What is Real Anymore? An AI/ML Image Dataset Using Authenticity Validation and Traceable Origins for Every Data Instance

Andrew McDonald
Computing
East Tennessee State University
Johnson City, United States
mcdonaldai@etsu.edu

***Abstract*—This project addresses the increasing challenge of detecting AI-generated images by creating a novel dataset titled "What Is Real Anymore?" (WIRA). WIRA comprises two subsets: the first includes over 2000 images, validated as authentically real by a set criterion and sourced from photographs on Flickr. The second subset consists of hyper-realistic AI-generated counterparts for each validated Flickr image, aggregated through the Leonardo.AI commercial API. All Flickr-validated images in WIRA are credited to their respective photographers and retain their associated rights. Commercial use of this dataset requires permission from the photographers or adherence to the copyright laws of each validated Flickr image used. This document details the rationale for image authentication, image categories, the motive for category selection, authenticity validation criterion, methodology for the creation of the dataset, the computational resources used, a review of included and excluded decision records, and potential enhancements to expand WIRA.**

***Index Terms*—AI-generated, AI-generated image detection, image dataset, image classification**

I. Introduction

In recent years, the rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) technologies has led to the proliferation of AI-generated content across various domains, such as text, images, and videos. While AI-generated content has the potential to revolutionize content creation and improve efficiency, it also poses significant challenges in terms of authenticity, trustworthiness, and potential misuse. The ability to distinguish between human-generated and AI-generated content has become increasingly important to maintain the integrity of information and prevent the spread of misinformation. Thankfully, researchers have tried to tackle this problem of detecting AI-generated content with trained AI/ML models [1]-[10]. All the trusted datasets used to train their models [9]-[36], however, do not validate or authenticate any of the training images. Due to the recent advancements of AI-generated content and unless an image's origin can be confirmed, it is impossible to argue whether an image is truly real or not. This unfortunately applies to all the images provided in the datasets mentioned, as there are not any validation methods or criteria used, other than web-scraping, to determine if these images are authentically real. This report aims to address the importance and explanation of the creation of a dataset consisting of authentically real and validated photographs along with AI-generated doppelgangers. The dataset created will serve as a precursor to future datasets that ensure each data instance representing an authentically real image is verified by its origin first before aggregating. To illustrate the complexity of distinguishing between authentically real and AI-generated images, Figure 1 proposes a challenge to identify which images are real. This visual exercise emphasizes the growing difficulty of human perception alone in validating image authenticity, further underscoring the importance of datasets with rigorous validation criteria for real images such as WIRA.

II. Rationale For Image Athentication

Many of the popular and otherwise trusted datasets used for the detection of AI-generated images such as LAION-400M [13], LSUN [15], and CIFAKE [19] were created over a decade from the writing of this document. These datasets contain subsets of images labeled as real. The curators of these datasets, however, did not employ any validation techniques that ensured the web-scraped images used were authentically real. During the period that these datasets were curated, AI-generated content contaminating real image aggregation within search engines was not an enormous problem as it is today, if even a problem at all. Research regarding the issue of detecting AI-generated images has been conducted only within the last decade, with an increase of related studies within the last 2 years [1]-[5], [7]-[9], [11], [12] due to the enormous performance gains of AI-generative image models. This performance gain is so impressive that now, most humans, even those with a trained-eye, may easily be deceived by the realism of AI-generated images. More now than ever before, AI-generated images are contaminating search engines, causing real images to be interspersed with artificial content, making it harder to find authentic visuals on real-life topics or things. If a dataset was curated today to detect AI-generated content by only web-scraping without any validation criteria, it is guaranteed any set of real-labeled images may be contaminated with AI-generated content. Due to the increasing photo-realism of the AI images,

Fig. 1. The challenge in identifying which images are authentically real and which are AI-generated from these shuffled pairs underscores the necessity of robust authenticity validation methods in datasets like WIRA.

human eye validation is becoming less effective to separate what is real and what is not.

Most researchers, therefore, depend on the foundation of established datasets such as the ones mentioned previously. One could argue, however, that the real images used in these datasets are not "real" since not one of them used any validation criteria to determine the origin of each data-instance labeled real. A counter argument against this, however, can be that AI-generated content did not start proliferating sources of the web-scraped images from these datasets during the time of their aggregation. While this counterargument is plausible, it remains unprovable since no validation criteria were applied to confirm whether the images scraped in the past were genuinely real or computer altered. Some of the first AI-generated photos can be traced back to over a decade from the writing of this report. Even so, one could argue that the timestamps in the image's metadata were forged. If these datasets of 'real' images lacked records of the origin and authenticity validation, it is now impossible to confirm the true authenticity of each image, leaving room for perpetual debate over their genuineness.

III. Image Categories

A. Landscapes and Environments

This hub of image subcategories conveys the beauty of the natural world exploring stunning landforms, ecosystems, and biomes. From vast mountain ranges to intricate forest ecosystems, each subcategory captures unique aspects of Earth's landscapes and environments, offering a comprehensive view of nature's complexity and unique patterns. Figure 2 shows the full tree of these four main categories and all nested leaf image subcategories within WIRA.

1) Cities: Offers exploration of urban life across the globe, featuring a range of subcategories dedicated to cities from every corner of the world. From towering skylines to bustling streets and iconic architecture, each subcategory provides a glimpse into the unique character and energy of different metropolitan landscapes.

2) Coastlines: Showcases the intersections of land and sea, with subcategories highlighting diverse coastal landscapes from around the world. From rugged cliffs and sandy beaches to tranquil bays, each collection captures the unique beauty and ecological richness.

3) Deserts: Delves into desert landscapes, featuring subcategories that explore arid regions across the globe. From sweeping sand dunes to rocky plateaus and resilient flora, each collection reveals the unique textures, colors, and ecosystems that define environments.

4) Forests: Shows lush and diverse forests worldwide, with subcategories showcasing everything from rainforests to serene temperate woodlands. Each collection highlights rich layered textures that define these green environments.

5) Mountains: Captures many mountain landscapes, featuring subcategories that span a range of peaks, valleys, and rugged terrains from around the world. From snow-capped summits to rolling alpine meadows and dramatic cliffs, each collection showcases the scale of these elevated landscapes.

B. Life and Portraits

This hub of image subcategories captures people in their everyday lives across cultures and environments. From portraits to candid moments, each collection reveals unique stories that make up human life, offering a rich number of faces and traditions around the world.

1) Adults: Highlights adult life, showcasing individuals from diverse backgrounds and cultures.

2) Children and Adults: Captures connections between children and adults, portraying moments of mentorship and family experiences.

3) Children: Highlights adolescent life across different cultures and settings around the world.

4) Culture: Explores rich culture, traditions, rituals, attire, and celebrations from unique communities around the world.

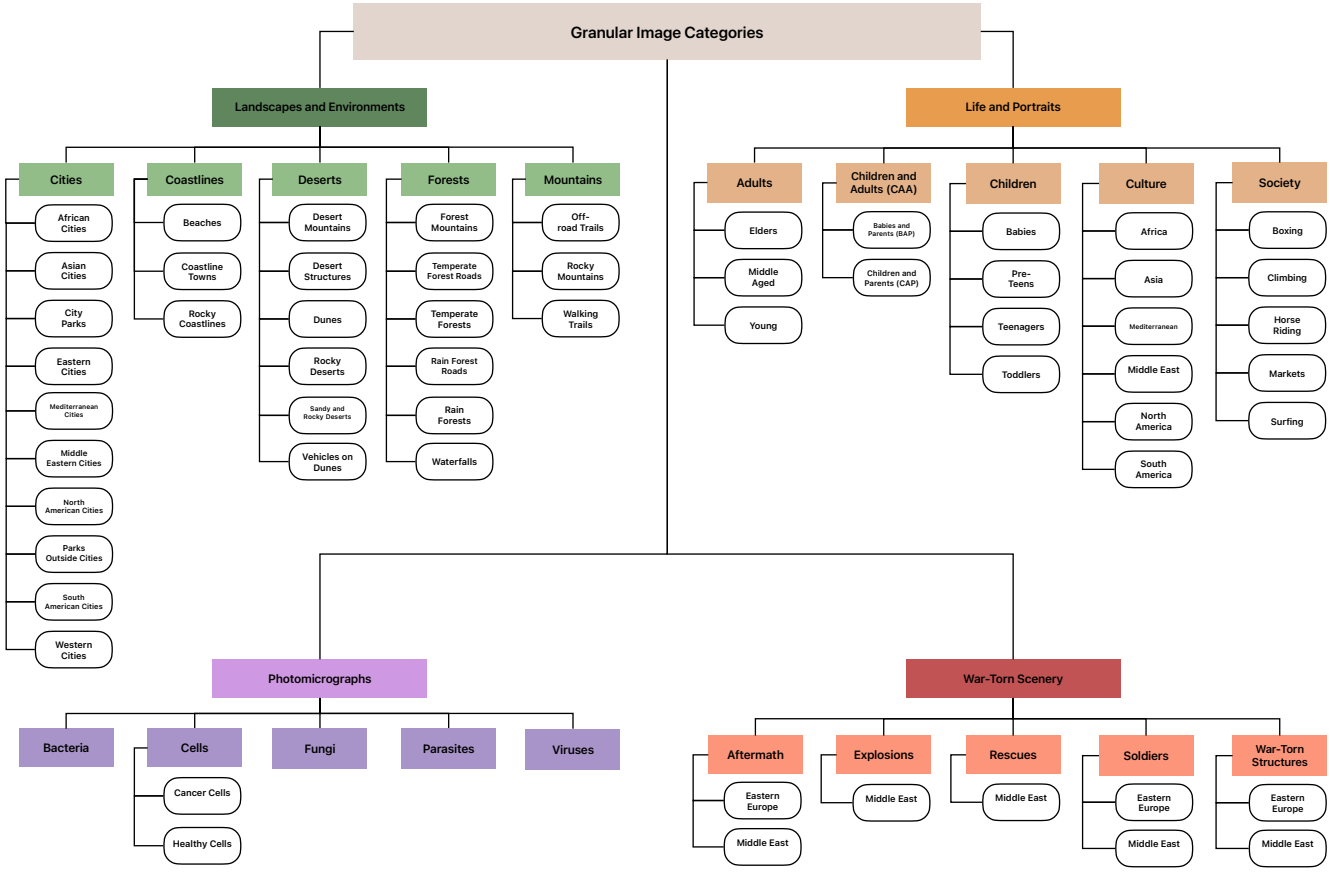

Fig. 2. Granular image categories for WIRA, showcasing the main hubs and their detailed subcategories.

5) Society: Reflects the structure of communities, capturing scenes of daily life, social interactions, and activities that illustrate how people live, work, and connect.

C. Photomicrographs

This hub of image subcategories delves into microorganisms. From detailed views of cellular organisms to intricate patterns in microscopic matter, each collection unveils otherwise hidden complexity.

1) Bacteria: Reveals the intricate forms and structures of bacteria.

2) Cells: Contains the patterns of healthy plant cell life as well as cancerous cells and the patterns they form.

3) Fungi: Captures the diversity of fungi at a microscopic level, showing unique structures of spores, hyphae, and fungal networks.

4) Parasites: Focuses on the structures of various parasitic organisms.

5) Viruses: Explores the varied structures of viruses, showcasing unique shapes for infecting host cells.

D. War-Torn Scenery

Offers an unflinching look at the devastating impact of war on societies, featuring subcategories that capture the harsh realities of conflict. These images confront viewers with graphic scenes of destruction, loss, and the human suffering that war leaves in its wake, providing a visceral portrayal of the profound toll that conflict exacts on people and places.

1) Aftermath: This subcategory captures remnants of conflict, illustrating the devastation it leaves behind.

2) Explosions: Focuses on the intense, destructive power of explosions, capturing their smoke clouds and fire.

3) Rescues: Highlights moments of bravery and compassion amidst chaos, capturing scenes of people helping others to safety, providing aid, and showing resilience.

4) Soldiers: Portrays the experiences of soldiers in various contexts, from moments of intense action to quieter scenes.

5) War-Torn Structures: Shows buildings and infrastructure bearing the destruction inflicted upon once-thriving structures in war zones.

IV. Motive For Image Category Selection

The selection of the four main image categories in WIRA; Landscapes and Environments, Life and Portraits, Photomicrographs, and War-Torn Scenery was driven by their potential for real-life misuse, particularly through the manipulation of AI-generated imagery. Each category represents a specific domain

where the realistic reproduction of false images could have significant adverse impacts.

A. Landscapes and Environments

With advances in AI, there is a risk of creating hyper-realistic but non-existent locations. Without verification, such synthetic images could deceive individuals, leading them to believe in the existence of fabricated places, potentially endangering lives if the information is used maliciously. This deception poses risks, including the potential for exploitation, manipulation, or even endangering individuals who are led to pursue non-existent destinations. WIRA's validation criterion ensures that each image in this category corresponds to an actual location, helping to prevent such misuse by allowing models to recognize patterns of authentically real places as well as find patterns in AI-generated versions of them.

B. Life and Portraits

The misuse of AI to generate non-existent individuals or to portray real people in uncharacteristic actions can create false narratives, impacting public trust. Such synthetic images could also contribute to identity manipulation and deception in social and political arenas. Furthermore, the ability of AI to fabricate human faces and situations could mislead viewers, making them believe in the existence of these fabricated entities. By validating every portrait or life scene included, WIRA helps uphold the integrity of personal identities and societal narratives for image classification models.

C. Photomicrographs

The public trust in scientific imagery is vulnerable to exploitation, as AI-generated images of non-existent pathogens could incite unnecessary fear or panic. This category emphasizes the importance of image authenticity to prevent such misinformation in scientific and medical fields. It is important to note that WIRA's validation criteria do not apply to the Photomicrographs category, as will be later detailed in section VII within the Photomicrograph Validation and Traceable Origins subheading.

D. War-Torn Scenery

In war-related contexts, synthetic images could be weaponized to mislead the public. AI could fabricate scenes of devastation or, conversely, mask ongoing conflict. This category is critical in identifying accurate conflict reporting and helps support humanitarian accountability. Adversarial AI could generate graphic war scenes to falsely depict conflict or suppress real events by portraying peaceful settings in active war zones. Authenticity validation within WIRA ensures that images genuinely reflect actual situations, preserving accurate historical records and protecting the public from manipulated portrayals of conflict when discerning what is real and what is not.

E. Conclusive Statements

These categories were chosen for their potential to safeguard society against the misuse of hyper-realistic AI-generated images in scenarios that directly impact public trust and safety. Through WIRA, the goal is to build a dataset that strengthens AI's capability to recognize genuine visual content along with hyper-realistic generated fabrications, thereby reinforcing the integrity of digital imagery in an era of increasing visual manipulation. WIRA's approach of aggregating authentically real images by a set validation criterion for these categories not only strengthens AI's ability to differentiate between real and synthetic content but also holds the digital world to a higher ethical standard for the validation of authentically real content.

V. Authenticity Validation Criterion

A comprehensive validation criterion was applied to ensure authenticity in each Flickr image used in the WIRA dataset. This multi-step process rigorously verifies each image's source, creator information, equipment used, and metadata. Metadata extraction for analytical purposes was extracted using ExifTool by Phil Harvey (version 12.96), available at https://exiftool.org. Through this process, any image that fails to meet verification standards is excluded from the dataset. The following outlines the steps in sequential order taken to confirm the authenticity of each image.

A. Initial Download and Metadata Collection

The image is initially downloaded using Flickr's API and its associated metadata is recorded by using ExifTool.

B. Creator Verification

The original creator of each image is identified using the creator's Flickr URL.

1) Creator Information Unavailable: If creator details are unavailable, the image is discarded, and the process resumes with the next image in sequence.

2) Creator Information Available: When creator information is verified, the origin URL is documented, and the image proceeds to the next stage of validation.

C. Ownership Validation

The metadata obtained from the Flickr API is reviewed to confirm that the identified creator is the legitimate owner of the image as indicated by the image's origin URL.

1) Unverifiable Image Owner: If ownership cannot be verified, the image is discarded, and the process resumes with the next image in sequence.

2) Image Ownership Verified: When ownership is confirmed, the creator's details are logged, and the image proceeds to the next stage of validation.

D. Camera Model Verification

The image's origin URL is reviewed to confirm the inclusion of the camera model used to capture the image, a critical indicator of authenticity.

1) Absent Camera Model Information: If camera model details are missing, the image is discarded, and the process resumes with the next image in sequence.

2) Available Camera Model Information: If the camera model is present, additional verification is performed through Flickr's camera database.

a) Unauthenticated Flickr Camera Model: If the camera model cannot be authenticated, the image is discarded, and the process resumes with the next image in sequence.

b) Authenticated Flickr Camera Model: When the camera model is validated, the information is recorded, and the image proceeds to the next stage of validation.

E. Image Similarity Comparison

The downloaded image is compared with the version on the Flickr origin page using a Structural Similarity Index Measure (SSIM) to confirm that no alterations have been made and to verify the photographer's ownership of the image associated with the URL.

1) Below 90% SSIM Scoring: If the SSIM score is below 90%, the image is discarded, and the process resumes with the next image in sequence.

2) Above 90% SSIM Scoring: A score above 90% indicates a strong match, allowing the image to pass validation and proceed to the final step.

F. Final Approval and Storage

After meeting all preceding criteria, the image is designated as authentically validated and stored. This process is repeated for each downloaded image until the dataset's required thresholds are achieved.

G. Conclusive Statements

Through the application of this comprehensive validation criterion, WIRA ensures that only authentically real images with verified origins are included in the dataset, distinguishing it from the previously mentioned existing datasets lacking any image verification protocols. Additionally, all image origins, metadata, records of images that passed or failed validation, and information on photographers who met or did not meet the criterion are meticulously documented for transparent analytical review. This analytical data is further detailed within section VI. Together, this rigorous approach and thorough analytical record-keeping enhances the dataset's reliability and its effectiveness in training AI/ML models for the purpose of accurately detecting hyper-realistic AI-generated content from reality.

VI. Methodology For Dataset Creation

The WIRA dataset[1] was developed through a comprehensive three-part application, the Real-To-AI Pipeline[2] is designed to aggregate authentic real images from Flickr as well as AI-generated images from Leonardo.AI's commercial API. This pipeline enables users to select models for AI-generated images along with customizable hyperparameters, offering a scalable and versatile tool for future dataset construction. Beyond simple aggregation, it ensures the authenticity of real images using the specific criterion detailed within this report. Outlined below are the three main components of WIRA's construction: the Real Image Scraper, the AI Image Captioner, and the Leonardo Image Generator. To provide an overview before discussing each part in detail: First, the Real Image Scraper collects images from Flickr, applying the previously outlined validation criterion to ensure authenticity for each image scraped. Second, the AI Image Captioner uses the open-source llama-3-vision-alpha-hf model to generate captions for each validated image, adding descriptive context. Third, these captions, along with their corresponding validated images, are forwarded to the Leonardo.AI commercial API with specified model information and custom parameters. The API then generates a hyper-realistic AI counterpart for each image-caption pair.

A. Real Image Scraper

The Real Image Scraper retrieves authentically validated images from Flickr through a structured multi-step process. First, it queries the Flickr API using customizable search parameters, such as keywords, tags, sorting preferences, and media types. Next, the image processing phase begins, checking each image to ensure it is not a duplicate. Following this, each image's origin and original metadata are documented using ExifTool, providing a traceable history for every image collected. Once the origins are recorded, the previously detailed validation criteria are applied, and if the image passes, it is saved for use.

1) Querying the Flickr API: The Image Scraper Curl class is instantiated, directing all images that meet the specified criteria to be saved in the "GranularImageCategories" directory, with additional subdirectories organized based on the Flickr query parameters. To manage the scraping process, image thresholds are set to automatically stop the scraper upon reaching the desired number of validated images. A query is then created using the Flickr API's photo search method, incorporating the API key, query terms, and relevant tags within the HTTP headers. Additional parameters, such as sorting, safe search, media types, and extra information flags, are specified to tailor the search results. The scraper also keeps track of its progress by maintaining the count of images requested from the Flickr API. Once images are returned, the owner's details and image URL are recorded, establishing both the origin and creator information needed to proceed with the image validation process.

2) Duplication Check: Before an image is downloaded, it must first pass the Real-To-AI-Pipeline's duplicate image check and the previously described validation criteria. For each image response from the Flickr API, the Python imagehash library calculates four types of hashes. These are the Average, Difference, Perceptual, and Wavelet hash calculations. These hashes are then cross-checked against their respective hash logs. If no match is found, the hashes are stored in their

[1]https://github.com/McDonaldAndrew-ETSU/WIRA.git

[2]https://github.com/McDonaldAndrew-ETSU/Real-To-AI-Pipeline.git

respective logs to prevent duplicate downloads and to avoid reprocessing the image through the intensive authenticity validation criterion again.

3) Traceable Origins: For each new image encountered by the Real-To-AI-Pipeline, all related information, including metadata with origins like URL and owner URL (used in Part 4: Authenticity Validation), as well as camera details if available, is saved to a directory. A manifest is created to facilitate transparent analysis, allowing for the identification of each unvalidated image. Each image's repository path is mapped to a JSON-formatted block containing the original metadata extracted from ExifTool, organized in a "Scraped Image Manifest" file. Additionally, the repository path is linked to the image's origin URL in an "All Links Checked" file. Together, these files provide a complete traceability record of each image's origin. If the image is unique (with no duplicate hashes found), its traceable origins are documented, and it is then ready for the authenticity validation process.

4) Authenticity Validation: This step is guided by the detailed criteria outlined in section V. This involves using Flickr's API to obtain initial attribute values for each image, followed by verification directly on Flickr's website. Python Selenium is employed to access and confirm these values on the Flickr page, ensuring the accuracy and authenticity of the data received from the API.

a) Validate Flickr Creator with Flickr Image: The validation process begins with instantiating a Validator class, which creates a headless Selenium instance. The primary driver for Selenium is the Microsoft Edge driver. The Validator class uses specific attributes from the current image's Flickr API response such as the image's Flickr ID, secret, user ID, and direct image URL (used for download if validation succeeds) to focus on authenticating that image on Flickr. The creator's Flickr page is located using the user ID obtained from the Flickr API. If the API provides the creator's URL, the process continues to the next step; if not, the process restarts from Step 1 with the next available image.

b) Validate Creator on Flickr Creator URL: Second, the validator instance accesses and verifies the creator's information on the Flickr page using Selenium. If the creator's name matches the account name displayed on the account page where the Flickr image is hosted, the validation process continues. If no match is found, the process restarts from Step 1 with the next available image.

c) Validate Camera from Flickr Image Metadata: Third, the validator retrieves the camera information from the EXIF data in the Flickr image's API response. If the camera attribute is present, it is recorded, and the process proceeds to the next step. If the camera attribute is missing, the creator is flagged in a "Watchlist" file, which records creators who failed validation along with their image URLs and reasons for failure. In this case, the reason, "No camera listed within image metadata," is appended after the image URL, separated by " - " to ensure the URL remains intact. This entry is added to a list of failed image URL-reason pairs associated with the specific creator. The image URL is also mapped to a local path in a "Failed" file, as the initial image is downloaded and saved separately from WIRA. The process then restarts from Step 1 with the next available image.

d) Validate Camera on Flickr Camera Database: Fourth, with a validated creator and recorded camera information, the validator verifies the camera's authenticity using Flickr's camera database. Flickr maintains a verified database with detailed descriptions for each recognized camera. For images that include a camera in the Flickr API response, there is typically a link to the camera's description on the Flickr image's origin page and the validator then searches for this link. In most cases, this link is present; however, some photographers may use unverified cameras not listed in Flickr's camera database. This step ensures that only images with Flickr-validated cameras proceed to the next step. If the camera cannot be verified in Flickr's camera database (i.e., the link is absent), the creator is added to the "Watchlist" file, noting the image URL and the reason for failure "Camera could not be validated on Flickr page". The image URL is also mapped to its local downloaded path in the "Failed" file. The process then restarts from Step 1 with the next available image.

e) Validate Local and Creator's Images by SSIM: Fifth, once the creator and camera are validated, the downloaded image must be confirmed as identical to the image displayed on the creator's Flickr account. Occasionally, discrepancies arise between the image provided by the Flickr API and the one displayed on its original page, often due to slight modifications such as watermarks, borders, or minor edits. To address this, a Structural Similarity Index Measure (SSIM) is calculated between the two images using the Python Scikit-Image Metrics library. An SSIM range of 95%-100% typically signifies that the images are visually identical, with any variations likely due to minor artifacts or compression. Scores between 85%-95% suggest small edits or adjustments, while scores below 85% indicate significant structural or visual differences, suggesting the images are not the same. For accurate comparison, both images are resized to match the dimensions of the smaller image. A threshold of 90% SSIM was selected to allow for minor modifications, such as watermarks or borders, that photographers might add for copyright purposes. If the SSIM score meets or exceeds 90%, the validation proceeds. If not, the creator is added to the "Watchlist" file, with the image URL recorded alongside the reason for failure "Image downloaded is not visually the same as the image on Creator page based on SSIM scoring". The image URL is also mapped to its local path in the "Failed" file. The process then restarts from Step 1 with the next available image.

f) Complete Validation and Traceable Origins: Finally, with the creator, camera, and image all successfully validated, the authenticity criterion is fully met. The creator is added to a file titled "Criteria Success List," with the validated image URL appended to the creator's list of previously validated image URLs. The image URL is also mapped to its local path in the "Passing" file. With the authenticity validation complete, the image is saved to the "GranularImageCategories" directory specified during the initial setup of the Image Scraper Curl

instance. This process continues until the Image Scraper Curl instance reaches the defined successful image threshold.

B. AI Image Captioner

The second component of the Real-To-AI-Pipeline is the AI Image Captioner, a standalone Flask API. This tool leverages the open-source model qresearch/llama-3-vision-alpha-hf from Hugging Face (available at https://huggingface.co/qresearch/llama-3-vision-alpha-hf).
The AI Image Captioner accepts image requests and returns captions, providing descriptive context for each image. The main components of the AI Image Captioner, which are further detailed in the following subheadings, include the Flask API, the AI captioning model, and the containerization process.

1) Creating the Flask API: The Python Flask library is used to build a simple API capable of handling HTTP requests. This API has two primary methods: one for general image captioning and another specifically for photomicrographs. The rationale for these separate methods lies in the need to provide contextual prompts. For most images, the API sends a prompt to the captioning model asking for a detailed description without specifying the image type, allowing the model to infer its content. However, the llama-3-vision-alpha-hf model requires specific context for accurate descriptions of photomicrographs, as it otherwise struggles to interpret the content correctly. The API operates by opening a web socket that receives HTTP POST requests with an image attachment. Upon startup, it initializes an instance of the AI Image Captioner, which will also be referenced in the Leonardo Image Generator section. The API route for general image captioning is "/caption," while photomicrograph images are sent to a dedicated route, "/caption-photomicrograph."

2) The Captioning Model: When the Flask API initializes, it creates an instance of the AI "Captioner" class. This Captioner instance is configured to run offline, ensuring that the llama-3-vision-alpha-hf model is fully tokenized and loaded within its container without needing an internet connection. If the cached model in the GitHub repository cache directory is missing or corrupted, the Captioner is designed to detect this issue and attempt to retrieve the latest version from its original Hugging Face repository. Upon successfully downloading the latest model version, it generates a new cache directory to store updated safe-tensor shards efficiently. Once the model is fully loaded within the API container, it offers two main methods corresponding to the API routes designated for general image captioning and photomicrograph captioning. The primary distinction between these methods is that the photomicrograph captioning route provides the Captioning model with context, specifying that the image type is a photomicrograph, allowing for more accurate descriptions.

3) Containerization: Docker is used to containerize the Flask API and AI Captioner components, creating a cohesive and scalable application. The Docker container is based on the official python:3.11.8-slim image. Once the base image is set up, necessary PyTorch and CUDA libraries for GPU interaction are installed from https://download.pytorch.org/whl/cu124. To support the synchronization of the Docker Container and the machine's NVIDIA GPU, WSL2 is used for the Docker Desktop backend. The project's virtual environment dependencies are defined in a requirements file, which the container uses to install additional libraries. After installing dependencies, the contents of the AI Image Captioner directory, including the cached model, are copied into the container. The Docker Compose file is then configured to ensure compatibility with an NVIDIA GPU on the host OS. Once setup is complete, the container is launched, starting the Flask API and instantiating an instance of the Captioner model, which then awaits image POST requests. Upon captioning an image, the Captioner model sends a response containing the generated caption, which can be stored for future use.

C. Leonardo Image Generator

The third and final component of the Real-To-AI-Pipeline is the Leonardo Image Generator. Once all images are aggregated into the "GranularImageCategories" directory according to the thresholds set by the Real Image Scraper, each image is sent via HTTP POST to the AI Image Captioner container. The captions generated for each image are recorded for later use. After all images are captioned, each image and its corresponding caption are submitted to the Leonardo.AI commercial API to produce a hyper-realistic AI-generated counterpart. The API is polled until the AI-generated image is ready, at which point it is saved locally to a designated "AI" directory. The following subheadings provide a detailed, sequential overview of this process.

1) Captioning Images from Image Directory Paths: All image directory paths for each subcategory are recorded in a file titled "Directory Paths." To handle photomicrographs separately in the captioning process, the paths for each subcategory within the Photomicrographs directory are specifically recorded in a file named "Photomicrograph Paths." Once the captioner generates a caption for a given image, the caption is saved in an "Images Captioned" file, mapped to the image's local path. This mapping ensures that all images are captioned and allows each image, along with its unique caption, to be sent to the Leonardo.AI commercial API.

2) Generating Hyper-Realistic AI Images: After all images have been captioned, the Directory Path and Photomicrograph Path files are used to locate each image and its associated caption for submission to the Leonardo.AI commercial API. First, a pre-signed URL is requested from Leonardo.AI to send an authenticated image generation request. The application sends the captioned image to Leonardo.AI's "Image to Image" generation feature with the image's caption as the prompt. To ensure consistency, the AI-generated image's dimensions are configured to match the original image's height and width, maintaining the aspect ratio between the authentic image and its AI-generated counterpart. The model selected for generating images is the Leonardo Vision XL model, with further details on model selection provided in section IX.

The application then polls the Leonardo.AI API to track the generation status. Once an AI-generated image is ready, it is downloaded and saved to the "AI" local directory. A "Main Manifest" file records the local path of each AI-generated image and maps it to the original image's local path, enabling analytical comparisons between paired images. This process iterates over all captioned images until each has a hyper-realistic AI counterpart, completing the Real-To-AI-Pipeline and finalizing the WIRA dataset creation.

VII. Authenticity Validation And Traceable Origins For Photomicrographs

As noted in section IV, the Photomicrographs category does not apply the main Authenticity Validation and Traceable Origins criterion described in sections V or VI. This decision was made due to the lack of mainstream capability on Flickr for photographers to record specific tools, such as microscopes, within Flickr's camera database. Consequently, a modified approach was applied for authenticity validation within the Photomicrographs category. For WIRA's transparency, all sources for the Bacteria, Cancer Cells, Healthy Cells, Fungi, Parasites, and Virus subcategories are all thoroughly cited for transparency, ensuring their Traceable Origins. These subcategories contain images exclusively aggregated by hand from reliable sources, including the CDC's Public Health Image Library (CDC PHIL - https://phil.cdc.gov) , the Broad Institute's Broad Bioimage Benchmark Collection (BBBC - https://bbbc.broadinstitute.org), the Image Data Resource for Open Microscopy (IDR - https://idr.openmicroscopy.org), and IAQ Consultants (https://www.iaqsg.com). Where available, each image from these sources is further documented with its original publication reference. In addition to recording these sources, the images are cited within the GitHub repository, ensuring that each photomicrograph meets the requirements for traceable origins. This adjusted authenticity validation method provides a transparent and traceable foundation for the photomicrographs included in WIRA.

A. Bacteria

All photomicrographs of bacteria were hand collected from the DAS+4tag_Trial2 images from IDR located on https://doi.org/10.17867/10000151b. This subset of images originates from Z. Ali, V. Parisutham, S. Choubey, and R. C. Brewster's study [37] containing photomicrographs of E. Coli bacteria. Other individual images were hand collected from the CDC PHIL with no direct links to original publications. The citations of the IDR image set, the IDR image set's origin publication, as well as the individual CDC PHIL URLs are cited within the GitHub repository.

B. Cancer Cells

All photomicrographs of cancer cells were hand collected from BBBC image datasets with the subset of images originating from BBBC001 and BBBC0018 from J. Moffat et al's study [38]. Also, BBBC006 is used but has no direct link to an original publication. The citation of these image sets as well as their available origin publication are cited within the GitHub repository.

C. Healthy Cells

All photomicrographs of healthy cells were hand collected from BBBC image datasets along with image datasets from IDR. The BBBC009 dataset is used but has no direct link to an original publication. Image sets AT1G02730, AT1G05570, were aggregated from IDR but originates from W. Yang et al's study [39]. Image sets Diplophyllum taxifolium and Scapania mucronate were aggregated from IDR originating from a study from K. Peters and B. König-Ries [40]. The citation of these image sets as well as their available origin publication are cited within the GitHub repository.

D. Fungi

All photomicrographs of fungi were hand collected from the CDC PHIL and IAQ Consultants with no direct links to original publications. The citations for the individual CDC PHIL and IAQ Consultants URLs are cited within the GitHub repository.

E. Parasites

All photomicrographs of parasites were hand collected from BBBC image datasets. The BBBC010 image dataset is used originating from Moy et al's study [41]. Also, BBBC041 is used but has no direct link to an original publication. The citation of these image sets as well as their available origin publication are cited within the GitHub repository.

F. Viruses

All photomicrographs of viruses were hand collected from IDR, some of which do not contain any original publication. These are the Zb_BSF019089, BSF019243-1A, and preScreen datasets on IDR. The BSF018307-4D image dataset is used originating from F. Georgi et al's study [42]. The citation of these image sets as well as their available origin publication are cited within the GitHub repository.

VIII. Computational Resources For WIRA Construction

This section presents the specific hardware and software resources used in the construction of the WIRA dataset, which was developed entirely on a local machine. Including these details ensures transparency and supports reproducibility for researchers who may wish to replicate or extend this work without relying on cloud resources. Table 1 displays the hardware specifications while Table 2 displays the software environmeent of the machine used to construct WIRA.

IX. WIRA Decsion Records

This section presents the decisions made throughout the creation of WIRA, detailing both accepted and rejected choices along with the rationale behind each. Organized chronologically, it provides an explanation for the inclusion or exclusion of each decision, offering a transparent view into the dataset's development process.

TABLE I
HARDWARE SPECIFICATIONS

Component	Specification
Machine	Dell Precision 7770
Processor	12th Gen Intel Core i7-12850HX 2.10GHz
Installed RAM	64.0 GB DDR5 4800MHz CAMM non-ECC
System Type	64-bit OS, x64-based processor
Integrated GPU	Intel UHD Graphics, 32.0 GB
Discrete GPU	NVIDIA RTX A1000 Laptop GPU, 4GB GDDR6

TABLE II
SOFTWARE ENVIRONMENT

Component	Specification
Operating System	Windows 11 Pro
OS Version / Build	23H2 / 22631.446
Editor	VS Code
Programming Language	Python 3.11.8
AI/ML Backend	PyTorch 2.5.1
CUDA Version	12.4
Containerization Platform	Docker Desktop v4.34.3 (170107)
Docker Official Image	python:3.11.8-slim
WSL version	2.2.4.0
AI Image Generator Platform	Leonardo.AI Production API v1.0
Metadata Tool	ExifTool by Phil Harvey 12.96

A. Third-Party Software to Validate Images

Third-party software, such as APIs like isitai.com, was initially considered to streamline the validation of web-scraped images by detecting anomalies indicative of AI-generated content, thereby potentially expediting the authenticity validation process. However, it was determined that such tools should not be part of the authenticity validation process, as they do not provide insight into an image's origin or verify its authenticity from a photographer. Additionally, relying on a third party for validation could compromise the credibility of authenticity, especially as the Real-To-AI Pipeline already depends on the search engine as a third party for initial image sourcing.

B. Image Captioning Websites

Websites like https://pallyy.com/tools/image-caption-generator can be used to automatically caption images, which is an essential component of the Real-To-AI Pipeline. However, these tools were found to produce subpar results when compared to outputs from open-source AI models, such as the llama-3-vision-alpha-hf model.

C. Source of Images Scraped

Initially, Google was selected as the primary source for scraping images. However, as the Authenticity Validation Criterion evolved, it became clear that using a Custom Google Search Engine would better streamline the web-scraping process. Despite this adjustment, challenges persisted in maintaining accountability for image sources on Google. It was rare to identify the original author of an image, and verifying whether an image was authentically captured by a camera proved difficult. Even reverse image searches often failed to provide the oldest publication date, as some entries lacked this information altogether. These limitations made the web-scraping process inefficient for compiling authentically validated images from photographers. Consequently, Flickr was chosen due to its robust API, which supports thorough investigation into the source and origin of each image. This approach ensures that if an image is later determined to be non-authentic despite passing the Authenticity Validation criteria, accountability rests solely with the photographer, not the search engine. The combination of the validation criteria and Flickr's platform reinforces the authenticity of passing images, enabling each to be traced back to its photographer, who has attested to its authenticity.

D. Image Metadata for Authenticity Validation

While tools like Phil Harvey's ExifTool make it easy to access an image's metadata, they equally allow for metadata manipulation. Initially, metadata was considered a primary factor for determining an image's authenticity; however, a malicious actor could use the same tool to alter metadata on an AI-generated image. Consequently, metadata is now utilized solely for analytical purposes and does not play a role in any stage of the Authenticity Validation criteria.

E. Using Cloud Computing Architecture

Due to time and funding constraints during the research and development of WIRA, implementing a cloud computing architecture was not feasible. However, as outlined in section X, future integration of cloud architecture could significantly enhance WIRA's capabilities.

F. Transparency of WIRA

WIRA is designed to maintain complete transparency, allowing users to easily critique or validate its contents. For each successfully aggregated image, the photographer assumes full responsibility for ensuring that the content they produce is authentically real. Critics and researchers can use the analytical files detailed in section VI, available in the GitHub repository, to analyze each image. Without such transparency, determining "what is real anymore" would not be possible.

G. Choice of Leonardo.AI Model for AI-Generated Images

Through rigorous testing of various parameter settings for the Leonardo.AI commercial API, many models were tested. Figure 3 shows a comparison of real images to a sample of model and parameter combinations used to determine the image-generation model for WIRA. A larger examination showcasing the comparison of different models and parameter settings can be found on the GitHub repository. As shown in Figure 3, The Leonardo Kino XL and Leonardo Vision XL models performed exceptionally well showing hyper-realistic counterpart images comparatively to the real images sampled. The Leonardo Vision XL model was selected after it was found to produce fewer anomalies compared to the Leonardo Kino XL model when both were analyzed side by side.

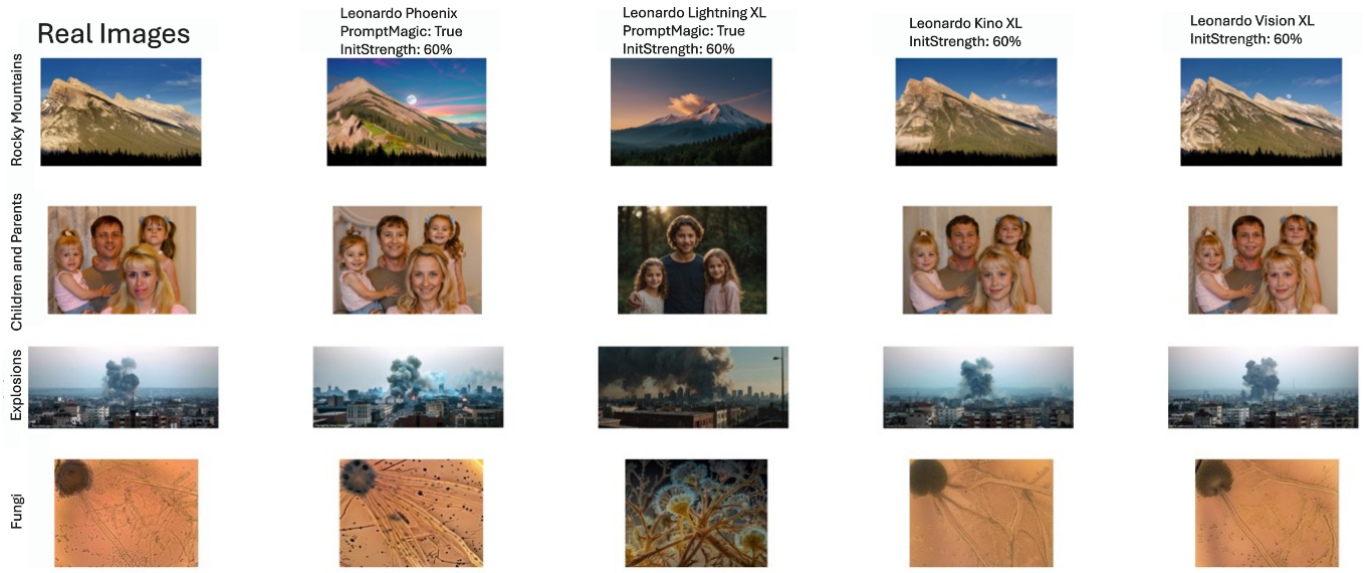

Fig. 3. Showcasing different categories within WIRA, the first column represents authentically real images validated from the Authenticity Validation steps. The following columns represent different model and parameter outputs used when generating hyper-realistic AI counterpart images from the Leonardo.AI commercial API.

X. Potential Enhancements For Expanding WIRA

This section outlines potential future enhancements for the WIRA dataset, focusing on upgrades that can extend its applicability beyond AI-generated image detection to a broader range of AI/ML solutions. Additionally, the following suggestions aim to increase the scalability and adaptability of WIRA, making it a versatile resource for diverse applications.

A. SSIM Scoring Optimization

To improve SSIM scoring accuracy, the comparison process should resize the local image to match the dimensions of the reference image, rather than resizing both images to the smallest dimensions of either image.

B. Cloud Computing Integration

Implementing a cloud computing architecture would support the dataset's scalability, facilitating large-scale processing and storage for extended applications.

C. Image Captioning Standards

For the image captioning process, ensure that the captioning model generates clear, contextually appropriate descriptions that adhere to AI moderation standards, such as those established by Leonardo.AI, to maintain ethical and safe content generation.

XI. Conclusion

The novel "What Is Real Anymore?" (WIRA) dataset marks a pivotal step in authentic image dataset curation for AI-generated image detection. By incorporating a rigorous authenticity validation process and traceable origins, WIRA addresses critical gaps in current datasets by ensuring that each image is authentically validated, and its source is transparent. This dataset not only supports more reliable training for AI/ML models but also establishes a foundation for ethical data use in the face of increasing AI-driven content generation. WIRA's comprehensive methodology, including the Real-To-AI Pipeline, ensures that every authentically real image and its hyper-realistic AI-generated counterpart meet high standards for authenticity and traceability. This dataset fills a vital need in AI research, where distinguishing between authentic and synthetic imagery becomes increasingly challenging due to the sophistication and hyper-realism of AI-generated visuals. Furthermore, WIRA opens doors to scalable and ethically grounded applications in AI/ML. Potential future enhancements, including cloud computing integration and refined image similarity measures, will enable the expansion of WIRA, making it adaptable to diverse research and practical applications. In conclusion, WIRA provides the AI/ML community with a trusted resource for advancing AI-generated content detection, promoting digital integrity, and setting a benchmark for the ethical curation of authentically real data. As the digital landscape continues to evolve, datasets like WIRA will remain instrumental in upholding public trust in visual content and contribute to the protection of innocent individuals against adverse uses of generative AI across the globe.

References

[1] G. Monkam, W. Xu, and J. Yan, "A GAN-based Approach to Detect AI-Generated Images", *2023 26th ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter)*, Taiyuan, China, Jul. 5-7, 2023, pp. 229-232, doi: 10.1109/SNPD-Winter57765.2023.10223798.

[2] Y. Luo, J. Du, K. Yan, and S. Ding, "LaRE2: Latent Reconstruction Error Based Method for Diffusion-Generated Image Detection", *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR), Seattle, WA, USA, Jun. 16-22, 2024, pp. 17006-17015, doi: 10.1109/CVPR52733.2024.01609.

[3] M. Zhang, H. Wang, P. He, A. Malik, and H. Liu, "Improving GAN-Generated Image Detection Generalization Using Unsupervised Domain Adaptation", *2022 IEEE International Conference on Multimedia and Expo (ICME)*, Taipei, Taiwan, Jul. 18-22, 2022, pp. 1-6, doi: 10.1109/ICME52920.2022.9859763.

[4] W. Xia, Y. Zhang, Y. Yang, J. H. Xue, B. Zhou, and M. H. Yang, "GAN Inversion: A Survey", *Computer Vision and Pattern Recognition*, Mar. 22, 2022, doi: 10.48550/arXiv.2101.05278.

[5] Y. Sun, X. Wang, and X. Tang, "Deep Learning Face Representation by Joint Identification-Verification", *Computer Vision and Pattern Recognition*, Jun. 18, 2014, doi: 10.48550/arXiv.1406.4773.

[6] I. Anokhin, K. Demochkin, T. Khakhulin, G. Sterkin, V. Lempitsky, and D. Korzhenkov, "Image Generators with Conditionally-Independent Pixel Synthesis", *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, Jun. 20-25, 2021, pp. 14273-14282, doi: 10.1109/CVPR46437.2021.01405.

[7] F. Zhan et al, "Multimodal Image Synthesis and Editing: The Generative AI Era", *Computer Vision and Pattern Recognition*, Aug. 24, 2023, doi: 10.48550/arXiv.2112.13592.

[8] H. Wang, J. Fei, Y. Dai, L. Leng, and Z. Xia, "General GAN-generated Image Detection by Data Augmentation in Fingerprint Domain", *2023 IEEE International Conference on Multimedia and Expo (ICME)*, Brisbane, Australia, Jul. 10-14, 2023, pp. 1187-1192, doi: 10.1109/ICME55011.2023.00207.

[9] P. Lorenz, R. Durall, and J. Keuper, "Detecting Images Generated by Deep Diffusion Models using their Local Intrinsic Dimensionality", *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Paris, France, Oct. 2-6, 2023, pp. 448-459, doi: 10.1109/ICCVW60793.2023.00051.

[10] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild", *Computer Vision and Pattern Recognition*, Sep. 24, 2015, doi: 10.48550/arXiv.1411.7766.

[11] M. Lin, L. Shang, and X. Gao, "Enhancing Interpretability in AI-Generated Image Detection with Genetic Programming", *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, Shanghai, China, Dec. 1-4, 2023, pp. 371-378, doi: 10.1109/ICDMW60847.2023.00053.

[12] D. C. Epstein, I. Jain, O. Wang, and R. Zhang, "Online Detection of AI-Generated Images", *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Paris, France, Oct. 2-6, 2023, pp. 382-392, doi: 10.1109/ICCVW60793.2023.00045.

[13] C. Schuhmann et al, "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs", *Computer Vision and Pattern Recognition, accepted at Data Centric AI NeurIPS Workshop 2021*, Sydney, Australia, Nov. 3, 2021, doi: 10.48550/arXiv.2111.02114.

[14] Z. Wang et al, "DIRE for Diffusion-Generated Image Detection", *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, Oct. 1-6, 2023, pp. 22388-22398, doi: 10.1109/ICCV51070.2023.02051.

[15] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop", *Computer Vision and Pattern Recognition*, Jun. 4, 2016, doi: 10.48550/arXiv.1506.03365.

[16] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database", *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, Jun. 20-25, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation", *Neural and Evolutionary Computing*, Feb. 26, 2018, doi: 10.48550/arXiv.1710.10196.

[18] M. Zhu et al, "GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image", *Computer Vision and Pattern Recognition*, Jun. 24, 2023, doi: 10.48550/arXiv.2306.08571.

[19] J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images", *Computer Vision and Pattern Recognition*, Mar. 24, 2023, doi: 10.48550/arXiv.2303.14126.

[20] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images", M.S. thesis, Dept. Computer Science, Univ. of Toronto, ON, CA, Apr. 8, 2009. [Online]. Available: https://www.cs.toronto.edu/ kriz/learning-features-2009-TR.pdf.

[21] A. Rahman, B. Paul, N. H. Sarker, Z. I. A. Hakim, and S. A. Fattah, "ArtiFact: A Large-Scale Dataset with Artificial and Factual Images for Generalizable and Robust Synthetic Image Detection", *Computer Vision and Pattern Recognition*, Feb. 24, 2023, doi: 10.48550/arXiv.2302.11970.

[22] Y. Sun, X. Wang, and X. Tang, "Hybrid Deep Learning for Face Verification", *2013 IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, Dec. 1-8, 2013, pp. 1489-1496, doi: 10.1109/ICCV.2013.188.

[23] Y. Sun, X. Wang, and X. Tang, "Deep Learning Face Representation from Predicting 10,000 Classes", *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, Jun. 23-28, 2014, pp. 1891-1898, doi: 10.1109/CVPR.2014.244.

[24] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4217-4228, Dec. 1, 2021, doi: 10.1109/TPAMI.2020.2970919.

[25] T. Karras et al, "Alias-Free Generative Adversarial Networks", *Computer Vision and Pattern Recognition*, Oct. 18, 2018, doi: 10.48550/arXiv.2106.12423.

[26] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training Generative Adversarial Networks with Limited Data", *Computer Vision and Pattern Recognition*, Oct. 7, 2020, doi: 10.48550/arXiv.2006.06676.

[27] A. Aksac, D. J. Demetrick, T. Ozyer, and R. Alhajj, "BreCaHAD: a dataset for breast cancer histopathological annotation and diagnosis", *BMC Research Notes*, vol. 12, no. 82, Feb. 12, 2019, doi: 10.1186/s13104-019-4121-7.

[28] Y. Choy, Y. Uh, J. Yoo, and J. W. Ha, "StarGAN v2: Diverse Image Synthesis for Multiple Domains", *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 13-19, 2020, pp. 8185-8194, doi: 10.1109/CVPR42600.2020.00821.

[29] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene Parsing through ADE20K Dataset", *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 21-26, 2017, pp. 5122-5130, doi: 10.1109/CVPR.2017.544.

[30] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: a database and web-based tool for image annotation", *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157-173, Oct. 31, 2007, doi: 10.1007/s11263-007-0090-8.

[31] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database", *ACM NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, vol. 1, pp. 487-495, Dec. 8, 2014, doi: 10.5555/2968826.2968881.

[32] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo", *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, Jun. 13-18, 2010, pp. 3485-3492, doi: 10.1109/CVPR.2010.5539970.

[33] T. Y. Lin et al, "Microsoft COCO: Common Objects in Context", *Computer Vision and Pattern Recognition*, Feb. 21, 2015, doi: 10.48550/arXiv.1405.0312.

[34] M. Cordts et al, "The Cityscapes Dataset for Semantic Urban Scene Understanding", *Computer Vision and Pattern Recognition*, Apr. 7, 2016, doi: 10.48550/arXiv.1604.01685.

[35] S. Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now", *Computer Vision and Pattern Recognition*, Apr. 4, 2020, doi: 10.48550/arXiv.1912.11035.

[36] O. Russakovsky et al, "ImageNet Large Scale Visual Recognition Challenge", *Computer Vision and Pattern Recognition*, Jan. 30, 2015, doi: 10.48550/arXiv.1409.0575.

[37] Z. Ali, V. Parisutham, S. Choubey, R. C. Brewster, "Inherent regulatory asymmetry emanating from network architecture in a prevalent autoregulatory motif", *Elife*, Aug. 18, 2020, doi: 10.7554/eLife.56517.

[38] J. Moffat et al, "A Lentiviral RNAi Library for Human and Mouse Genes Applied to an Arrayed Viral High-Content Screen", *Cell*, vol. 124, no. 6, pp. 1283-1298, Mar. 24, 2006, doi: 10.1016/j.cell.2006.01.040.

[39] W. Yang et al, "Regulation of Meristem Morphogenesis by Cell Wall Synthases in Arabidopsis", *Current Biology*, vol. 26, no. 11, pp. 1404-1415, Jun. 6, 2016, doi: 10.1016/j.cub.2016.04.026.

[40] K. Peters and B. König-Ries, "Reference bioimaging to assess the phenotypic trait diversity of bryophytes within the family Scapaniaceae", *Scientific Data*, no. 598, Oct. 4, 2022, doi: 10.1038/s41597-022-01691-x.

[41] T. I. Moy et al, “High-Throughput Screen for Novel Antimicrobials using a Whole Animal Infection Model”, *ACS Chemical Biology*, vol. 4, no. 7, pp. 527-533, Jun. 29, 2009, doi: 10.1021/cb900084v.
[42] F. Georgi et al, “A high-content image-based drug screen of clinical compounds against cell transmission of adenovirus”, *Scientific Data*, no. 265, Aug. 12, 2020, doi: 10.1038/s41597-020-00604-0.