Deepfake Technology: Emerging Threats and Security Implications

Chuck Easttom
Georgetown University and Vanderbilt University, USA

william.easttom@Vanderbilt.Edu

Abstract. Deepfake technology is advancing rapidly and poses a range of cybersecurity concerns. Deepfakes have been used to perpetrate elaborate financial frauds. There is also the concern of deepfakes being used to influence elections. Deepfakes can fabricate statements or actions by public figures, influencing elections, public opinion, or policy decisions or simply to amplify disinformation. Adversaries can use deepfakes to spread propaganda or misinformation, destabilizing political or military scenarios. As deepfakes become more prevalent, individuals may begin to doubt authentic content, creating a "reality apathy" where distinguishing truth from fiction becomes difficult.

Keywords: LLM, AI malware, Deepfakes, AI CSAM, StyleGAN, Neural networks, Digital forensics

1. Introduction

Deepfakes are a type of synthetic media, often in the form of manipulated videos, images, or audio, which use artificial intelligence (AI) techniques to replace or mimic a person's likeness or voice. The name "deepfake" comes from "deep learning," a subset of AI that uses neural networks to analyze and generate data, and "fake," reflecting the fact that the media created is deceptive. The United States Government Accountability Office defines Deepfakes as “A deepfake is a video, photo, or audio recording that seems real but has been manipulated with AI. The underlying technology can replace faces, manipulate facial expressions, synthesize faces, and synthesize speech. Deepfakes can depict someone appearing to say or do something that they in fact never said or did.” (GAO, 2022).

Deepfake algorithms, particularly those based on deep learning, are trained using a large dataset of images or videos of the person being imitated. The AI learns the specific facial features, expressions, and voice patterns of the individual. One of the most common techniques for creating deepfakes involves GANs. A GAN consists of two neural networks:

- Generator: This network tries to create realistic fake images or videos.
- Discriminator: This network evaluates the output and attempts to determine whether it is fake or real. These two networks work together in a feedback loop, gradually improving the quality of the generated content.

In the case of deepfake videos, the AI replaces or alters the face of a person in a video with that of another person, matching facial expressions, movements, and even speech patterns. For deepfake audio, AI can synthesize a person’s voice and make them say things they never actually said. Once the face swap or voice generation is complete, further refinement techniques are used to ensure that the deepfake looks or sounds as realistic as possible, minimizing visual or auditory discrepancies.

2. Deepfake Technology

There are numerous tools available for creating deepfake images and videos. Some of these are more suited to entertainment, however there are tools that have the sophistication to create deepfakes for malicious purposes. DeepFace (https://faceswap.dev/) is a widely used open-source Python tool. It can be quite effective in creating videos and images. This tool does require some knowledge of machine learning as well as some experience with the software, but tutorials are provided.

Various tools also can be utilized to generate images. A few of the common web tools are listed here:

- https://www.craiyon.com/
- https://www.canva.com/
- https://www.vidnoz.com
- https://app.vidau.ai
- https://imagen.playground.modelslab.com/

There are also free, online tools for generating deepfake voices:

- https://www.fineshare.com/ai-voice-generator/deepfake.html
- https://www.voicebooking.com/en/free-voice-over-generator

- https://speechify.com/ai-voice-generator/
- https://app.aistudios.com

These free online tools are not the most sophisticated deepfake tools but do provide a starting point for someone wishing to learn the technology of deepfakes.

Newer architectures are also being developed specifically for high quality imagery and videos. StyleGAN is a generative adversarial network (GAN) architecture developed by NVIDIA for generating high-quality, realistic images (NVIDIA, 2024). It was introduced in 2018 and significantly improved the state of generative models by allowing for better control over the style and features of generated images. Unlike traditional GANs, which generate images by progressively upsampling random noise, StyleGAN uses a "style-based" generator (Noema, 2022). It employs a mapping network that converts a latent code (random input vector) into an intermediate latent space, known as "style space." The style space controls various features at different levels of the image generation process, such as coarse, middle, and fine details (Karras, & Aila, 2019). StyleGAN uses adaptive instance normalization to control the influence of different styles. AdaIN normalizes feature maps in the generator network and then scales them according to the learned style parameters, allowing for flexible and precise style manipulation. StyleGAN3 (released in 2021) further enhanced the architecture by addressing issues related to translation and rotation invariance, making the model even more suitable for tasks that require consistent spatial properties.

Another technology being applied to deepfakes are diffusion models (Thiel, Stroebel, & Portnoff, 2023). Diffusion models are a class of generative models used in machine learning, particularly in the field of image and audio synthesis. They have gained significant attention for their ability to generate high-quality data by modeling the process of gradually adding noise to data and then reversing this process to generate new samples. The core idea behind diffusion models is to model the process of data generation as a series of transformations. These transformations gradually add random noise to the data (diffusion process) and then learn to reverse this process (denoising process) to generate new samples from noise. The diffusion process involves adding small amounts of Gaussian noise to the original data step by step until it becomes indistinguishable from pure noise. The model then learns to reverse these steps, reconstructing the original data from the noisy versions.

Yet another technology being used to generate deepfakes is DreamBooth (Ruiz, 2023). DreamBooth is a deep learning technique used to fine-tune text-to-image models like Stable Diffusion to generate customized images of specific subjects. Developed by researchers from Google Research and Boston University, it allows for personalized content generation by incorporating unique objects, individuals, or styles into an existing generative model. DreamBooth starts with a pre-trained text-to-image diffusion model (e.g., Stable Diffusion) (Zhang, 2023). These models are already capable of generating a wide variety of images from textual prompts. The fine-tuning process involves adding new concepts to the model, allowing it to generate images of specific subjects that the original model was not explicitly trained on. DreamBooth requires a small set of reference images (usually 3-5) of the subject you want to personalize. These images are used to teach the model how the subject looks.

2.1 Creating Deepfakes

There are a range of software products available for creating deepfakes. There appears to be an inverse relationship between ease of use and sophistication. Many of the products that are very easy to use, lack sophistication.

The more robust tools require substantial effort. Creating a deepfake requires a significant amount of data, such as images, videos, or audio recordings of the target person. The quality and quantity of this training data greatly impact the realism of the final deepfake. High-quality and high-resolution content is preferred to ensure the deepfake appears as realistic as possible. The training data often includes a variety of facial expressions, head angles, lighting conditions, and speaking styles. The collected images or video frames are preprocessed to align the face and standardize size and orientation. This step ensures consistency and improves the training process. Facial landmarks (e.g., eyes, nose, mouth) are identified to better understand the structure of the face in different positions and expressions.

In video deepfakes, once the model is trained, the face of the target person is swapped onto the face of another person in a video. This process includes matching facial expressions, head positions, and eye movements to the original footage. To make the face look more realistic, techniques such as color correction, edge blending, and smoothing are applied. These techniques reduce visual artifacts and ensure that the deepfake face integrates seamlessly with the background. If the deepfake involves speech, additional steps are taken to ensure that the

mouth movements match the audio. Lip-syncing techniques are used to synchronize the mouth movements with the audio to make the video appear more convincing.

Creating a deepfake of a person's voice involves training text-to-speech or voice conversion models on audio samples from the target speaker. Techniques such as WaveNet or Tacotron can be used to generate realistic synthetic speech. After the deepfake is generated, manual editing tools (e.g., Adobe After Effects, DeepFaceLab) may be used to refine the output, correct imperfections, or add special effects.

DeepFace is a robust open-source tool available from Windows, Macintosh, or Linux from https://faceswap.dev/. It has a robust set of features, but also there is a learning curve in mastering this software. DeepFace uses a deep neural network architecture with nine layers, including convolutional layers, which are designed to process and analyze image data. Before feeding images into the neural network, DeepFace performs a 3D alignment of the face. This step involves transforming the face in the image to a frontal view by mapping it to a 3D model. This alignment helps standardize the input images and improves recognition accuracy by reducing variations due to different facial poses and angles. Deepface is a Python tool, but also has a graphical user interface, shown in figure 1.

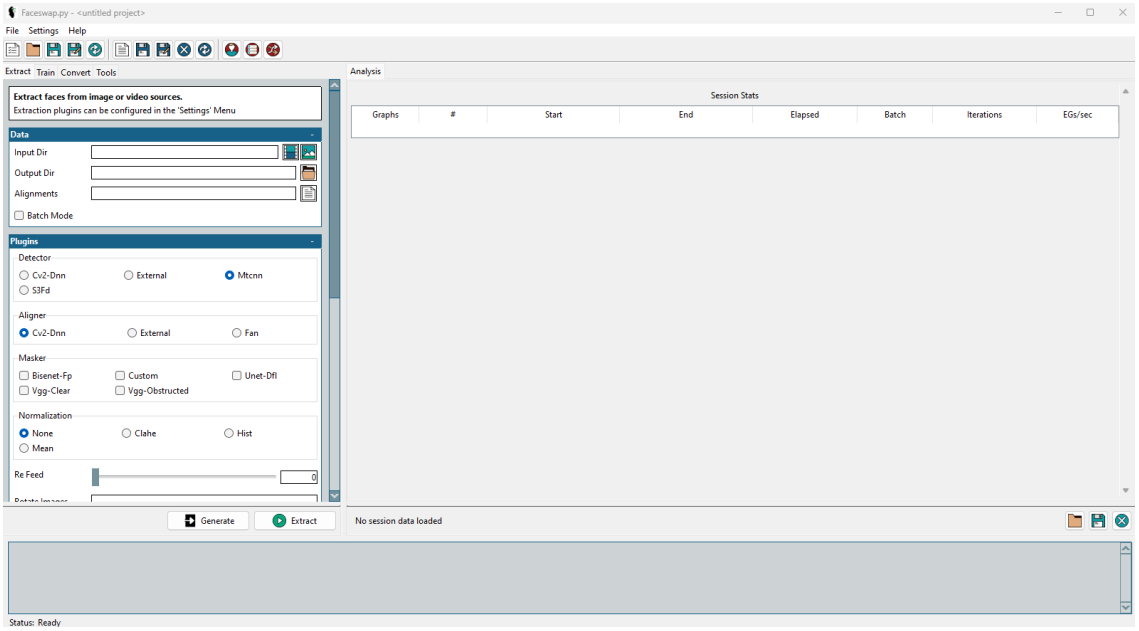

Figure 1: DeepFace

DeepFace is also available as a Python package. Python programmers can add the package just by using pip install DeepFace. The Python package does not have as many features as the full application. DeepFaceLab is a related application for designing and training deepfake models. DeepFaceLive is another related application, this one for live conferencing.

Table 1 provides a summary of widely used deepfake software:

Table 1: Deepfake Software

Name	Platform	Strengths	Weaknesses
DeepFace Lab	Windows	Highly customizable	Difficult Learning Curve
FaceSwap	Windows/Linux/MacOS	Free and beginner friendly	Runs slow unless you use a powerful computer.
DeepSwap	Online	Easy to use	Limited features
Wav2Lip	Windows/Linux	Works great for lip synching to videos.	Only useable for lip synching.
Reface	Mobile App	Easy to use	Limited features.
Synthesia	Online	High quality	Expensive
Face2Face	Windows, Linux, MacOS	High quality, works in real-time.	Difficult Learning Curve

2.2 Detecting Deepfakes

Forensically detecting deepfakes involves identifying subtle artifacts or inconsistencies left behind by the deepfake generation process. These detection techniques can be divided into several categories, each focusing on different types of analysis. Deepfakes often involve warping the face to match the target's expressions. Deepfake detection depends on searching for artifacts of such warping. Artifacts such as inconsistent lighting, unnatural eye blinking, or irregular facial movements can be detected in the generated video. The boundary between the face and the background may also appear distorted or blurred. Examining these face warping artifacts can indicate a possible deepfake. Checking audio and visual synchronization can also sometimes be used to detect deepfakes.

Skin textures in deepfake videos may lack the fine-grained details found in real images, especially in high-resolution footage. Techniques like magnifying fine skin textures can help identify artificially smoothed or irregular patterns can help identify deepfakes. Deepfake algorithms may struggle to accurately replicate realistic head poses and facial geometries, especially when the person moves their head in complex ways. Anomalies in the head's position relative to the body can indicate tampering.

Deepfake videos may exhibit unnatural motion, such as jittery or abrupt movements. These inconsistencies can be detected by analyzing motion vectors over time. Differences in texture, lighting, or other visual elements may vary from frame to frame in deepfake videos. Techniques that track these changes over time can help detect inconsistencies. In many early deepfakes, the generated subject's blinking patterns were abnormal (e.g., blinking too frequently or not enough). Analyzing eye movements and blink rates over time can help identify deepfakes. The reflection of light in a person's eyes can provide clues about the authenticity of the video. Deepfake models may not perfectly replicate the light reflections seen in real eyes, which can be a giveaway.

Machine learning models, especially CNNs, can be trained to distinguish between real and fake content by learning patterns in datasets of genuine and deepfake videos. These models can automatically detect artifacts or inconsistencies introduced by deepfake generation techniques. Deepfake generation often introduces artifacts in the frequency domain that are not visible in the spatial domain. Techniques like Fourier transform can be used to analyze frequency patterns and detect abnormalities.

Different machine learning algorithms can be useful for different deepfake detection modalities. Convolutional Neural Networks can be trained to recognize inconsistencies in facial patterns, lighting, and texture (Ahmed, et al., 2022). Recurrent Neural Networks are effective for analyzing temporal inconsistencies in videos (Albazony, 2023). Autoencoders are often able to detect artifacts left by deepfake generation techniques.

There are somewhat manual forensic methods that can also be used. One approach is analyzing the image/video in the frequency domain to spot compression artifacts and inconsistencies caused by generative models. Techniques like Photoplethysmography (PPG) measure subtle color changes in skin caused by blood flow, which are hard for deepfake algorithms to replicate (Yilmaz & Vatansever, 2024). Analyzing the image/video in the frequency domain to spot compression artifacts and inconsistencies caused by generative models (Frank, et al., 2020). Detection of blurry edges, pixel inconsistencies, or sudden jumps between frames in deepfake videos. These are the predominant forensic methods for detecting deepfakes.

While these detection techniques can be effective, deepfake generation technology is continuously evolving, leading to an ongoing arms race between deepfake creators and detection methods. Combining multiple forensic techniques often improves the accuracy of detecting deepfakes. There are some online tools designed to detect AI generated content (images, videos, text, etc.). The efficacy of these tools is not clear:

- https://deepfakedetector.ai/
- https://deepfake-detect.com/
- https://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/home_login
- https://deepfakedetector.pro/

3. Malicious Use of Deepfakes

Deepfakes are already being used in crime. February 2, 2024, a finance worker was tricked into paying out 25 million in transfers to criminals using Deepfake video. “The elaborate scam saw the worker duped into attending a video call with what he thought were several other members of staff, but all of whom were in fact deepfake recreations, Hong Kong police said at a briefing on Friday.” (Chen & Magroma, 2024)

Deepfakes are also being used for even more heinous crimes. October 24, 2023, Wired Magazine published an article regarding AI Generated Child Abuse. “AI-generated, child sexual abuse images is now underway, experts warn. Offenders are using downloadable open-source generative AI models, which can produce images, to devastating effects. The technology is being used to create hundreds of new images of children who have previously been abused. Offenders are sharing datasets of abuse images that can be used to customize AI models, and they’re starting to sell monthly subscriptions to AI-generated child sexual abuse material (CSAM).” (Burgess, 2023).

In September 2023, the Guardian reported pedophiles using generative AI to create child sexual abuse materials (CSAM). From that article “Dan Sexton, chief technology officer at the Internet Watch Foundation, told the Guardian paedophile discussion forums on the dark web were discussing matters such as which open-source models to use and how to achieve the most realistic images.” (Milmo, 2023). The National Institute of Standards and Technology has developed some basic techniques that can be used to mitigate this threat (NIST, 2023):

- Scan input prompts for abusive content.
- Assess models for their potential to generate CSAM.
- Increase options for reporting and flagging.

These steps are not a panacea for the problem of deepfake CSAM but can be an important first step.

Deepfakes can be a substantial part of Fifth-Generation Warfare. Fifth-generation warfare (5GW) refers to a form of conflict that goes beyond traditional military tactics and physical confrontations, focusing on a more diffuse and unconventional approach (Krishnan, 2024). Unlike earlier generations of warfare, which involved state-on-state conflict or well-defined groups fighting for territory, 5GW operates in the cognitive and information spaces, targeting the beliefs, perceptions, and social dynamics of adversaries. The primary battleground in 5GW is the minds of individuals, communities, and societies. This warfare aims to shape, influence, or disrupt the perception and behavior of a target population, often using misinformation, propaganda, and social engineering (Gillani, Nazir, & Pirzada, 2021). 5GW often involves decentralized or loosely organized actors rather than state-sponsored armies. It may include insurgents, terrorist groups, hacktivists, and even individuals or small networks of people who use unconventional methods to achieve their objective. The rise of the internet and advanced communication technologies has enabled the spread of 5GW tactics. Cyberattacks, hacking, data breaches, and other forms of digital disruption are common strategies.

The sophistication of deepfake technology is also leading to doubt regarding legitimate videos. In litigation against Tesla over the claims of the car being self-driving, the plaintiffs point to a 2016 video of Elon Musk at a tech conference stating that "A Model S and Model X at this point can drive autonomously with greater safety than a person. Right now." The defendants are claiming the video is a deepfake (Bond, 2023). However, in the case of this video, it has been widely available on the internet since 2016 with no claims of it being a deepfake until this litigation. Judge Dixon writing for the ABA (Dixon, 2024) discusses the issues with deepfakes being proffered as evidence. In that article, Judge Dixon cites the case of a recording of a high school principal making racists comments about students and faculty. The recording went viral and lead to threats and a need for police protection for the principle. It was later found the recording was a deepfake created by the athletic director who was being terminated.

4. National Security Concerns

National security concerns surrounding deepfakes are significant due to their potential to destabilize governments, military operations, and international relations. These concerns are rooted in the technology's ability to manipulate public perception, spread disinformation, and exploit vulnerabilities in critical systems. Deepfakes can be used to impersonate political leaders, fabricating inflammatory statements or actions that incite unrest or undermine public trust in government institutions. Foreign adversaries could deploy deepfakes to spread false narratives or discredit candidates, influencing the outcome of elections. In the recent 2024 U.S. presidential elections, deep fakes were a substantial concern. In January, an AI-generated robocall mimicking President Joe Biden's voice was used to dissuade New Hampshire Democrats from voting in the primary. The Federal Communications Commission assessed a $6 million fine against the political consultant who made it. (Dwyer & Herndon, 2024).

Adversaries can use deepfakes to mislead military personnel, such as simulating commands from high-ranking officers, potentially disrupting operations. Deepfake audio or video can mimic trusted individuals to extract sensitive information from military or intelligence personnel. Deepfakes can be used to create false evidence of

military actions or atrocities, manipulating global opinion or justifying retaliatory measures. Deepfakes can create fake statements or actions by foreign leaders, exacerbating tensions between nations and triggering conflict. Deepfakes can be used by extremist groups to produce propaganda that recruits members or incites violence, such as fake martyrdom videos or inflammatory speeches.

In the intelligence gathering and analysis domain, video evidence has long been considered the gold standard, as it has been in court proceedings. However, the widespread availability of deepfake technology should lead any analyst to question video intelligence. From the beginning of the conflict in Gaza, deepfakes have been used to sway public opinion (Klepper, 2023). Examples of AI-generated images include videos showing supposed Israeli missile strikes, or tanks rolling through ruined neighbourhoods, or families combing through rubble for survivors.

5. Conclusions

Deepfake technology is widely available and the sophistication of such technology is increasing. The ability to detect deepfakes has been outpaced by the ability to create sophisticated deepfakes. This technology has already been used to perpetrate financial fraud, generate CSAM, and product political propaganda. It is expected that this will increase in coming years. Being aware of the ease with which deepfakes can be generated is the first step in mitigating this threat. Even as better detection tools are developed, it is reasonable to expect the technology to create deepfakes will also advance. Deepfakes pose a growing threat to many areas of society, including national security.

References

Ahmed, S. R., Sonuç, E., Ahmed, M. R., & Duru, A. D. (2022, June). Analysis survey on deepfake detection and recognition with convolutional neural networks. In 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) (pp. 1-7). IEEE.

Albazony, A. A. M., Al-Wzwazy, H. A., Al-Khaleefa, A. S., Alazzawi, M. A., Almohamadi, M., & Alavi, S. E. (2023, July). DeepFake Videos Detection by Using Recurrent Neural Network (RNN). In 2023 Al-Sadiq International Conference on Communication and Information Technology (AICCIT) (pp. 103-107). IEEE.

Bond, S. (2023). "People are trying to claim real videos are deepfakes. The courts are not amused." *NPR*. Retrieved from https://www.npr.org/2023/05/08/1174132413/people-are-trying-to-claim-real-videos-are-deepfakes-the-courts-are-not-amused

Burgess, M. (2023). The AI-Generated Child Abuse Nightmare Is Here. Wired. Retrieved from https://www.wired.com/story/generative-ai-images-child-sexual-abuse/

Chen, H., Magroma, K. (2024). "Finance worker pays out $25 million after video call with deepfake 'chief financial officer'." Retrieved from https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

Dixon, H. (2024). "The "Deepfake Defense": An Evidentiary Conundrum." *ABA*. Retrieved from https://www.americanbar.org/groups/judicial/publications/judges_journal/2024/spring/deepfake-defense-evidentiary-conundrum/

Dwyer, D., Herndon, S. (2024). "AI deepfakes a top concern for election officials with voting underway." *ABC*. Retrieved from https://abcnews.go.com/Politics/ai-deepfakes-top-concern-election-officials-voting-underway/story?id=114202574

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020, November). Leveraging frequency analysis for deep fake image recognition. In International conference on machine learning (pp. 3247-3258). PMLR.

GAO (2020) "Deepfakes." *U.S. GAO*. Retrieved from https://www.gao.gov/assets/gao-20-379sp.pdf

Karras, T., Laine, S., & Aila, T. (2019). « A style-based generator architecture for generative adversarial networks." *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401-4410).

Klepper, D. (2023). "Fake babies, real horror: Deepfakes from the Gaza war increase fears about AI's power to mislead." *AP*. Retrieved from https://apnews.com/article/artificial-intelligence-hamas-israel-misinformation-ai-gaza-a1bb303b637ffbbb9cbc3aa1e000db47

Krishnan, A. (2024). "Fifth Generation Warfare: Dominating the Human Domain." Taylor & Francis.

Milmo, D. (2023). Paedophiles using open source AI to create child sexual abuse content, says watchdog. The Guardian. Retrieved from https://www.theguardian.com/society/2023/sep/12/paedophiles-using-open-source-ai-to-create-child-sexual-abuse-content-says-watchdog

NIST (2024). "Reducing the Risk of Synthetic Content: Preventing generative AI from producing child sexual abuse material." Retrieved from https://www.nist.gov/system/files/documents/2024/02/15/ID012%20-%202024-02-01%2C%20Thorn%20and%20ATIH%2C%20Comments%20on%20AI%20EO%20RFI.pdf

Noema, Y. (2022). "A brief overview of NVIDIA StyleGAN." Retrieved from https://medium.com/imagescv/a-brief-overview-of-nvidia-stylegan-61cb24ec01f5

NVIDIA (2024). "StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators". Retrieved from https://research.nvidia.com/publication/2022-05_stylegan-nada-clip-guided-domain-adaptation-image-generators

Romero Moreno, F. (2024). "Generative AI and deepfakes: a human rights approach to tackling harmful content." *International Review of Law, Computers & Technology*, 1-30.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation". *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22500-22510).

Thiel, D., Stroebel, M., & Portnoff, R. (2023). "Generative ML and CSAM: Implications and Mitigations" Retrieved from https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf

Yilmaz, B., & Vatansever, S. (2024, September). An Overview of Deepfake Video Detection Using Remote Photoplethysmography. In 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1-8). IEEE

Zhang, S. (2023). "Dreambooth-based image generation methods for improving the performance of CNN*." In 2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI)* (pp. 1181-1184). IEEE.