AI Vs AI

By

Abdallah Mamdouh Fathy	*20200308*
Abdelhamid Adel Abdelhamid	*20200275*
Hagar Mohammed Hassan	*20201205*
Tasneem Yaser Al-Maghawry	*20201047*
Waad Ragab Barakat	*20201216*
Zainab Gamal Mohammed	*20200811*

Supervised by

Dr. Mohammed Wahby

TA. Asmaa Ahmed

Artificial Intelligence Department

Cairo University

2023-2024

Table of Contents

List of Tables

List of Figures

Abstract

The extraordinary ability of generative models to generate photographic images has intensified concerns about the spread of disinformation, thereby leading to the demand for detectors capable of distinguishing between AI-generated fake images and real images. In an era dominated by digital information, the emergence of AI-generated content has introduced unprecedented challenges to the authenticity and integrity of online discourse. Our project delves into the critical importance of detecting content generated by AI, as it explores the evolving landscape of AI technologies and their implications for digital content creation. The project underscores the pressing need to develop robust detection mechanisms to discern between AI-generated and human-authored content, aiming to mitigate the risks of misinformation, manipulation, and malicious use. Through a comprehensive examination of the challenges posed by AI-generated content and the strategies employed to detect it, this paper sheds light on the ongoing battle between AI and detection methods. By emphasizing transparency, accountability, and ethical considerations, this research contributes to the advancement of responsible AI usage and fosters a more trustworthy digital environment.

Chapter 1: Introduction

Artificial intelligence (AI) has revolutionized content creation, ushering in an era where AI-generated content plays a pivotal role in diverse scenarios. This content, produced by sophisticated generative AI models, encompasses a wide range of formats including text, images, audio, video, and cross-modal transformations such as text-to-image and text-to-audio. These advancements have enabled AI to generate realistic and creative content that was once solely the domain of human creators.

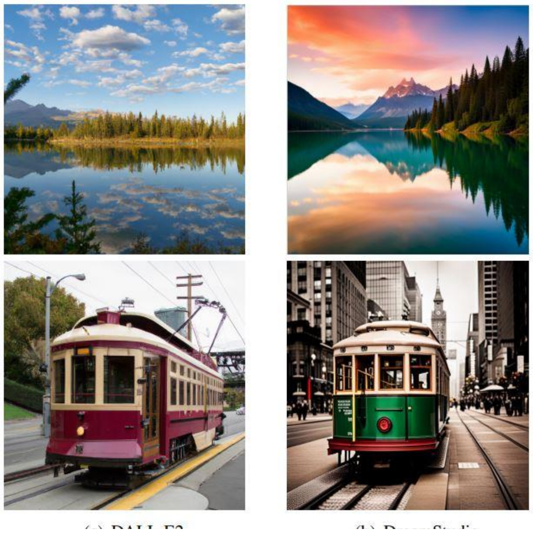


Figure 1 *Generated images utilizing two different AI systems, DALL·E2 and DreamStudio.*

AI-generated text, for example, has found applications in automated journalism, content marketing, and creative writing. Models like GPT-4 can generate coherent and contextually relevant articles, stories, and even poetry. In the realm of images, generative adversarial networks (GANs) and other deep learning techniques produce photorealistic images that are indistinguishable from those captured by cameras. These images are used in various fields such as advertising, entertainment, and even in creating synthetic datasets for training other AI models.

Audio and video content generation through AI has similarly seen substantial progress. AI systems can create realistic voiceovers, music compositions, and even full-length movies or animations. Text-to-speech models convert written text into natural-sounding speech, enhancing applications in virtual assistants and accessibility tools. Video generation and deepfake technologies allow for the creation of lifelike animations and simulations, which have applications ranging from entertainment to education and beyond.

Cross-modal transformations, such as text-to-image and text-to-audio, further expand the creative possibilities. These technologies enable users to generate images from textual descriptions, opening up new avenues for visual storytelling and design. Similarly, text-to-audio models can generate sound effects and audio descriptions from written content, enhancing multimedia experiences.

The influence of AI-generated content is profound, driving innovation and efficiency in content creation while also raising important ethical and regulatory considerations. As these technologies continue to evolve, they promise to reshape the landscape of media and communication, offering unprecedented opportunities for creativity and productivity.

1.1. Problem Definition

Detecting AI-generated content is of paramount importance due to the growing threat posed by the dissemination of manipulated visual content across various online platforms. The ability to identify AI-generated content is essential for preventing the misuse of visual media for harmful purposes. AI-generated content can be employed in spreading propaganda, creating fake identities, and manipulating public opinion, as well as in various forms of cybercrime. These actions can deceive individuals and have serious implications for society.

As AI algorithms continue to evolve, they can produce incredibly realistic content that is nearly indistinguishable from authentic material. This presents a significant challenge, as it becomes increasingly difficult for users to differentiate between genuine and AI-generated content. The potential for disseminating false or misleading information grows, making effective detection mechanisms crucial. Without these safeguards, individuals may be deceived into believing manipulated content is real, leading to decisions based on inaccurate information.

The consequences of unchecked AI-generated content are far-reaching. The spread of misinformation can undermine trust in media sources, exacerbate societal tensions, and contribute to the erosion of public confidence. Additionally, the pervasive influence of AI-generated content can skew public perception and disrupt democratic processes by manipulating information flow and public discourse. Therefore, robust detection methods are necessary to mitigate these risks and ensure the integrity of information consumed by the public.

1.2. Objectives

Our objective is to develop robust detection methods that can effectively mitigate the risks associated with manipulated visual content and prevent its harmful impact. These methods aim to identify AI-generated content, thus upholding ethical standards and promoting transparency in digital communication. By ensuring that users are informed about the origin and authenticity of the visual content they encounter online, we contribute to a more trustworthy digital environment.

Achieving this goal involves tackling a range of challenges. First, we must contend with the unprecedented sophistication of modern AI models, which are increasingly capable of creating highly convincing fake images and videos. Additionally, adversarial techniques are continually evolving, designed specifically to bypass detection systems and evade scrutiny. Furthermore, the vast and diverse nature of digital content circulating across online platforms adds another layer of complexity to our task. Each of these challenges requires innovative approaches and advanced technologies to ensure our detection methods remain effective and reliable.

1.3. Challenges

Detecting content generated by AI presents numerous challenges, primarily due to the rapid evolution of AI algorithms, especially generative models like GANs (Generative Adversarial Networks). These models can produce highly realistic content that is often indistinguishable from real content to the naked eye.

Firstly, acquiring a diverse and comprehensive dataset of AI-generated photos is a significant challenge. This process often requires collaboration with researchers and institutions, access to specialized datasets, and the generation of synthetic data to represent various AI algorithms, manipulation techniques, and stylistic variations. This ensures that the detection models are exposed to a wide range of scenarios and can generalize well across different types of AI-generated content.

Secondly, selecting and training appropriate models, such as Convolutional Neural Networks (CNNs) or even specialized GANs for detection purposes, is crucial for achieving high detection accuracy. This involves employing advanced strategies like transfer learning, where pre-trained models are fine-tuned on new data, and data augmentation, which artificially increases the diversity of the training dataset by applying transformations like rotation, scaling, and flipping.

Additionally, AI-generated photos can be deliberately crafted to evade detection models through adversarial attacks. This necessitates the development of resilient detection systems that can withstand such attacks. Techniques like robust training, where models are exposed to adversarial examples during the training process, are essential for building systems that can detect manipulated content even under adversarial conditions.

Meaningful feature engineering is also a critical component in detecting AI-generated content. This involves identifying and extracting features that highlight the unique characteristics of AI-generated images. Analyzing statistical properties, texture patterns, and noise characteristics can provide valuable insights that differentiate synthetic content from authentic images.

Lastly, generating accurate ground truth labels for AI-generated photos poses a significant challenge due to their high resemblance to real images. Effective labeling strategies are required, often involving expert knowledge and rigorous adversarial testing to ensure the reliability of the training datasets. This meticulous process helps create robust training datasets that can enhance the overall performance of detection models.

So, the detection of AI-generated content requires a multifaceted approach that includes dataset acquisition, model selection and training, adversarial robustness, feature engineering, and precise labeling. Each of these components plays a crucial role in developing reliable detection systems that can effectively identify manipulated visual content and uphold the integrity of digital media.

1.4. Model Overview

After reviewing similar products and research papers in the market, as explained in Chapter 2, we identified the need for a more generalized dataset with a greater variety of generators to address existing limitations. To achieve our goal of creating a model capable of detecting fake images from most well-known generators and performing acceptably on unseen ones, we selected a dataset encompassing images from 25 different generators. This dataset includes 13 GAN-based, 7 diffusion-based, and 5 other miscellaneous generators, ensuring a diverse range of generator types. Additionally, we prioritized choosing a large-scale dataset with varied content, as images depicting inanimate subjects tend to be more believable and, therefore, more challenging to distinguish from real ones.

Existing products often struggle to detect fake images generated by models not seen during the training phase. To address this issue, we propose several enhancements to improve model performance. Firstly, we aim to ensure more accurate and reliable results through a carefully designed model architecture. Our model employs a capsule network for detecting fake images, where the pre-processed images are passed through a portion of the pre-trained VGG-19 network before entering the capsule network. We use a Capsule-Forensics network with ten primary capsules, maintaining a consistent design across all primary capsules.

Capsule networks offer several advantages over traditional Convolutional Neural Networks (CNNs) when it comes to detecting fake images from real ones. Unlike CNNs, which rely heavily on max-pooling layers that can lose valuable spatial hierarchies, capsule networks preserve spatial relationships between features through the use of dynamic routing between capsules. This capability allows capsule networks to maintain a more detailed and holistic understanding of the image structure, making them particularly adept at recognizing subtle manipulations and anomalies that are characteristic of AI-generated images. Additionally, capsule networks are designed to handle variations in viewpoint and orientation more effectively than CNNs, further enhancing their ability to distinguish between real and fake images, even when the manipulations are sophisticated. By leveraging these unique strengths, capsule networks provide a more robust and accurate approach to detecting visual content generated by advanced AI techniques.

Additionally, we apply regularization techniques, including random noise and dropout during training, and use images resized to 200×200 pixels. For binary classification, we utilize two output capsules that process the output of the primary capsules via dynamic routing, followed by a softmax layer. The training phase involves using the cross-entropy loss function and the Adam optimizer to enhance model accuracy.

In summary, our approach combines a comprehensive and diverse dataset with a robust model architecture and advanced training techniques to address the shortcomings of existing products, providing a more effective solution for detecting AI-generated content.

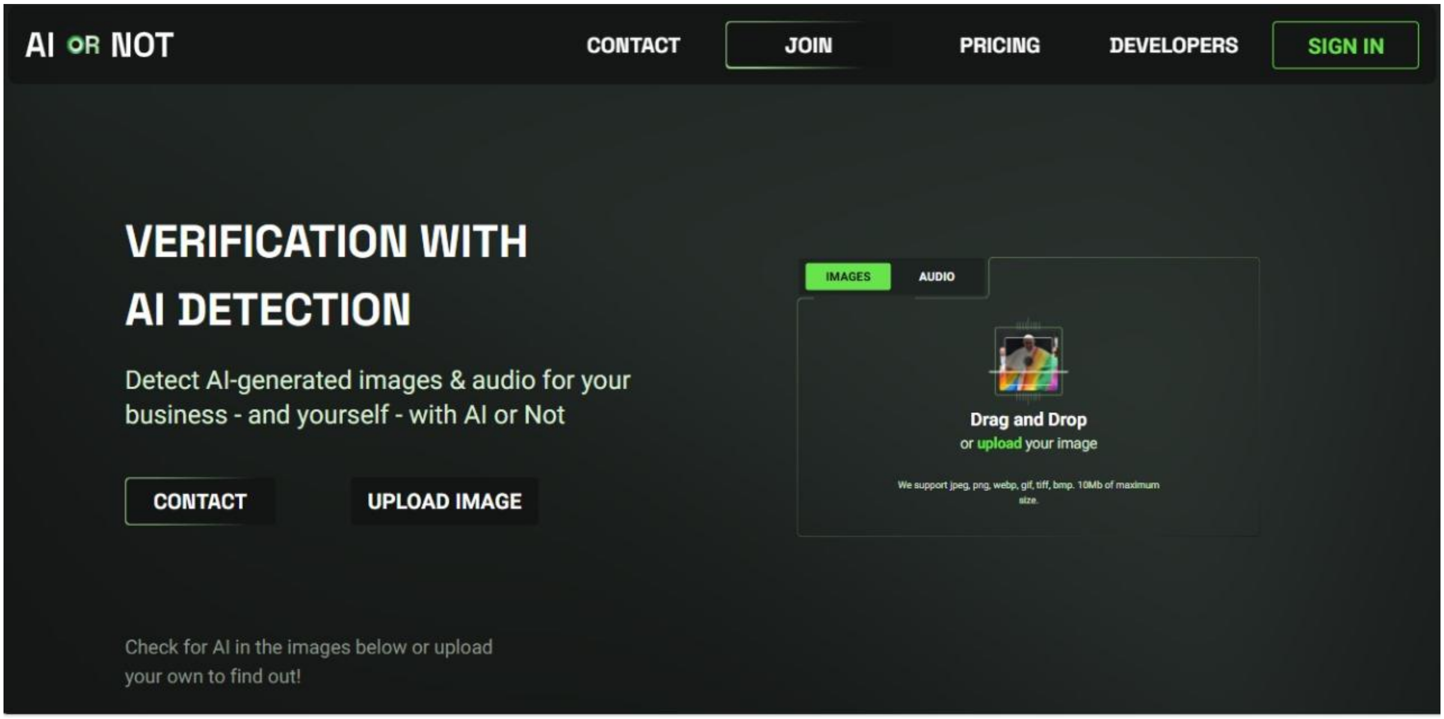
Chapter 2: Literature Review

In this section, we first introduce similar products already existing in the market. We then mention several major research efforts focused on forgery detection preceding our research.

2.1. Similar Products

In order to understand the landscape of fake image detection technology and its market potential to know how exactly a value would be added, we had assessed the existing similar products, the market already offers several products aimed at detecting fake or AI-generated images.

Some of them AI or Not [1] is a web-based tool designed to determine whether an image has been generated by artificial intelligence.AI or Not support only PNG and JPEG images. It uses advanced algorithms and machine learning techniques to analyze images and detect signs of Al generation. Our service compares the input image to known patterns, artifacts, and characteristics of various Al models and human-made images to determine the origin of the content. It identifies image from 5 AI Generating Models which are MidJourney [2], DALL-E2 [3], GAN, Generated faces and Stable Diffusion [4].



VERIFICATION WITH AI DETECTION

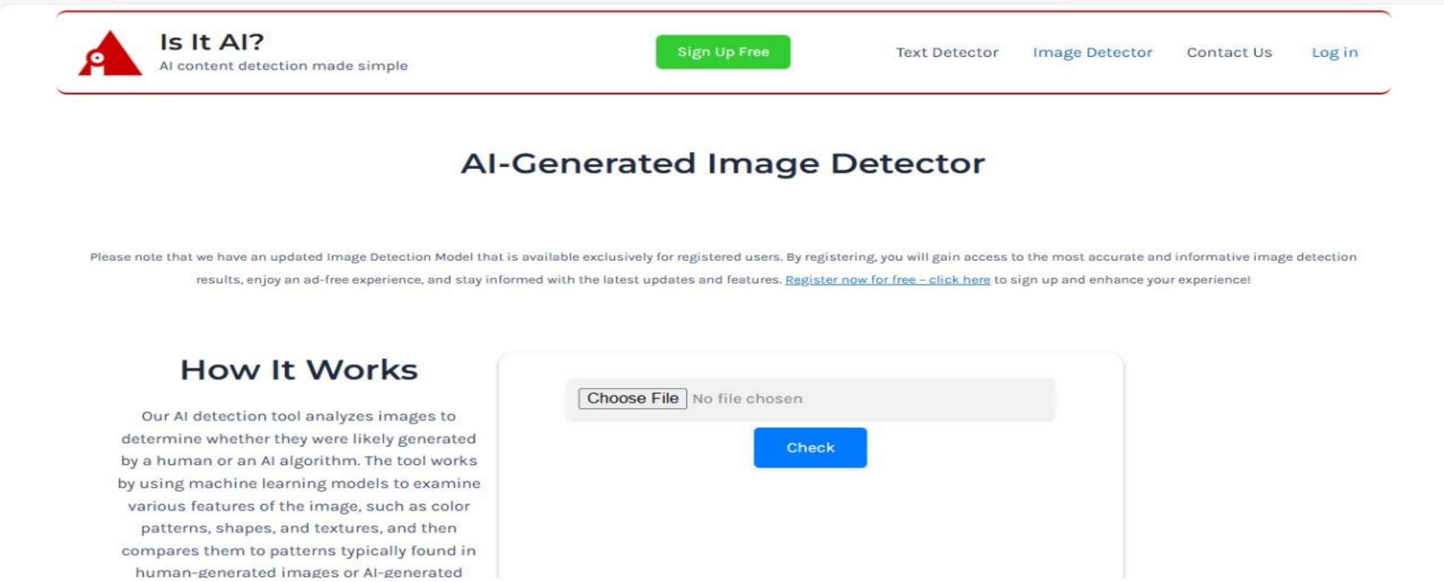
Detect AI-generated images & audio for your business - and yourself - with AI or Not

We support jpeg, png, webp, gif, tiff, bmp. 10Mb of maximum size.

Check for AI in the images below or upload your own to find out!

Figure 2 *AI OR NOT Website*

Fake Image Detector [5] is another software solution developed for detecting and flagging fake or manipulated images in real-time. The platform offers an intuitive user interface that enabling users to quickly identify suspicious JPG images, it classifies images using techniques like metadata analysis and ELA achieving poor performance for real images with high quality.



AI-Generated Image Detector

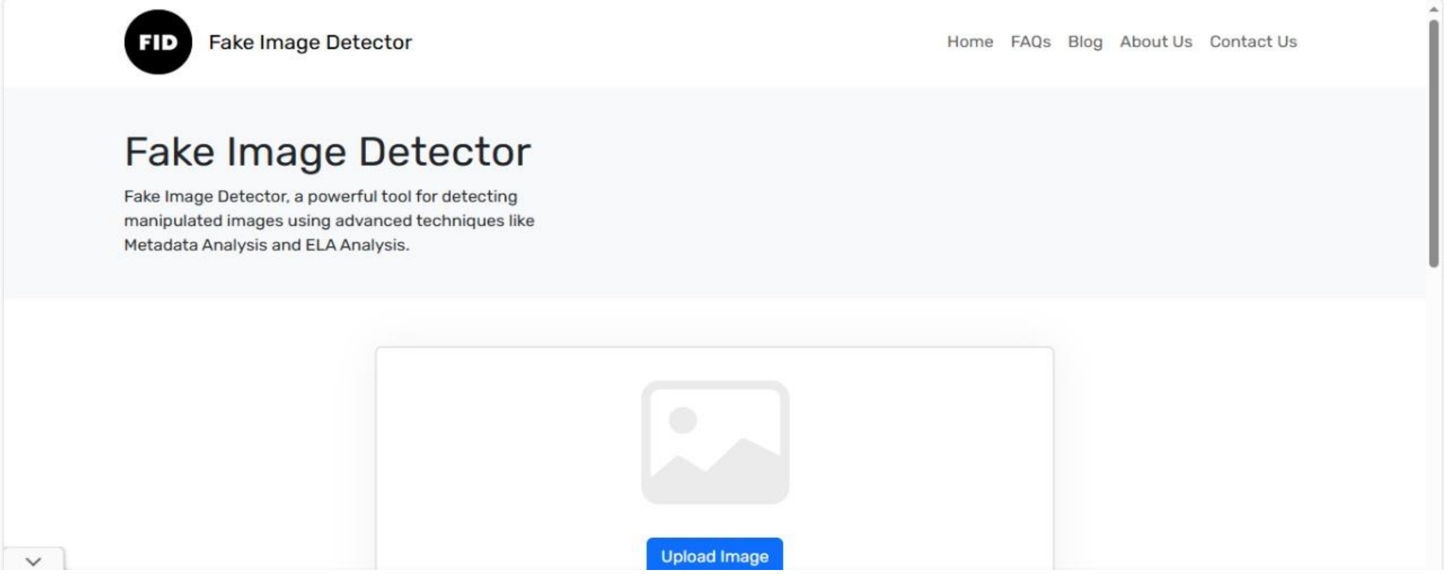
Please note that we have an updated Image Detection Model that is available exclusively for registered users. By registering, you will gain access to the most accurate and informative image detection results, enjoy an ad-free experience, and stay informed with the latest updates and features. Register now for free - click here to sign up and enhance your experience!

How It Works

Our AI detection tool analyzes images to determine whether they were likely generated by a human or an AI algorithm. The tool works by using machine learning models to examine various features of the image, such as color patterns, shapes, and textures, and then compares them to patterns typically found in human-generated images or AI-generated

Figure 3 *Is It AI Website*

Is It Ai [6] is another used AI detection tool analyzes images to determine whether they were likely generated by a human or an AI algorithm. The tool works by using machine learning models to examine various features of the content., such as color patterns, shapes, and textures, and then compares them to patterns typically found in human-generated images or AI-generated content, but it also has limitations with low-quality images and images that have been heavily edited or manipulated. It is also possible for human-generated images to be incorrectly but confidently labeled as AI-generated, if the image contains certain features that are similar to those found in AI-generated images. Additionally, the Is It AI image detector should not be used as a primary decision-making tool, but rather as a complement to other methods of determining the source of an image.

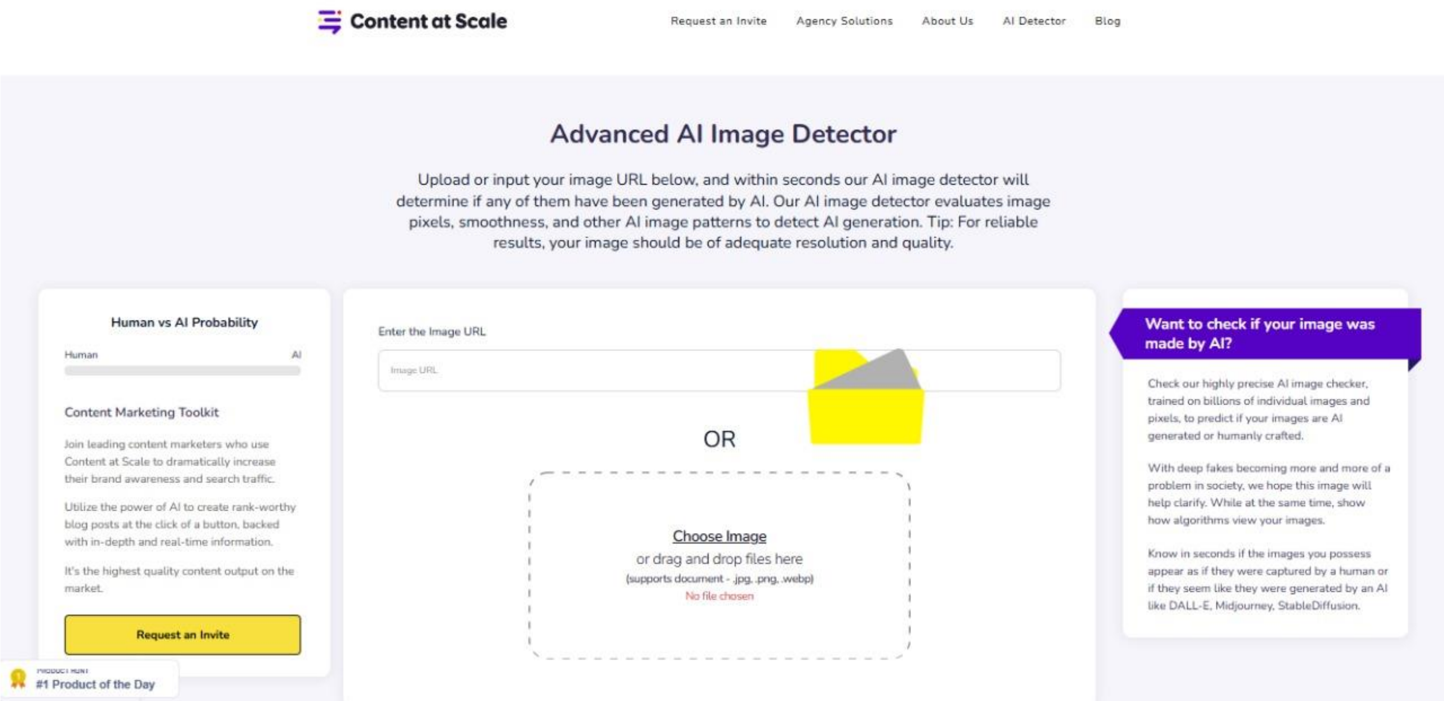


Fake Image Detector

Fake Image Detector, a powerful tool for detecting manipulated images using advanced techniques like Metadata Analysis and ELA Analysis.

Figure 4 *Fake Image Detector Website*

Finally, we have Content at Scale [7] which is an AI-powered platform specializing in image moderation and content verification services. The platform offers a comprehensive suite of tools for detecting fake images and make the user know if the images you possess appear as if they were captured by a human or if they seem like they were generated by an AI like DALL-E, Midjourney, StableDiffusion which are very few generating image sources in a world having new every day. We found also some similar projects and implementations but almost all of them have an issue with the generalization on unseen AI-generating models, here are some of them: Fake image detection with 91% accuracy [8] and Detecting-Images-Generated-by-Diffusers with 94% accuracy [9].



Advanced AI Image Detector

Upload or input your image URL below, and within seconds our AI image detector will determine if any of them have been generated by AI. Our AI image detector evaluates image pixels, smoothness, and other AI image patterns to detect AI generation. Tip: For reliable results, your image should be of adequate resolution and quality.

Human vs AI Probability

Human AI

Content Marketing Toolkit

Join leading content marketers who use Content at Scale to dramatically increase their brand awareness and search traffic.

Utilize the power of AI to create rank-worthy blog posts at the click of a button, backed with in-depth and real-time information.

It's the highest quality content output on the market.

Request an Invite

PRODUCT HUNT
#1 Product of the Day

Enter the Image URL

Image URL

OR

Choose Image
or drag and drop files here
(supports document - .jpg, .png, .webp)
No file chosen

Want to check if your image was made by AI?

Check our highly precise AI image checker, trained on billions of individual images and pixels, to predict if your images are AI generated or humanly crafted.

With deep fakes becoming more and more of a problem in society, we hope this image will help clarify. While at the same time, show how algorithms view your images.

Know in seconds if the images you possess appear as if they were captured by a human or if they seem like they were generated by an AI like DALL-E, Midjourney, StableDiffusion.

Figure 5 *Content at Scale Website*

2.2. Research Papers

The rapid advancement of artificial intelligence (AI), particularly in the domain of generative models, has led to a new era of content creation and manipulation. With the propagation of progressing techniques such as Generative Adversarial Networks (GANs), Diffusion Models and deep learning architectures, the production of AI-generated images and videos has become remarkably convincing, With the previously mentioned technologies and the proliferation of users as it has become easy for anyone to create whatever unreal content they want, whether it is a completely new content or manipulating existing one, this leads to increased concerns about the authenticity of digital content which made the research and studies in detecting the synthetic content also have increased. to identify misleading or fake content, researchers have created computer vision methods and tuned existing well-known architectures to fit the problem. **AI-Generated Image Detection using a Cross-Attention Enhanced Dual-Stream Network** [10] propose a novel cross-attention enhanced dual-stream network specifically designed for Text to Image (T2I) and Photographs (PG) generated by a digital camera detection, they randomly select 20,000 uncompressed PG images (256×256) from the Alaska [11] database, extracting specific scenes like sports fields or Gothic architecture. using ChatGPT then expand keywords into 5,000 prompts, generating images with DALL-E2 [3] and DreamStudio [12] The resulting DALL-E2 and DreamStudio databases each contain 20,000 PG images and corresponding T2I images at three resolutions (256×256, 128×128, 64×64),

applying random JPEG compression (75 to 95 quality). Six distinct databases from two AI systems are created, with 12,000 pairs for training, 3,000 for validation, and 5,000 for testing in each. the approach integrates two streams: a residual stream for texture extraction and a content stream emphasizing low-frequency aspects through a dedicated network. Extracted information undergoes downscaling via CNN modules, fused using cross multi-head attention, and combined through channel concatenation. A classifier distinguishes whether the image is Text to Image (T2I), or a Photograph (PG) as represented in Figure 2.

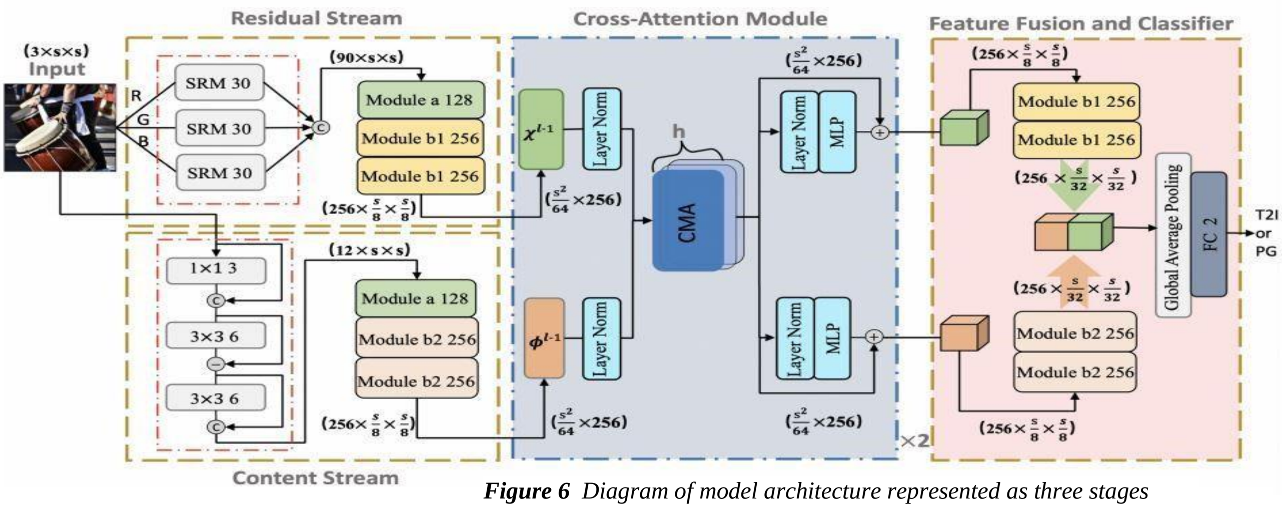

Residual Stream
Cross-Attention Module
Feature Fusion and Classifier
(3×s×s)
Input
R
G
B
SRM 30
SRM 30
SRM 30
(90×s×s)
Module a 128
Module b1 256
Module b1 256
$(256\times\frac{s}{8}\times\frac{s}{8})$
χ^{l-1}
Layer Norm
$(\frac{s^2}{64}\times256)$
h
CMA
Layer Norm
MLP
Layer Norm
MLP
$(\frac{s^2}{64}\times256)$
$(256\times\frac{s}{8}\times\frac{s}{8})$
Module b1 256
Module b1 256
$(256\times\frac{s}{32}\times\frac{s}{32})$
Global Average Pooling
FC 2
T2I
or
PG
$(256\times\frac{s}{32}\times\frac{s}{32})$
Module b2 256
Module b2 256
$(256\times\frac{s}{8}\times\frac{s}{8})$
1×1 3
3×3 6
3×3 6
(12×s×s)
Module a 128
Module b2 256
Module b2 256
$(256\times\frac{s}{8}\times\frac{s}{8})$
ϕ^{l-1}
Layer Norm
$(\frac{s^2}{64}\times256)$
$(\frac{s^2}{64}\times256)$
×2
Content Stream

Figure 6 *Diagram of model architecture represented as three stages*

Although the model achieved better performance than other models as represented in table1, the study reveals that the proposed model consistently outperforms alternatives, even at 64×64 resolution, the model maintains impressive accuracy, achieving 93.1% and 97.8%, but the lake of the fake image's sources variation in dataset, only depends on DALL.E2 and DreamStudio make the model provide very low accuracy to the fake images from unseen generators, it totally ignores the challenging problem of detector generalization which we targeted in our work.

Table 1 *Comparative results for different methods on DALL.E2 and DreamStudio*

Methods	**DALL.E2**								
	256×256			**128×128**			**64×64**		
	TPR	**TNR**	**ACC**	**TPR**	**TNR**	**ACC**	**TPR**	**TNR**	**ACC**
ResNet18	96.3	94.8	95.6	94.0	92.3	93.2	87.4	85.8	86.6
Quan	97.6	97.4	97.5	96.8	95.9	96.3	-	-	-
Yao	96.3	94.5	95.3	96.3	94.2	95.3	86.1	88.3	87.2
SPL2018	98.3	98.1	98.2	97.6	96.8	97.2	92.2	90.3	91.3
He	98.4	98.1	98.3	97.2	97.1	97.2	93.1	90.4	91.8
HcNet	98.6	98.6	98.6	97.5	96.0	96.8	90.9	88.1	89.5
QuanNet	98.6	98.4	98.5	98.5	97.2	97.9	93.2	87.9	90.6

CGNet	98.4	98.1	98.3	97.9	98.0	98.0	92.5	93.1	92.8
Ours	99.3	99.1	99.2	98.6	97.9	98.3	93.1	93.1	93.1

DreamStudio								
256×256			128×128			64×64		
TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR	ACC
97.8	96.3	97.1	97.5	97.0	97.2	93.3	93.7	93.5
98.2	98.6	98.4	98.5	97.3	97.9	-	-	-
97.1	97.5	97.3	97.6	97.0	97.3	91.3	92.3	91.8
98.9	99.0	99.0	99.3	98.8	99.1	96.5	95.3	95.9
99.0	98.1	98.5	99.2	98.7	99.0	97.2	96.2	96.7
98.8	99.0	98.9	99.2	97.9	98.5	96.0	95.9	96.0
98.3	98.4	98.4	99.2	98.9	99.1	97.6	97.0	97.3
99.4	99.2	99.3	99.3	98.9	99.1	98.0	97.6	97.9
99.5	99.6	99.5	99.4	99.5	99.5	97.7	97.8	97.8

Detecting Images Generated by Diffusers [13] is another research relative to our problem, Diffusion Models enable the creation of realistic images. The paper introduces an initial effort to differentiate between generated and real images based on both the image content and associated text and explore the characteristics influencing the credibility of such images, making their identification challenging, they considered two starting datasets, namely MSCOCO [14] and Wikipedia Image-Caption Matching dataset [15]. For each of the two datasets, 6000 images were extracted from the training set and a further 6000 images were generated using two text-to-image methods, namely Stable Diffusion and GLIDE. The same was done with 1500 images from the validation set and another 6000 from the test, they find that identifying generated images is feasible using basic Multi-Layer Perceptrons (MLPs) based on features from CLIP or traditional Convolutional Neural Networks (CNNs), for the experiments employ simple Deep Learning architectures as binary classifiers for real or generated images.

The first model uses an MLP with CLIP-extracted features, leveraging its ability to produce expressive features for both text and images. The second category involves standard CNNs like ResNet50 and XceptionNet, each pertained for ImageNet classification, ensuring a fair parameterized comparison across all models.

Table 2 *Model Dataset Mod Features Accuracy AUC Params Pretrain*

Model	Dataset	Mode	Features	Accuracy	AUC	Params	Pretrain
MLP-Base	MSCOCO	Image Only	CLIP VIT	79.5	88.8	23M	N/A
MLP-Base	MSCOCO	Text + Image	CLIP-VIT	78.5	88.8	23M	N/A
MLP-Base	MSCOCO	Image Only	CLIP-R50	67.5	75.0	23M	N/A
MLP-Base	MSCOCO	Text +Image	CLIP-R50	66.5	74.2	23M	N/A
XceptionNet	MSCOCO	Image Only	XceptionNet	94.6	98.9	20M	ImageNet
Resnet50	MSCOCO	Image Only	Reset50	97.1	99.6	23M	ImageNet
MLP-Base	Wikipedia	Image Only	CLIP-VIT	72.8	81.4	23M	N/A
MLP-Base	Wikipedia	Text +Image	CLIP-VIT	73.1	80.8	23M	N/A
MLP-Base	Wikipedia	Image Only	CLIP-R50	65.9	74.2	23M	N/A
MLP Base	Wikipedia	Text +Image	CLIP-R50	64.5	73.5	23M	N/A
XceptionNet	Wikipedia	Image Only	XceptionNet	90.7	97.1	20M	Image Net
Resnet50	Wikipedia	Image Only	Resne150	94.5	98.1	23M	Image Net

Table 3 *Result of the various classifiers and the fake images generated with GLIDE.*

Model	Dataset	Mode	Features	Accuracy	AUC	Params	Pretrain
MLP-Base	MSCOCO	Image Only	CLIP VIT	95.8	99.2	23M	N/A
MLP-Base	MSCOCO	Text +Image	CLIP-VIT	95.8	99.2	23M	N/A
MLP-Base	MSCOCO	Image Only	CLIP-R50	79.0	81.4	23M	N/A
MLP-Base	MSCOCO	Text +Image	CLIP-R50	78.0	86.4	23M	N/A
XceptionNet	MSCOCO	Image Only	XceptionNet	98.9	99.9	20M	ImageNet
Resnet 50	MSCOCO	Image Only	Reset50	99.3	99.9	23M	ImageNet
MLP-Base	Wikipedia	Image Only	CLIP-VIT	93.7	98.4	23M	N/A
MLP-Base	Wikipedia	Text +Image	CLIP-VIT	94.3	98.4	23M	N/A
MLP-Base	Wikipedia	Image Only	CLIP-R50	77.1	83.2	23M	N/A
MLP-Base	Wikipedia	Text +Image	CLIP-R50	75.5	84.5	23M	N/A
XceptionNet	Wikipedia	Image Only	XceptionNet	99.2	99.9	20M	ImageNet
Resnet50	Wikipedia	Image Only	Resnet50	99.5	99.9	23M	ImageNet

The result shown in table 2 and 3 show that detecting images is viable when the generator method is included in the training set, but a significant generalization issue exists. Model trained on Stable Diffusion-generated images can detect GLIDE-generated images, but the reverse is not true. Cross-testing images generated with Stable Diffusion on those generated with GLIDE, and vice versa, yields suboptimal results. On the MSCOCO dataset, training on Stable Diffusion and testing on GLIDE achieves 69.7% accuracy, while the reverse scenario obtains a lower accuracy of 52.3%.

Another technique used for detecting the manipulated visual content discussed in **Can AI Detect AI Generated Images** [16] The paper explores image synthesis challenges post-Generative Adversarial Networks (GANs), acknowledging advancements but expressing concerns about misuse threats. It stresses the need for tools to differentiate AI-generated from real images in forensics. The proposed solution involves fine-tuning a pre-trained CNN on a diverse dataset to effectively detect GAN-generated images, addressing potential security risks.

The paper's structure covers related works, methodology, experiments, results, limitations, and concludes with a call for robust automated models to ensure image integrity across domains.

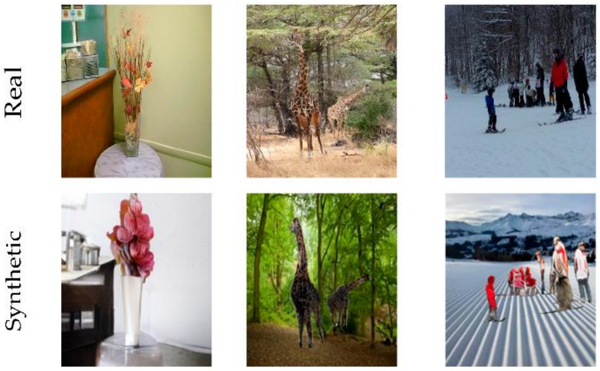


Figure 7 *Examples of real images and their corresponding generated ones produced by various GAN-models based on different tasks.*

The "Data Collection" section introduces the Real or Synthetic Images (RSI) dataset, derived from 12 diverse image synthesis models spanning tasks such as image-to-image, sketch-to-image, and text-to-image synthesis. Originating from COCO-Stuff, an extension of MS COCO, the dataset includes pixel-level annotations for 172 classes, enhancing applicability in various tasks. The RSI dataset consists of 164k images from MS COCO, distributed into training, validation, and testing sets with 80 thing classes, 91 stuff classes, and an 'unlabeled' class. For experiments, subsets of 24k, 12k, and 12k images were curated, maintaining a balance between real and synthetic images. Figure 4 visually demonstrates the dataset's diversity, showcasing synthesized images from different tasks and GAN models. The dataset creation involved gathering real images, generating inputs like text and sketches, and processing them with 12 synthesis models, standardizing outputs to 256 × 256 resolution.

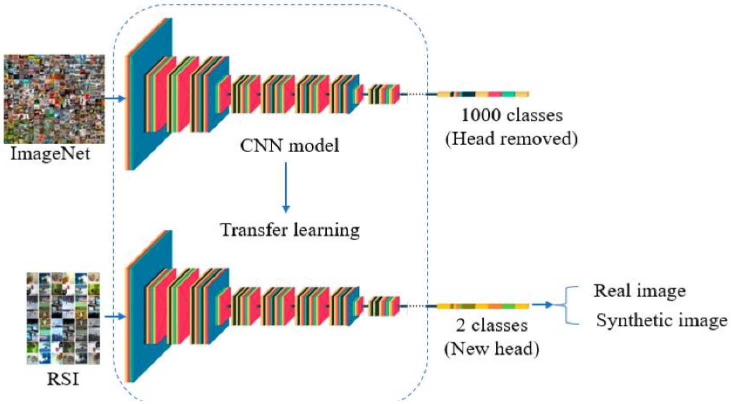


Figure 8 The pipeline of our proposed framework for GAN-generated image recognition.

The paper focused on developing an automated tool to detect GAN-generated images after assembling a diverse dataset. They employed transfer learning, fine-tuning classifiers pre-trained on ImageNet to distinguish between real and synthetic images efficiently. Experimentation involved various classifiers VGG19 [17], ResNet [18] (with different layers),InceptionV3 [19],Xception [20], DenseNet121 [21], InceptionResNetV2 [22],MixConv,MaxViT,and EfficientNetB4 by modifying the classifier heads. Training involved 20 epochs, saving the best models based on validation loss improvements, using a batch size of 64, with adaptive learning rate reduction and data augmentation, primarily horizontal flips, using optimizers like Adam or RMSprop.

This section outlines the evaluation metrics for assessing the efficiency of proposed methods in identifying GAN-generated images. Eight metrics, including Precision, Recall, F1 score, Accuracy, AP, ROC-AUC, FPR, and FNR, were used. EfficientNetB4 achieved 100% accuracy on the RSI dataset, containing 12,000 images equally split between GAN-generated and real images, outperforming InceptionV3 with 98% accuracy. The proper classification is demonstrated through examples, and Class Activation Maps (CAM) methods (GradCAM, AblationCAM, LayerCAM, and Faster ScoreCAM) were integrated to visualize the regions influencing the model's classification decisions during evaluation.

Table 4 *Performance of different classifiers on testing set*

	Precision	**Recall**	**F1**	**Accuracy**	**Ap**	**ROC-AUC**	**FPR**	**FNR**
VGG19	0.94	0.94	0.94	0.94	0.9819	0.9803	0.053	0.064
ResNets50	0.93	0.91	0.91	0.91	0.9933	0.9927	0.0035	0.168
ResNets101	0.95	0.95	0.95	0.95	0.9879	0.9877	0.028	0.08
ResNets152	0.92	0.92	0.92	0.92	0.9743	0.9718	0.042	0.118
Inception V3	0.98	0.98	0.98	0.98	0.9976	0.9974	0.016	0.03
Xception	0.97	0.97	0.97	0.97	0.9995	0.9994	0.0003	0.054
DenseNet121	0.97	0.97	0.97	0.97	0.9969	0.9966	0.012	0.044
InceptionResNet V3	0.96	0.96	0.96	0.96	0.9942	0.9943	0.037	0.036
MixConv	0.94	0.94	0.94	0.94	0.9411	0.9412	0.057	0.056
MixViT	0.92	0.86	0.89	0.89	0.9375	0.9375	0.087	0.137
EfficientNetV4	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00

The model's effectiveness was tested across different input modalities: sketch-to image, text-to-image, and image-to-image models. Three experiments were conducted where models trained on specific modalities were tested on others. The results showed high accuracies, with the S2I_T2I model achieving 0.99 accuracy, I2I_S2I reaching 0.95, and I2I_T2I obtaining 0.83, despite challenges in the image improvement step for sketch-to-image models, showcasing the model's robustness in detecting GAN generated images across diverse synthesis methods.

To enhance generalization problem, we study this research paper: **Online Detection of AI-Generated Images** [23] With advancements in AI-generated images coming on a continuous basis, it is increasingly difficult to distinguish traditionally sourced images from AI-generated ones, AI-generated Images detector preferred to generalize as now we have too many generating models, this paper study generalization in this setting, training on N models and testing on the next (N + k), following the historical release dates of well-known generation methods. As images increasingly consist of both real and generated components extend the approach to pixel prediction, demonstrating strong performance using automatically generated in-painted data, so the paper provides pixel level detector using Cut-Mix augmentation.

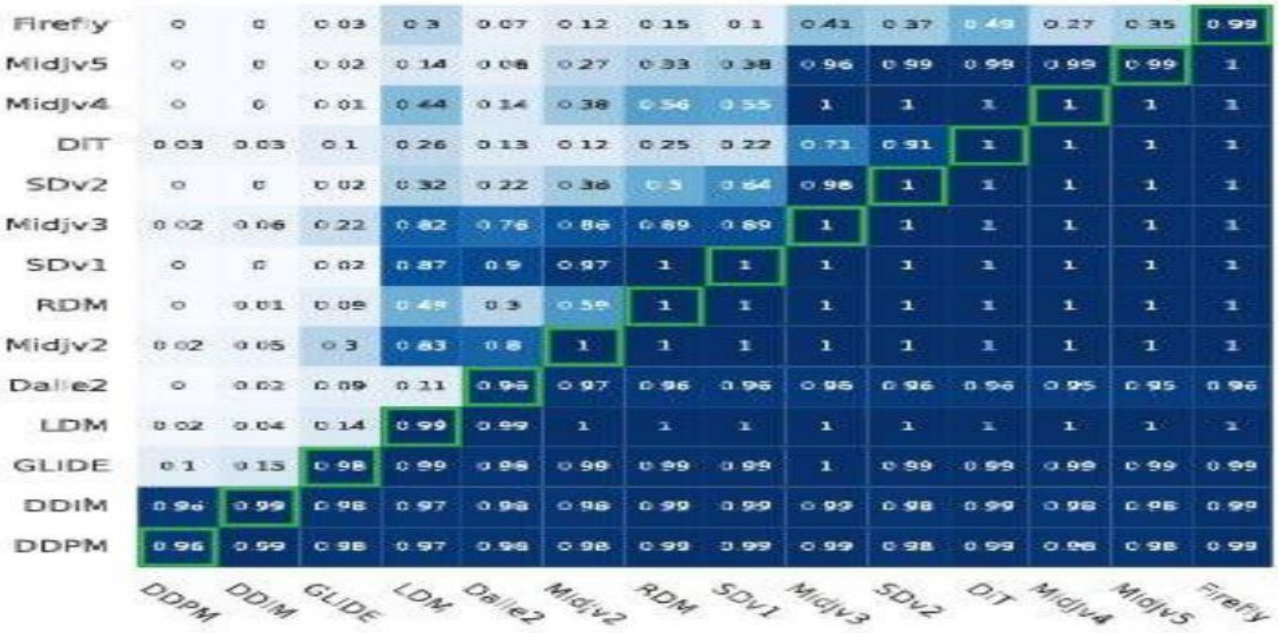
Dataset is customized in this paper; they collect a dataset of 14 well-known generative models using different generating architecture and released between June 2020 and March 2023 for the fake images class (AI- generated images). They use LAION-400M as source of non-generated images. Dataset is composed of 570,221 images (405,862 for train, 48,057 for validation, 116,302 for test).

__Table 5__ The training sequence with the dataset size based on the publication date of each generator

Generation architecture	**Method/Dataset**	**Training set**	**Paper**	**Open-src**	**#**	**Train size**	**Val size**	**Test size**
Real images	LAION-400M		-	-		179,900	22,479	22490
Diffusion U-net	Denoising Diffusion Prob Model (DDPM)	LSUN	✓	✓	1	6,271	784	785
	Denoising Diffusion Implicit Model (DDIM)	LSUN	✓	✓	2	8,000	1,000	1,000
	GLIDE	Private	✓	✓	3	7,442	929	931
	DALL.E2 [3]	Private	✓		5	2,000	954	2,000
Diffusion + Decoder	Latent Diffusion (LDM)	LAION-400M	✓	✓	4	8,172	1,021	1,022
	Retrieval-Augmented Diffusion (RDM)	LAION-400M	✓	✓	7	8,528	1,066	1,066
	Stable Diffusion v1 [4]	LAION-2B	✓	✓	8	34,508	3,807	3,838
	Stable Diffusion v2 [4]	LAION-5B	✓	✓	10	35,997	4,000	4,000
Diffusion U-Vit	Diffusion with Transformers (DIT)	ImageNet	✓	✓	11	3,199	400	401
Unknown (product release)	Midjourney v2 [2]	Unknown	✗	✗	6	42,875	5,358	5,369
	Midjourney v3 [2]	Unknown	✓	✓	9	70,035	8,754	8,755
	Midjourney v4 [2]	Unknown	✗	✗	12	100,000	10,000	76,122
	Midjourney v5 [2]	Unknown	✗	✗	13	63,310	7,914	7,918
	Adobe Firefly [24]	Unknown	✗	✗	14	15,525	2,070	3,105

They progressively train a binary classifier with a cross-entropy loss to distinguish between naturally sourced "real" images and those generated by AI. following best practices, which show that a simple classifier can generalize across generators. using a common CNN architecture, ResNet-50, pre-trained on ImageNet as the backbone for the detector model. The training sequence follows Table 1, simulating the real-world release dates of generative models. Each detector training step in the sequence continues from the previous model weights and incorporates all historical images seen to date. Release dates are determined by paper publication date, service launch announcement, or public release of model weights as relevant for the generative source. Non-generated images from LAION-400M are consistently included in training, ensuring a diverse dataset. A class-balanced random sampler maintains distribution balance between generated and non-generated images during each training stage. Augmentations, enhances generalization, including random 256×256 cropping, Gaussian blur, grayscale transformation, and invisible watermarks. These augmentations are applied with specified probabilities during training. For evaluation, a center crop of 256×256 is utilized without any augmentation.

Training performance on N models and testing on the (N + 1) th reveals high accuracy scores (0.98 or above) for methods like DDIM, Midjourney, RDM, Stable Diffusion, and DiT, indicating early capture of relevant features. DALL·E 2 in the 5th model's accuracy scores vary, with GLIDE achieving 15% accuracy, suggesting potential calibration needs. The last model, Adobe Firefly, exhibits relatively low accuracy even when trained on all previous models.



	DDPM	DDIM	GLIDE	LDM	Dalle2	Midjv2	RDM	SDv1	Midjv3	SDv2	DiT	Midjv4	Midjv5	Firefly
Firefly	0	0	0.03	0.3	0.07	0.12	0.15	0.1	0.41	0.37	0.49	0.27	0.35	0.99
Midjv5	0	0	0.02	0.14	0.08	0.27	0.33	0.38	0.96	0.99	0.99	0.99	0.99	1
Midjv4	0	0	0.01	0.44	0.14	0.38	0.56	0.55	1	1	1	1	1	1
DIT	0.03	0.03	0.1	0.26	0.13	0.12	0.25	0.22	0.71	0.91	1	1	1	1
SDv2	0	0	0.02	0.32	0.22	0.36	0.5	0.64	0.98	1	1	1	1	1
Midjv3	0.02	0.06	0.22	0.82	0.76	0.86	0.89	0.89	1	1	1	1	1	1
SDv1	0	0	0.02	0.87	0.9	0.97	1	1	1	1	1	1	1	1
RDM	0	0.01	0.09	0.49	0.3	0.59	1	1	1	1	1	1	1	1
Midjv2	0.02	0.05	0.3	0.83	0.8	1	1	1	1	1	1	1	1	1
Dalle2	0	0.02	0.09	0.11	0.96	0.97	0.96	0.96	0.96	0.96	0.96	0.95	0.95	0.96
LDM	0.02	0.04	0.14	0.99	0.99	1	1	1	1	1	1	1	1	1
GLIDE	0.1	0.15	0.98	0.99	0.98	0.99	0.99	0.99	1	0.99	0.99	0.99	0.99	0.99
DDIM	0.96	0.99	0.98	0.97	0.98	0.98	0.99	0.99	0.99	0.98	0.99	0.98	0.98	0.99
DDPM	0.96	0.99	0.98	0.97	0.98	0.98	0.99	0.99	0.99	0.98	0.99	0.98	0.98	0.99

Figure 9 *Accuracy results of our online detector on the synthetic images.*

Another technique used for detecting the manipulated visual content discussed in **Detection of AI-Created Images Using Pixel-Wise Feature Extraction and Convolutional Neural Networks** [25] The paper discusses the increasing use of generative artificial intelligence, particularly transformer-based models like ChatGPT, DALL E, Stable Diffusion, etc., for creating realistic images. The authors propose a method for AI image detection using a complex feature extraction based on two parallel deep learning processes utilizing pixel-wise feature extraction techniques such as Photo Response Non-Uniformity (PRNU) and Error Level Analysis (ELA). PRNU is a kind of noise used for source camera identification, AI images should have no PRNU at all, while ELA is a special image (or pattern) that detects irregular errors in JPEG-coded images, detect editing in images, ELA has also been applied for forged face detection. The paper aims to contribute to the detection of AI generated images and discusses potential applications in image forensics, addressing concerns related to misinformation and deep fake technology.

In this work, several models are used to create the dataset, and other different models are tried for final tests. The dataset of the paper comprises two classes: AI-generated images and real camera photographs. AI images were created using DALL E, Stable Diffusion, and OpenArt, ensuring photorealism through visual inspection.

Table 6 *Numerical results for both methods.*

Method (Pattern type)	Accuracy	Precision	Recall	F1 Score
PRNU	0.95	0.93	0.97	0.95
ELA	0.98	0.97	0.99	0.98

Real photos were randomly selected from databases like Dresden Image Database, VISION dataset, and authors' provided images, covering various cameras and smartphones. The initial dataset had 459 images in each class, totaling 918. Later, an extended dataset of 1252 images were tested. the direct application of CNNs could not be very useful (besides the experience from the Deep fake case). For this reason, pixel-wise feature extraction was used**.** at first discussed *PRNU Extraction* which comes from the different light sensitivity of the different pixels, due to manufacturing imperfections. PRNU is seen as a multiplicative noise that responds to the following equation:

$$\boldsymbol{Im_{out}} = (\boldsymbol{I_{ones}} + \boldsymbol{Noise_{cam}}).\,\boldsymbol{Im_{in}} + \boldsymbol{Nosie_{add}}$$

Note that, in this application, researchers will always compute PRNU fingerprints with a single image (N = 1) both for AI-generated and real images. From each image, a centered square 512 × 512 region is extracted to work with smaller images and to avoid logos or visible watermarks, and PRNU is computed from the sub-image. The results of this process are noise-like images. Secondly discussed Error Level Analysis (***ELA***) pattern is computed to detect irregular distributions of quantization noise.

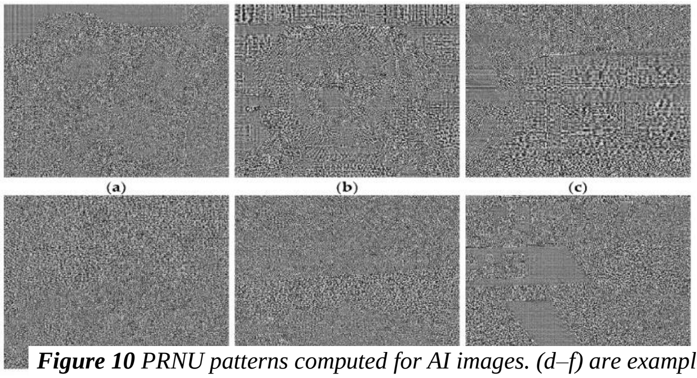


Figure 10 *PRNU patterns computed for AI images. (d–f) are examples of PRNU patterns for real images.*

$$\boldsymbol{ELA_{img}} = \boldsymbol{img} - \boldsymbol{JPEG^{-1}}[\boldsymbol{JPEG}(\boldsymbol{img}, \boldsymbol{95}\%)]$$

If we are facing an edited image, an irregular pattern with different intensities will appear**.** Note that, in this case, patterns are color images. For PRNU computation, images are converted to grayscale before any processing.

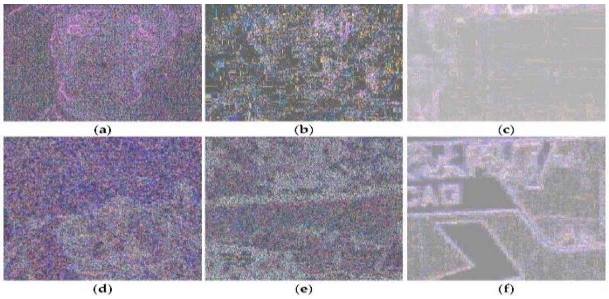


Figure 11 *ELA patterns computed for AI images. (d–f) are examples of ELA patterns for real images.*

Thirdly *CNNs—Convolutional Neural Networks*: CNNs are a cascade of convolutional (or linear filtering) stages accompanied by others of non-linear activation, normalization, and decimation. These stages extract high-level features from low-level. The final result is a numerical vector of as many components as classes to be recognized. The SoftMax normalization (the most frequently used at the final stage of CNNs) makes vector coefficients lie in the range of 0.0-1.0. The training algorithm is Stochastic Gradient Descent with Momentum (SGDM). In this paper, a previous image-to-image transformation is performed that acts as a pixel-wise feature extraction. CNN nets were trained and tested for both types of features extraction. In both cases, a good result is achieved: accuracy is 0.95 for PRNU and 0.98 for ELA. Both trainings were performed with 100 epochs. Training time is longer for the ELA case (167 min versus 109); this is reasonable because ELA images are color ones with three times more information.

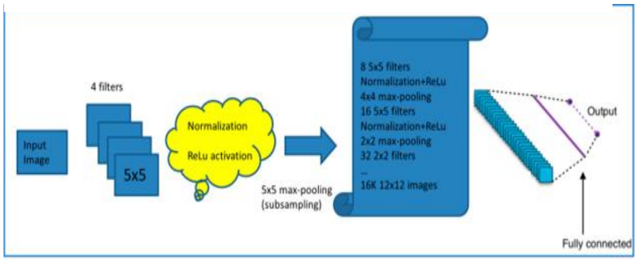


Figure 12 *CNN structure used.*

To understand more the generalization approach with the GANs models we had deep understood to **CNN-generated images are surprisingly easy to spot** [26]. The study investigates the possibility of creating a universal detector capable of discerning real images from those generated by various Convolutional Neural Network (CNN) models, regardless of their architecture or dataset. To explore this, they compile a dataset comprising fake images from 11 diverse CNN-based image generator models, including ProGAN, StyleGAN, BigGAN, and others, chosen to represent commonly used architectures. Despite differences in architectures, datasets, and training methods, the study reveals that a standard image classifier trained exclusively on one specific CNN generator, ProGAN, demonstrates remarkable generalization to unseen architectures, datasets, and even the newly released StyleGAN2. These findings hint at the intriguing possibility that CNN-generated images may possess shared systematic flaws that hinder realistic image synthesis.

Table 7 *Generation models. Assessment of forensic classifiers on a variety of CNN-based image generation methods.*

Family	Method	Image Source	# Images
Unconditional GAN	ProGAN	LSUN	8.0k
	StyleGAN	LSUN	12.0k
	BigGAN	ImageNet	4.0k
Conditional GAN	CycleGAN	Style/object transfer	2.6k
	StarGAN	CelebA	4.0k
	GauGAN	COCO	10.0k
Perceptual loss	CRN	GTA	12.8K
	IMLE	GTA	12.8K
Low-level vision	SITD	Raw camera	360
	SAN	Standard SR benchmark	440
Deepfake	FaceForensics++	Videos of faces	5.4k

To assess the transferability of classifiers trained to detect CNN-generated images, the researchers gather a dataset consisting of images produced by various CNN models, focusing on 11 synthesis models spanning a variety of architectures, datasets, and losses. They specifically examine the performance of classifiers trained on a single model, ProGAN, to detect fake images across diverse CNN generators. This choice represents real-world detection scenarios where the diversity or number of models for generalization is uncertain during training. The study employs ProGAN due to its capacity to generate high-quality images and its straightforward convolutional network structure and constructs a sizable dataset consisting solely of ProGAN generated images alongside real images, including 20 ProGAN models trained on different LSUN object categories.

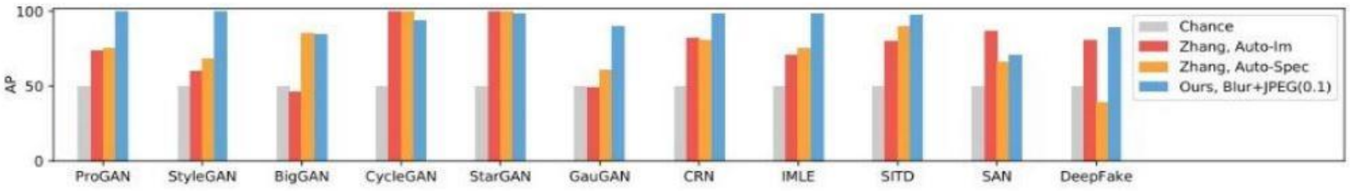


Figure 13 *Models Comparison*

Table 8 *Cross-generator generalization results*

Family	**Name**				**Individual test generators**											**total**
		Train	**Input**	**No. Class**												
					ProGAN	**StyleGAN**	**BigGAN**	**CycleGAN**	**StarGAN**	**GauGAN**	**CRN**	**IMLE**	**SITD**	**SAN**	**DeepFake**	**MAP**
Ours	DIP	ProGAN DIP	RGB	-	62.0	52.3	61.7	62.3	100.	49.0	98.2	38.6	92.8	93.1	63.1	70.3
	2-class	ProGAN	RGB	2	99.9	78.3	66.4	88.7	87.4	87.4	94.0	97.3	85.2	52.9	58.1	81.3
	4-class	ProGAN	RGB	4	99.8	87.0	74,0	93.2	92.3	94.1	95.8	97.5	87.8	58.5	59.6	85.4
	10% data	ProGAN	RGB	20	100	95.5	85.5	94.1	93.2	97.1	96.8	994	885	381	65	878
	20% data	ProGAN	RGB	20	100.	96.8	85.9	95.9	93.6	97.9	98.7	99.5	90.2	61.8	65.2	89.6
	40% data	ProGAN	RGB	20	100.	97.8	87.5	96.0	95.3	98.1	98.2	99.3	91.2	61.4	67.9	90.2
	80% data	ProGAN	RGB	20	100.	98.1	88.1	964	95.4	98.0	98.9	99.4	93.0	63.8	65.1	90.6

Furthermore, the researchers detail their experimental approach, emphasizing the training of a "real-or-fake" classifier on the ProGAN dataset and assessing its generalization to other CNN-synthesized images. They utilize ResNet-50 pretrained with ImageNet, training it in a binary classification setting. The training procedure involves using the Adam optimizer with specific parameters and adapting the learning rate based on validation accuracy. Additionally, the study evaluates the choice of training architecture by comparing results with models trained solely on BigGAN, considering both image generation and data augmentation techniques. Notably, their method achieves a 92% Area Under the Curve (AUC) on the recently released StyleGAN3 model, indicating promising outcomes for universal detection of CNN-generated images.

Chapter 3: Methodology

After reviewing similar products and research papers in the market, detailed in Chapter 2, we identified critical limitations in existing approaches, such as their inability to handle specific image formats and their challenges in accurately classifying images as real or AI-generated, especially when confronted with images from unseen generating models during training. These shortcomings underscored the need for innovative solutions. To address these issues, our approach focuses on several enhancements aimed at significantly improving model performance. Firstly, we have curated a comprehensive dataset featuring images from 25 diverse generators, encompassing

13 GAN-based, 7 diffusion-based, and 5 other miscellaneous generators. This dataset ensures that our model can effectively detect fake images across a wide spectrum of generating techniques.

Secondly, our model architecture utilizes a Capsule-Forensics network with ten primary capsules, leveraging dynamic routing to maintain spatial relationships and capture subtle manipulations inherent in AI-generated images. This design not only enhances the model's ability to discern between real and fake images but also mitigates the performance degradation observed with unseen generating models.

Additionally, we employ rigorous training techniques, including regularization with random noise and dropout, along with optimized image preprocessing and binary classification through dynamic routing and softmax layers. These methodological advancements collectively empower our model to deliver more accurate and reliable detection of AI-generated content, setting a new standard in the field.

3.1. Data Collection and Preparation

To address the limitations identified in previous approaches and improve model generalization, a diverse dataset of both real and AI-generated images from various sources will be used, we will use a dataset named ArtiFact [27], it is a large-scale dataset replete with diverse generators including GAN, Diffusion, fully manipulating, and partially manipulating images, many categories including human, human faces, animal, animal faces, places, vehicles, art, and many other real-life objects, comprising diverse generators, real-world impairments, object categories, and real-world challenges. Moreover, the proposed classification model addresses social platform impairments and effectively detects synthetic images from both seen and unseen generators, With the rapid development of sophisticated generative models, it is necessary to train include as much as possible variants in a detector’s training set. The dataset specifically includes 13 GANs, 7 Diffusion, and 5 other miscellaneous generators. On the other hand, in terms of synthetics, there are 20 fully manipulating and 5 partially manipulating generators, thus providing a broad spectrum of diversity in terms of generators used. The distribution of real and fake data with different sources is shown in Figure.5 and Figure.6, respectively. The dataset contains a total of 2,496,738 images, comprising 964,989 real images and 1,531,749 fake images. The most frequently occurring categories in the dataset are human, human faces, animal, animal faces, places, vehicles and art.

Figure 15 *Distribution of different methods for fake class.*

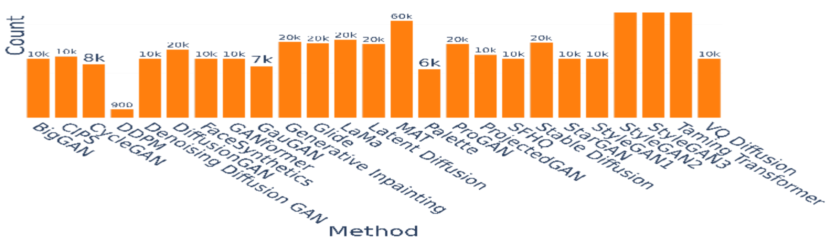
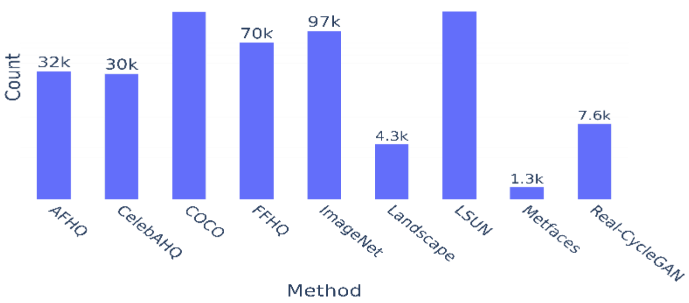


Figure 14 *Distribution of different methods for real class.*



3.2. Proposed Model

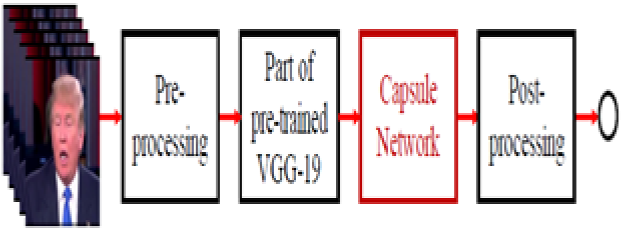


Figure 16 *Capsule-Forensics pipeline.*

We had chosen a model architecture that is suitable for our objectives and our plan of future work, believing that with a good fine-tuning we would be able to achieve the expected performance. In the next section we will discuss all the details step by step. We studied the use of a capsule network to detect fake images [28]. Capsule network is not the new term as it was first introduced in 2011 by Hinton et al. [29] They argued that CNNs have limited applicability to the "inverse graphics" problem and introduced a more robust architecture comprising several capsules. The agreements between low- and high-level capsules that encode the hierarchical relationships between objects and their parts with pose information enables a capsule network to preserve more information than a CNN while using only a fraction of the data used by a CNN. The pipeline of the Capsule-Forensics method is illustrated in Fig.12. The pre-processing task depends on the input. Each image is divided into patches. There is no strict requirement about the size of the input image. In general, the larger the input, the better the result, at the cost of more computational power. Using an image size of 300×300 as it is an even number (making it easy to perform cropping and scaling) and large enough to provide sufficient information for detecting fake content. The pre-processed image then passes through a part of the VGG-19 network pre-trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset before entering the capsule network. The VGG-19 network is used from the first layer to the third max pooling layer, which is not too deep to obtain biases from the object detection task which is the original purpose of this pre-trained network). This VGG-19 part is equivalent to the CNN part before the primary capsules in the design of the original capsule network. Using a pre-trained CNN as a feature extractor rather than training it from scratch provides the benefit of using it to guide the training and to reduce overfitting as well as that of transfer learning. The final part is the post processing unit, which works in accordance with the pre-processing one, the scores of the extracted patches are averaged. This average score is the final output. The capsule network includes 10 primary capsules and two output capsules (“real” and “fake”), as illustrated in Fig. 13 in the next page. Three capsules are typically used for light networks which require less memory and computation.

Each primary capsule is divided into three parts: a 2D convolutional part, a statistical pooling layer, and a 1D convolutional part. The statistical pooling layer has been proven to be effective in the forensics task. Moreover, it helps make the network independent of the input image size. This means that one Capsule-Forensics architecture can be applied to different problems with different input sizes without having to redesign the network. The mean

it helps make the network independent of the input image size. This means that one Capsule-Forensics architecture can be applied to different problems with different input sizes without having to redesign the network. The mean

and variance of each filter are calculated in the statistical pooling layer. After going through the following 1D convolutional part, it is sent through dynamic routing to the output capsules. The final result is calculated on the basis of the activation of the output capsules. The dynamic routing algorithm is used to calculate agreement between the features extracted by the primary capsules. Agreement is dynamically calculated at run-time and the results are routed to the appropriate output capsule (real or fake one for binary classification). The output probabilities are determined based on the activations of the output capsules. This dynamic routing algorithm differs from the classical fusion one in that it combines classification outputs from different classifiers.

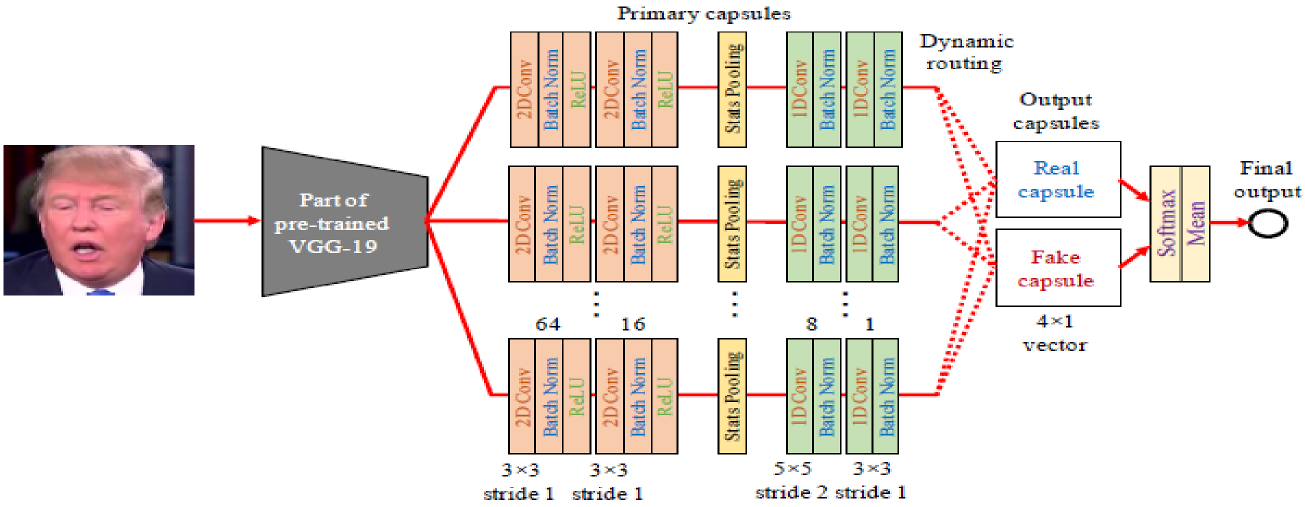
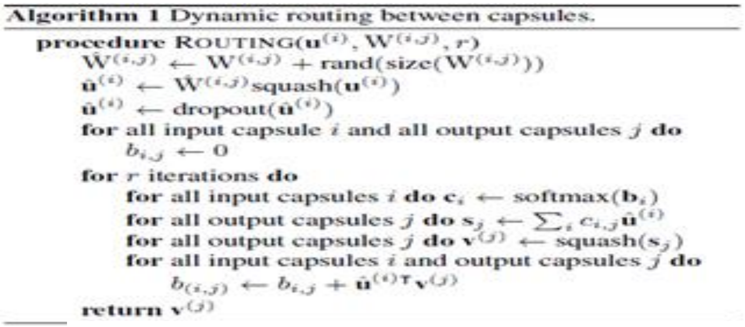


Figure 17 *Capsule-Forensics architecture.*

introducing two regularizations: adding random noise to the routing matrix and adding a dropout operation. Furthermore, a squash function is applied to the output vector of each primary capsule **u**$^{(i)}$ before routing to normalize it, which helps stabilize the training process. The squash function is used to scale the vector magnitude to unit length. Let us call the real and fake vector capsules **v** $^{(1)}$ and **v** $^{(2)}$ respectively. **W**$^{(i,j)}$ is the



```
Algorithm 1 Dynamic routing between capsules.
  procedure ROUTING($\mathbf{u}^{(i)}, \mathbf{W}^{(i,j)}, r$)
      $\hat{\mathbf{W}}^{(i,j)} \leftarrow \mathbf{W}^{(i,j)} + \mathrm{rand}(\mathrm{size}(\mathbf{W}^{(i,j)}))$
      $\hat{\mathbf{u}}^{(i)} \leftarrow \hat{\mathbf{W}}^{(i,j)}\mathrm{squash}(\mathbf{u}^{(i)})$
      $\hat{\mathbf{u}}^{(i)} \leftarrow \mathrm{dropout}(\hat{\mathbf{u}}^{(i)})$
      for all input capsule $i$ and all output capsules $j$ do
          $b_{i,j} \leftarrow 0$
      for $r$ iterations do
          for all input capsules $i$ do $\mathbf{c}_i \leftarrow \mathrm{softmax}(\mathbf{b}_i)$
          for all output capsules $j$ do $\mathbf{s}_j \leftarrow \sum_i c_{i,j}\hat{\mathbf{u}}^{(i)}$
          for all output capsules $j$ do $\mathbf{v}^{(j)} \leftarrow \mathrm{squash}(\mathbf{s}_j)$
          for all input capsules $i$ and output capsules $j$ do
              $b_{(i,j)} \leftarrow b_{i,j} + \hat{\mathbf{u}}^{(i)\mathsf{T}}\mathbf{v}^{(j)}$
      return $\mathbf{v}^{(j)}$
```

Figure 18 *Dynamic routing algorithm.*

matrix used to route $\mathbf{u}^{(i)}$ to $\mathbf{v}^{(i)}$ and r is the number of iterations. The dynamic routing algorithm follow fig.15. In practice, to stabilize the training process, the random noise should be sampled from a normal distribution (N (0; 0:01)), the dropout ratio should not be greater than 0.05 and two iterations (r = 2) should be used in the dynamic routing algorithm. The two regularizations are used along with random weight initialization to increase the level of randomness, which helps the primary capsules to learn with different parameters, then apply softmax. the final results are the means of all softmax outputs using the cross-entropy loss function and the Adam optimizer in the training phase.

$$squash(\mathbf{u}) = \frac{\|\mathbf{u}\|_2^2}{1 +| \|\mathbf{u}\|_2^2} \frac{\mathbf{u}}{\|\mathbf{u}\|_2}$$

Figure 19 *Squash function.*

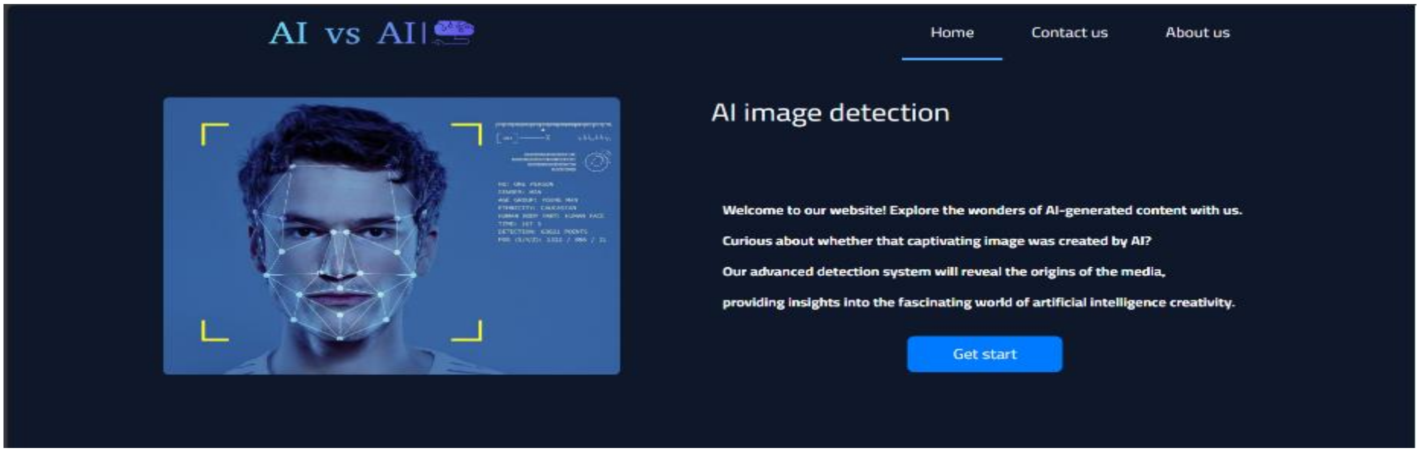
3.3. User Interface

In the realm of user interface design, our decision to develop an AI Image Detection Platform as a website rather than a mobile application stems from several key advantages. Primarily, websites offer unparalleled accessibility, eliminating the need for users to download or install software, which can be a significant barrier to entry. Additionally, websites are inherently platform-independent, functioning seamlessly across different devices and operating systems without the compatibility issues often encountered with mobile apps. This cross-platform compatibility is crucial for reaching a broader audience and providing a consistent user experience regardless of the device being used. Moreover, websites can be updated more easily than mobile apps, allowing for quicker deployment of new features and improvements without requiring users to manually update their software. This flexibility and ease of maintenance contribute to a more robust and agile development process.

3.3.1. Website Development

The AI Image Detection Platform website was meticulously crafted to offer an intuitive and highly efficient user interface, leveraging a comprehensive suite of modern web technologies designed to significantly enhance the overall user experience. The development process commenced with the design of a straightforward and user-friendly interface, prominently featuring a 'Get Started' button that streamlines the initiation process. This allows users to effortlessly begin their interaction with the platform. Users can either upload images for analysis or simply drag and drop them into the designated area, with functionalities to upload additional files consecutively or cancel the upload if necessary. This feature improves the operation of testing multiple images by making it faster and easier for users, encouraging them to spend more time on the website. The core functionality of the platform revolves around the 'Check Model' button, which activates the AI model to predict whether the uploaded image is AI-generated. Upon completion of the analysis, the platform displays detailed image information, including format, bits per pixel, color type, dimensions, interlacing, and resolution. To provide faster responses, we also offer a caching service for previously checked images. The website's structure and styling were meticulously achieved using HTML and CSS, ensuring a clean, accessible, and aesthetically pleasing design. Bootstrap was integrated to guarantee a responsive design and a consistent look across various devices, thus enhancing the overall user experience. JavaScript was employed to introduce interactive elements, making the interface more dynamic and engaging. The backend of the website is powered by Flask, a lightweight WSGI web application framework in Python, which efficiently handles server-side logic and image analysis requests. By integrating these advanced technologies, the website ensures a smooth and efficient process for users to determine the authenticity of their images, clearly exemplifying the advantages of a web-based solution over a mobile application

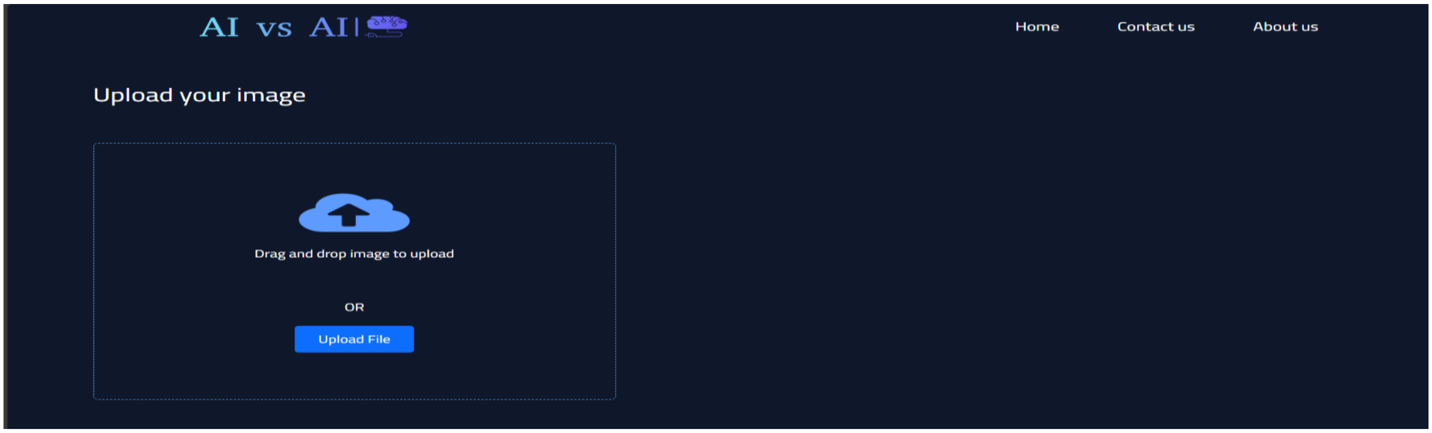
3.3.2. Website Demo



AI image detection

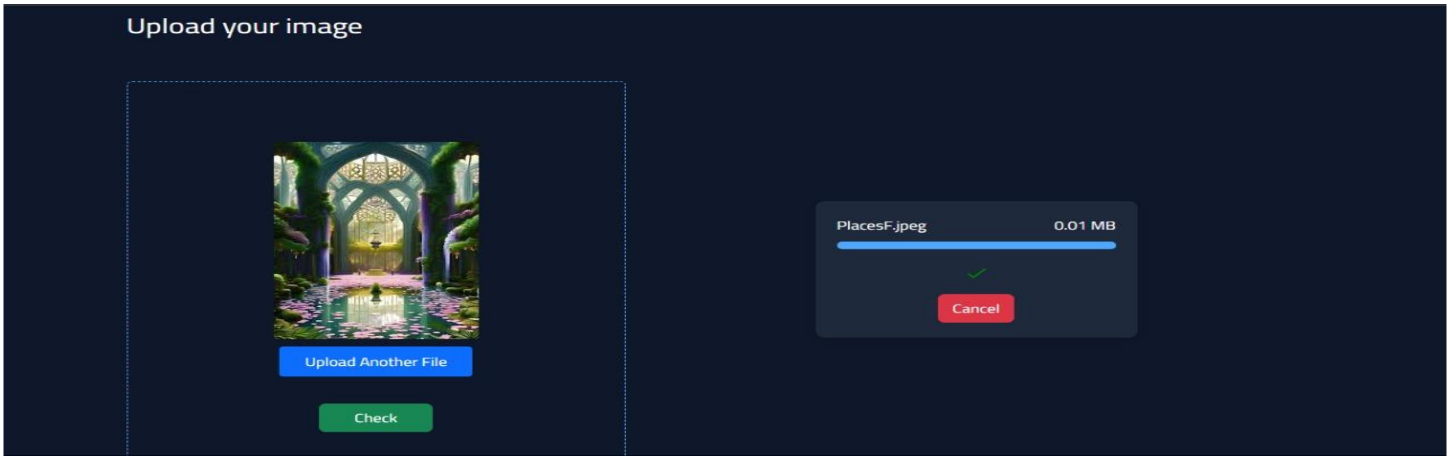
Welcome to our website! Explore the wonders of AI-generated content with us.
Curious about whether that captivating image was created by AI?
Our advanced detection system will reveal the origins of the media,
providing insights into the fascinating world of artificial intelligence creativity.

After Pressing Get Stat Button:



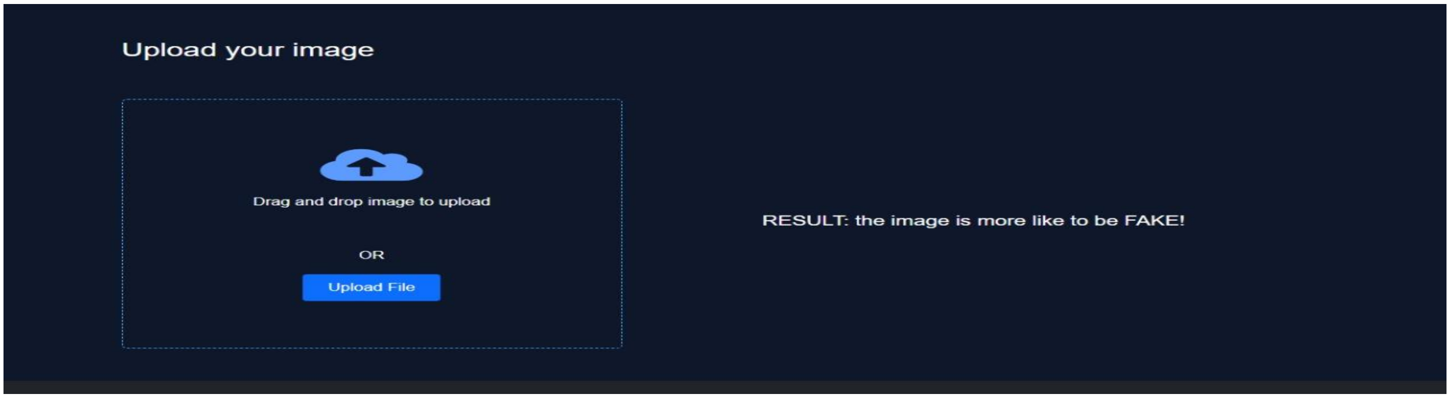
Upload your image

Then press Upload File Button:



Upload your image

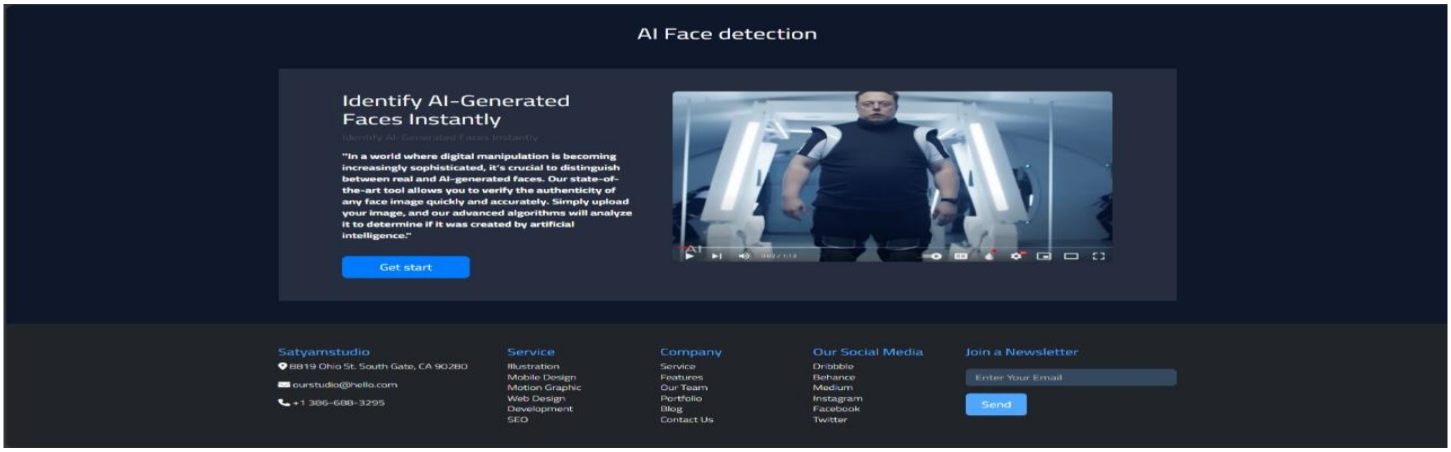
Press Check Button to check if the image is Real or AI-generated:



Upload your image

RESULT: the image is more like to be FAKE!

Face Detection:



AI Face detection

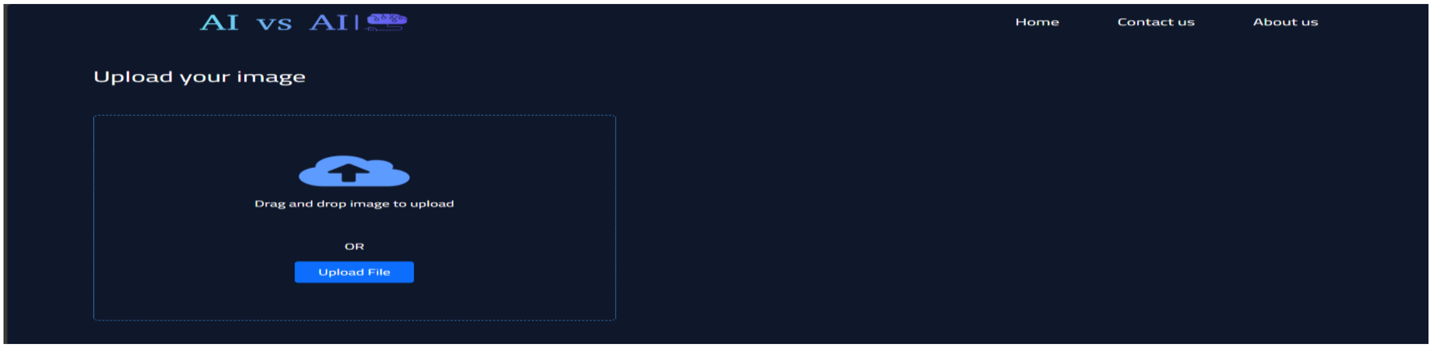
Identify AI-Generated Faces Instantly

Identify AI-Generated Faces Instantly

"In a world where digital manipulation is becoming increasingly sophisticated, it's crucial to distinguish between real and AI-generated faces. Our state-of-the-art tool allows you to verify the authenticity of any face image quickly and accurately. Simply upload your image, and our advanced algorithms will analyze it to determine if it was created by artificial intelligence."

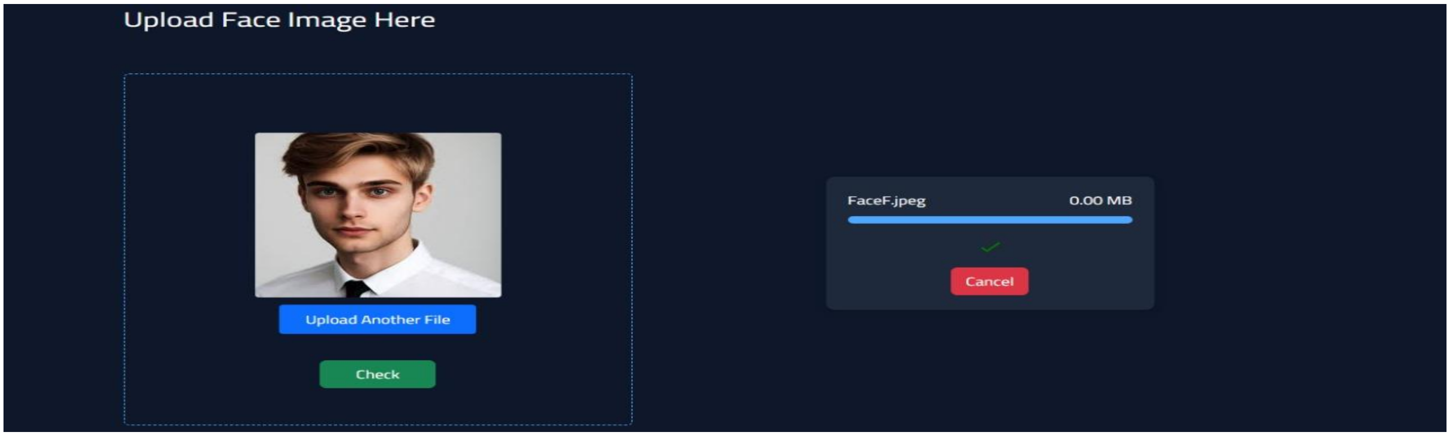
Satyamstudio
8819 Ohio St. South Gate, CA 90280
ourstudio@hello.com
+1 386-688-3295

After Press Get Start Button:



Upload your image

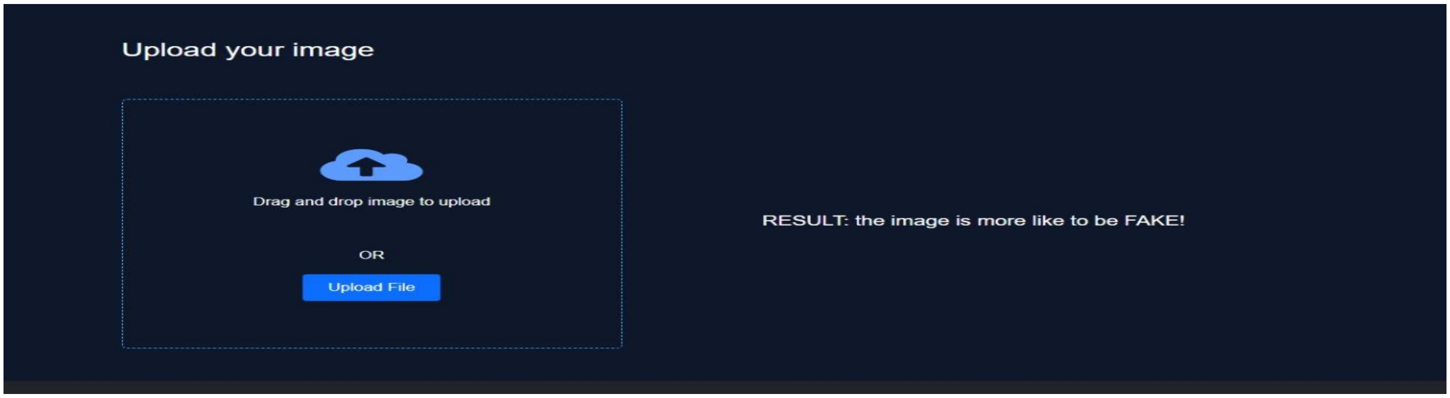
Then Press Upload File Button:



Upload Face Image Here

FaceF.jpeg 0.00 MB

Press Check Button to check if the Face image is Real or AI-generated:

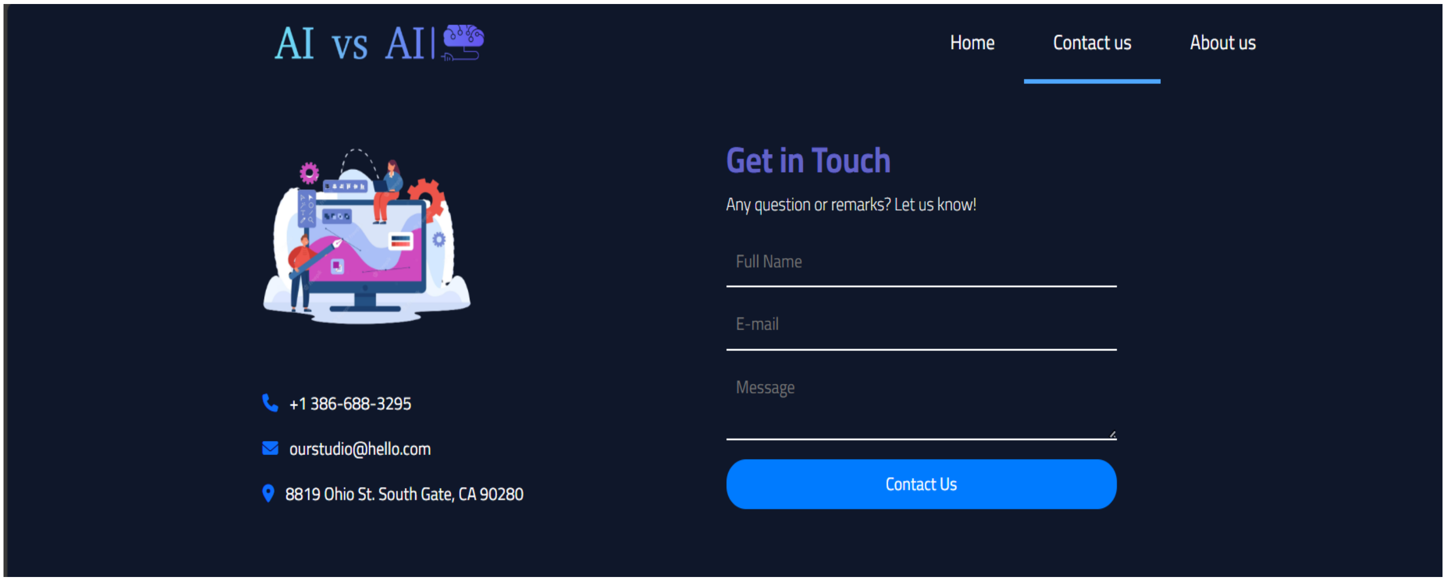


Upload your image

Drag and drop image to upload

OR

RESULT: the image is more like to be FAKE!



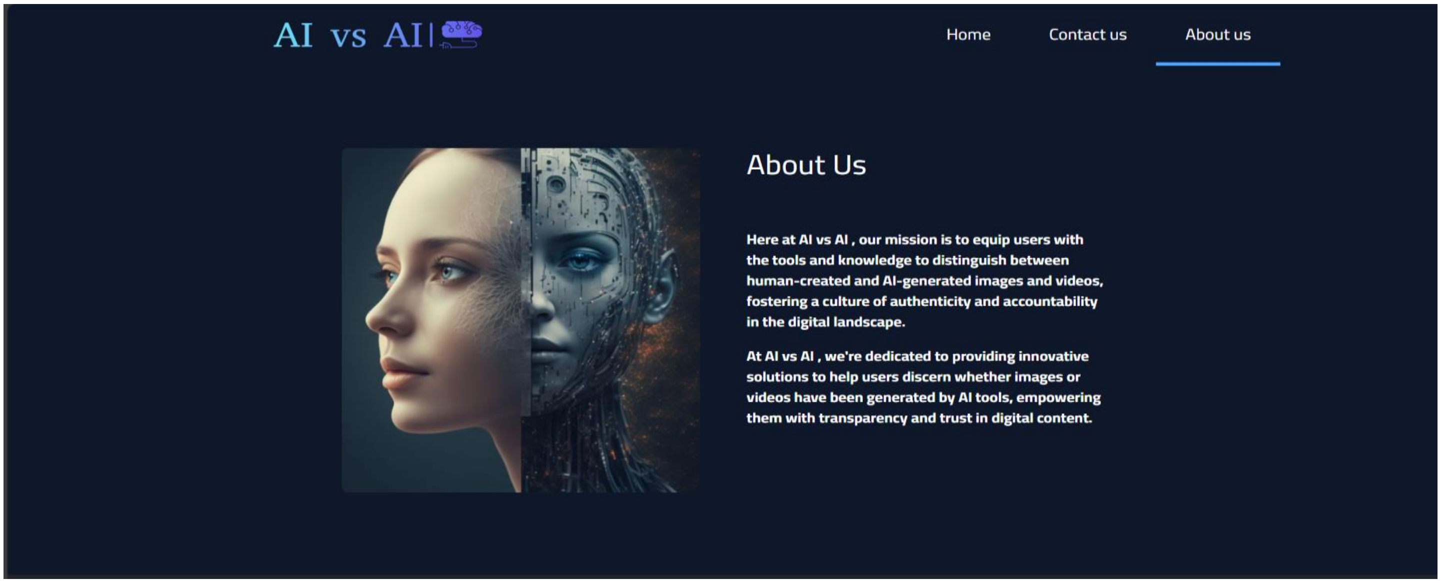
Get in Touch

Any question or remarks? Let us know!

☎ +1 386-688-3295

ourstudio@hello.com

8819 Ohio St. South Gate, CA 90280



About Us

Here at AI vs AI , our mission is to equip users with the tools and knowledge to distinguish between human-created and AI-generated images and videos, fostering a culture of authenticity and accountability in the digital landscape.

At AI vs AI , we're dedicated to providing innovative solutions to help users discern whether images or videos have been generated by AI tools, empowering them with transparency and trust in digital content.

Chapter 4: Results and Analysis

Our approach was rigorously evaluated using a diverse artifact dataset comprising images from 25 different generators, covering comprehensive categories to ensure a robust and representative sample. The total dataset included 1,936,738 samples, with 20% reserved for testing. The model achieved an accuracy of 86% during training and 86.24% during testing through 17 epochs of training. These results highlight the model's consistency and effectiveness in detecting AI-generated content across various generator types. The training process utilized a learning rate of 1e-4 and the Adam optimizer, ensuring optimal convergence and performance. Our model architecture incorporated 10 primary capsules and employed dropout with a rate of 0.3. Additionally, 2 iterations were used in the dynamic routing algorithm to enhance the model's robustness and accuracy.

We also try the same architecture with different datasets, each has its own variety of generators and the number of categories included to have insight about the model performance on small scale datasets that had been used in research papers and mentioned earlier in the literature review to show how much good the model comparing to the others exist.

We try subset of artifact dataset, we utilized 15,000 samples from the cycle generator and input them into our model achieving an accuracy of 99.88% on the training set and 88.39% on the test set. Additionally, we fed real images from the COCO dataset and generated images using the Taming Transformer on the COCO dataset into our model. The performance metrics were 93.37% for training and 91.85% for testing, upon combining these datasets, our model achieved an overall accuracy of 88%, indicate that the model is better at different datasets and types of generated images.

For the CIFake dataset, real images collected from CIFAR-10 dataset, and fake images generated from the equivalent of CIFAR-10 using Stable Diffusion version 1.4. The dataset contains 100,000 images for training (50k real and 50k fake) and 20,000 for testing (10k real and 10k fake). Our approach was evaluated on this dataset, achieving an accuracy of 97.2% on the training set and 96% on the test set.

Our approach was also evaluated using Faces dataset comprises 130,000 real faces from the Flickr dataset collected by Nvidia, along side130,000 fake faces generated by StyleGAN and Stable Diffusion. Our model was evaluated on this dataset, achieving an impressive accuracy of 95% on the training set and 93.5% on the test set.

These results highlight the model's robustness and effectiveness in distinguishing between real and AI-generated content. The results gave us a good imputation about the model performance as shown in table

Dataset	Generators No.	Categories Varity	Training Accuracy	Testing Accuracy
Artifact **(ours)**	25	Comprehensive	86%	86.24%
CiFake	1	Limited	97.2%	96%
260k Real and Fake Faces (GAN)	1	Absent	95%	93.5%
Subset of Artifact (cycleGan Generator)	1	Comprehensive	0.9988	0.8839
Subset of Artifact (Taming Transformer on the COCO dataset	1	Comprehensive	93.37%	91.85%
Combined	2	Comprehensive	88%	88%

***Table 9** Results Of Capsule Model on different Datasets*

The following table summarizes the performance of our model and others across different datasets. In the Artifact dataset, a CNN model achieved an accuracy of 87%. Our approach, evaluated on the Artifact dataset, utilized a Capsules model and achieved an accuracy of 86.24%. Transitioning to the Faces Dataset, a Capsule model achieved a higher accuracy of 92%. On the Faces Dataset, our Capsules model demonstrated improved performance with an accuracy of 93.5%. These results illustrate the comparative effectiveness of different models in detecting AI-generated content across specific datasets, highlighting the strengths of Capsule models particularly in the context of face image detection.

Dataset	Model	Performance
Artifact [27]	CNN Model	87%
Faces Dataset [28]	Capsule Model	92%
Artifact(**ours**)	Capsules Model	86.24%
Faces Dataset(**ours**)	Capsules Model	93.5%

***Table 10** Performance Comparison between our model and other*

Chapter 5: Conclusion and Recommendation

We have addressed the significant limitations of existing approaches in the detection of AI-generated images, particularly those stemming from their inability to handle specific image formats and challenges in accurately classifying images from unseen generating models. Our innovative approach has focused on several key enhancements to significantly improve model performance.

We curated a comprehensive dataset featuring images from 25 diverse generators, encompassing 13 GAN-based, 7 diffusion-based, and 5 other miscellaneous generators. This extensive dataset has enabled our model to effectively detect fake images across a wide spectrum of generating techniques. Our model architecture utilizes a Capsule-Forensics network with ten primary capsules and employs dynamic routing to maintain spatial relationships and capture subtle manipulations inherent in AI-generated images. This design enhances the model's ability to discern between real and fake images, mitigating performance degradation observed with unseen generating models.

Additionally, we employed rigorous training techniques, including regularization with random noise and dropout, optimized image preprocessing, and binary classification through dynamic routing and softmax layers. These methodological advancements collectively empower our model to deliver more accurate and reliable detection of AI-generated content.

We meticulously crafted the AI Image Detection Platform as a web-based solution to provide unparalleled accessibility, cross-platform compatibility, and ease of maintenance. This platform offers an intuitive and efficient user interface, significantly enhancing the overall user experience.

Our approach was rigorously evaluated using the Artifact dataset, achieving an accuracy of 86.24% during testing. We further evaluated our model on various datasets, demonstrating its robustness and effectiveness across different types of generated images. Notably, our model achieved impressive accuracy rates, such as 88.39% on the cycle generator subset, 91.85% on the COCO dataset, 96% on the CIFake dataset, and 93.5% on the Faces dataset.

The development of our AI Image Detection Platform aimed at accurately discerning fake images across diverse sources and categories was a formidable undertaking, bolstered by the implementation of a state-of-the-art Capsule Network model. Unlike traditional convolutional neural networks, Capsule Networks offer several distinct advantages in this project. They excel in capturing hierarchical relationships within data, which is particularly beneficial for detecting subtle variations between real and AI-generated images. This capability enhances the model's ability to generalize effectively across different contexts and image types, thus elevating standards in accuracy and reliability for image detection tasks. The significance of our Capsule Network-based approach cannot be overstated in the context of combating misinformation and safeguarding the integrity of digital content. By harnessing advanced neural network architectures like Capsule Networks, we not only address current challenges but also anticipate future demands in the rapidly evolving landscape of AI-generated imagery. Moving forward, to further advance the field of AI-driven content verification, we recommend expanding our capabilities beyond image detection. Future efforts could focus on developing AI systems capable of detecting manipulated videos and text, extending our platform's reach to address multimedia content authenticity comprehensively. Incorporating techniques such as deep learning for video analysis and natural language processing for text verification would be crucial in achieving this goal. Additionally, enhancing the performance and robustness of our model remains a priority. Strategies such as expanding the diversity and scale of our training dataset, integrating data from a wide array of sources and categories, and leveraging transfer learning from pretrained models are pivotal. Employing sophisticated techniques like data augmentation and adversarial training, coupled with continuous refinement through real-world feedback loops, will further fortify our model against emerging image manipulation techniques. Our commitment to advancing AI technologies underscores our dedication to developing systems that are not only accurate and reliable but also adaptive and resilient in the face of evolving challenges.

In conclusion, our research sets a new standard in the field of AI-generated content detection. The innovative use of a Capsule-Forensics network, comprehensive dataset curation, and rigorous training techniques have collectively resulted in a model that is highly accurate and reliable. Our web-based AI Image Detection Platform exemplifies the advantages of this approach, offering an accessible and user-friendly solution to the challenge of distinguishing between real and AI-generated images.

References

[1] "AI or Not," [Online]. Available: https://www.aiornot.com/.

[2] "Mid Journey," [Online]. Available: https://www.midjourney.com.

[3] "DALL.E.2," [Online]. Available: https://openai.com/dall-e-2.

[4] "stablediffusion," [Online]. Available: https://github.com/Stability-AI/stablediffusion.

[5] "Fake Image Detector," [Online]. Available: https://www.fakeimagedetector.com/.

[6] "Is It Ai," [Online]. Available: https://isitai.com/ai-image-detector/.

[7] "Content at scale," [Online]. Available: https://contentatscale.ai/ai-image-detector/.

[8] "githubFakeImageDetection," [Online]. Available: https://github.com/agusgun/FakeImageDetector.

[9] "Detecting-Images-Generated-by-Diffusers," [Online]. Available: https://github.com/davide-coccomini/Detecting-Images-Generated-by-Diffusers.

[10] W. H. K. W. W. L. a. P. Z. Ziyi Xi, "AI-Generated Image Detection using a Cross-Attention Enhanced Dual-Stream Network," *arXiv,* p. 8, 2023.

[11] [Online]. Available: https://alaska.utt.fr/#leaderboard.

[12] "DreamStudio," stability.ai, [Online]. Available: https://beta.dreamstudio.ai/generate.

[13] Davide Alessandro Coccomini, Andrea Esuli, Fabrizio Falchi and Claudio Gennaro, "Detecting Images Generated by Diffusers," *arXiv*, p. 8, 2023.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie and James Zitnick, "Microsoft coco: Common objects in context.," *European conference on computer vision ,* p. 740–755, 2014.

[15] K. Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky and Marc Najork, "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," *n Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval,* 2021.

[16] Samah S. Baraheem and Tam V. Nguyen., "AI vs. AI: Can AI Detect AI-Generated Images.," *Mdpi,* 2023.

[17] Simonyan, K. and Zisserman, A, "Very deep convolutional networks for large-scale image recognition," *arXiv* , p. 1409, 2014.

[18] Sun, J, Ren, S, Zhang, X and He, K, "Deep residual learning for image recognition," *arXiv,* 2015.

[19] Wojna, Z, Ioffe, S.; Shlens, Vanhoucke, V and Szegedy, C., "Rethinking the inception architecture for computer vision.," *arXiv,* 2015.

[20] F. X. Chollet, "Deep learning with depthwise separable convolutions.," *In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.*

[21] Weinberger, K.Q., van der Maaten, L, Liu, Z and Huang, G, "Densely connected convolutional networks," *arXiv,* 2016.

[22] Alemi, A, Vanhoucke, V, Szegedy, C and Ioffe, S., "Inception-v4, Inception-ResNet and the impact of residual connections on learning," *arXiv,* 2016.

[23] Richard Zhang, Ishan Jain and David C. Epstein, "Online Detection of AI-Generated Images. ," *arXiv,* 2023.

[24] "Adobe Firefly," [Online]. Available: https://firefly.adobe.com/.

[25] Fernando Martin-Rodriguez, Rocio Garcia-Mojon and Monica Fernandez-Barciela, "Detection of AI-Created Images Using Pixel-Wise Feature Extraction and Convolutional Neural Networks," *mdpi,* p. 10, 2023.

[26] Richard Zhang, Oliver Wang, Sheng-Yu Wang, Alexei A. Efros and Andrew Owens, "CNN-generated images are surprisingly easy to spot... for now," *ArXiv,* 2019.

[27] Najibul Haque Sarker, Bishmoy Paul, Md Awsafur Rahman, Shaikh Anowarul Fattah and Zaber Ibn Abdul Hakim, "ARTIFACT: A LARGE-SCALE DATASET WITH ARTIFICIAL AND FACTUAL IMAGES FOR," *arXiv,* 2023.

[28] Huy H. Nguyen, Junichi Yamagishi and Isao Echizen, "USE OF A CAPSULE NETWORK TO DETECT FAKE IMAGES AND VIDEOS," *arXiv,* p. 14, 2019.

[29] G. E. Hinton, A. Krizhevsky and S. D. Wang, "Transforming auto-encoders," *International Conference on Artificial Neural Networks (ICANN) Springer,* 2011.