diffusion Online*, https://stablediffusionweb.com/, Accessed on May 5, 2023.

[14] NovelAI, *Image Generation*, https://docs.novelai.net/, Accessed on May 5, 2023.

[15] The Verge, *Anyone Can Use This AI Art Generator – That's the Risk*, https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data, Accessed on May 5, 2023.

[16] Vice, *This AI Tool Is Being Used to Make Freaky, Machine-Generated Porn*, https://www.vice.com/en/article/xgygy4/stable-diffusion-stability-ai-nsfw-ai-generated-porn, Accessed on May 5, 2023.

[17] AIrtist Project, *Is AI Art Threatening Human Artists*, https://sites.duke.edu/airtist5/2023/02/11/is-ai-art-threatening-human-artists, Accessed on May 5, 2023.

[18] The Verge, *Getty Images Bans AI-Generated Content over Fears of Legal Challenges*, https://www.theverge.com/2022/9/21/23364696/getty-images-ai-ban-generated-artwork-illustration-copyright, Accessed on May 5, 2023.

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, High-resolution image synthesis with latent diffusion models, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.10684-10695, 2022.

[20] C. Schuhmann, *LAION-AESTHETICS*, https://laion.ai/blog/laion-aesthetics/, Accessed on May 24, 2023.

[21] R. Beaumont, *LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets*, https://laion.ai/blog/laion-5b/, Accessed on May 24, 2023.

[22] NovelAI, *The Magic behind NovelAIDiffusion*, https://blog.novelai.net/the-magic-behind-novelaidiffusion-b4797e0d27b2, Accessed on May 24, 2023.

[23] M. Mahyoub, F. Natalia, S. Sudirman and J. Mustafina, Sign language recognition using deep learning, *2023 15th International Conference on Developments in eSystems Engineering (DeSE)*, pp.184-189, 2023.

[24] M. Mahyoub, F. Natalia, S. Sudirman, P. Liatsis and A. H. J. Al-Jumaily, Data augmentation using generative adversarial networks to reduce data imbalance with application in car damage detection, *2023 15th International Conference on Developments in eSystems Engineering (DeSE)*, pp.480-485, 2023.

[25] J. Zhang, Q. Qi, H. Zhang, Q. Du, Z. Guo and Y. Tian, Detection of bird's nest on transmission lines from aerial images based on deep learning model, *International Journal of Innovative Computing, Information and Control*, vol.18, no.6, pp.1755-1768, 2022.

[26] Z. Liu, X. Qi and P. H. S. Torr, Global texture enhancement for fake face detection in the wild, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.8057-8066, 2020.

[27] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang and N. Yu, Multi-attentional deepfake detection, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.2185-2194, 2021.

[28] B. Zi, M. Chang, J. Chen, X. Ma and Y. G. Jiang, WildDeepfake: A challenging real-world dataset for deepfake detection, *Proc. of the 28th ACM Int. Conf. Multimed.*, pp.2382-2390, 2020.

[29] G. Branwen, *Danbooru2019: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset*, https://www.gwern.net/Danbooru2021, Accessed on May 24, 2023.