National College of Ireland

Spotting the Synthetic: Vision Transformers vs Diffusion-Based Image Generation

MSc Research Project
Data Analytics

Brian Riera
Student ID: 23186771

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Brian Riera……

Student ID: 23186771……

Programme: …MSc Data Analytics…… **Year:** 2025……

Module: …MSc Research Project……

Supervisor: Jorge Basilio……

Submission Due Date: 11/08/2025……

Project Title: Spotting the Synthetic: Vision Transformers vs Diffusion-Based Image Generation……

Word Count: 6273…… **Page Count ..19..** (excluding AI Form)….

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Brian Riera……

Date: …09/08/2025……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

Your Name/Student Number	Course	Date
Brian Riera	MSc Research Project	08/08/2025

AI Acknowledgment

Tool Name	Brief Description	Link to tool
ChatGPT	1. ChatGPT helped restructure existing notebook-based training code into Python script, adding command-line argument parsing, automatic mixed precision integration, and clearer training loop organisation as well as handling debugging with back and forth messages. 2. After researching Grad-CAM and Vision Transformer attention methods, I used ChatGPT to generate the working code. It generated the first versions of the Grad-CAM functions for my ResNet model and the attention rollout for the ViT, and I then worked with it to fix errors and adjust the code so everything integrated properly. This involved sharing error messages and test results until both visualisations worked as intended. 3. Used to assist in rewording and rephrasing certain sections of the Introduction and Literature Review. I was instructed to adapt material I had previously written and submitted for the CA2 in the Research in Computing module to avoid duplication in Turnitin. I chose to use ChatGPT to help vary wording from the original, with all ideas, references, analyses, and conclusions remaining entirely my own. In several cases, the reworded text from ChatGPT was further edited and refined by myself before inclusion in the thesis.	

Description of AI Usage

ChatGPT	
For the training scripts - helped turn notebook into scripts, and then add specific parts to training script as well as error debugging	

Can you edit this script to make it print something every so often so I can see if it's working?	[Training script +
	for batch_idx, (images, labels) in enumerate(train_loader):
	if batch_idx % 10 == 0:
	print(f'Processing batch {batch_idx}')

ChatGPT	
For GradCam + ViT attention rollout	
I want to implement Grad-CAM to see which parts of the image my trained ResNet-50 is focusing on. Can you give me a PyTorch function that takes an image tensor and returns a heatmap overlay, working for both my real and synthetic classes?"	[Code chunk given, then repeat back and forth with error message/outputs

ChatGPT	
For rewording example of original text vs new worded	
The rise of generative artificial intelligence (AI) has led to an unprecedented surge in synthetic image creation, transforming industries such as digital art, entertainment, and media forensics. Although innovative, this progress has brought with it serious implications, including ethical concerns, particularly regarding misinformation, privacy violations, and deepfake technology (Balasubramaniam *et al.* 2024)	The emergence of generative artificial intelligence (AI) has led to a dramatic increase in the production of synthetic images, significantly impacting sectors like digital art, entertainment, and media forensics. While this advancement has been ground-breaking, it has also led to serious challenges especially in the realm of ethics, with regards to the spread of misinformation, breaches of privacy, and the proliferation of deepfake content (Balasubramaniam *et al.*, 2024).

Spotting the Synthetic: Vision Transformers vs Diffusion-Based Image Generation

Brian Riera
23186771

Abstract

As AI-generated images become more realistic and widespread particularly those produced by diffusion models like DALL·E 3 and MidJourney, there is an increasing need for effective detection methods. This study investigates the potential of Vision Transformers (ViTs) to distinguish between real and diffusion-generated images, addressing a gap in current research that has largely focused on GAN-based detection. Using the SuSy dataset, which contains 25,561 images, a ViT model was fine-tuned for binary classification. A ResNet-50 model was also trained as a CNN baseline for comparison. The ViT achieved strong performance with an accuracy of 96.2%, F1-score of 91.6%, and an AUC of 0.9898, demonstrating that transformer-based models are competitive even in limited data scenarios. While ResNet-50 slightly outperformed the ViT across most metrics, the results support the viability of ViTs in synthetic image detection and highlight their potential for future applications as generative models continue to evolve.

1 Introduction

The emergence of generative artificial intelligence (AI) has led to a dramatic increase in the production of synthetic images, significantly impacting sectors like digital art, entertainment, and media forensics. While this advancement has been ground-breaking, it has also led to serious challenges especially in the realm of ethics, with regards to the spread of misinformation, breaches of privacy, and the spread of deepfake content (Balasubramaniam *et al.*, 2024). With the growing presence of AI-generated content, the capability to reliably identify synthetic imagery has become crucial in tackling associated ethical issues.

However, with the rapid progress in generative technologies ranging from the initial introduction of Generative Adversarial Networks (GANs) by Goodfellow *et al.* (2014) to the more recent rise of diffusion models by Ho, Jain, and Abbeel (2020) it has become increasingly challenging to tell apart artificial images from those created by humans. This was underscored by findings from Lu *et al.* (2023), which reported that people incorrectly identified AI-generated visuals 39% of the time.

Since the introduction of the first GAN over ten years ago, a lot of previous research, such as the work by Baraheem and Nguyen (2023) has focused on identifying images created using GANs. However, with diffusion models emerging as a more recent and increasingly dominant force in generative AI, particularly through applications like DALL·E and Stable Diffusion, there is a shortfall in current research, underscored by Guarnera, Giudice, and Battiato (2024), who advocate for improved techniques to identify diffusion-generated media.

Due to their strong performance in pattern recognition and feature extraction, Convolutional Neural Networks (CNNs) have long been the traditional go-to approach for detecting AI-generated content. Yet, even with their effectiveness in identifying obvious inconsistencies, CNNs face limitations when dealing with the complex and diverse artefacts created by newer generative models (Essa, 2024).

In contrast to CNNs, which analyse images by focusing on localised regions, Vision Transformers (ViTs) utilise self-attention mechanisms that allow them to capture broader contextual relationships within an image (Dosovitskiy *et al.*, 2020), and have shown promising results in detecting deepfakes, particularly those generated by GANs.

Essa (2024) demonstrated that ViTs and MLP-Mixers achieved high accuracy when applied to GAN-generated images, it is reasonable to hypothesise that Vision Transformers may also be effective in identifying diffusion-generated content. Their ability to model long-range dependencies across image regions could offer an advantage in detecting the more subtle and complex artefacts associated with diffusion models. However, their use in detecting images created by diffusion models remains limited, indicating an area for further research.

This study addresses that gap by assessing how effectively Vision Transformers can detect diffusion-generated images with their performance compared to a baseline model based on Convolutional Neural Networks. To carry this out, this study will use a publicly available dataset consisting of 25,561 images from Bernabeu-Pérez *et al.* (2024) and models will be assessed using various performance metrics such as Accuracy, Recall, Precision, F1-score, and Area Under the Curve (AUC).

This research seeks to answer the following question: "How effectively can Vision Transformers classify real images from diffusion-generated images?"

By answering this question, this study makes the following contributions: It evaluates the effectiveness of Vision Transformers in classifying diffusion-generated images, addressing a relatively underexplored area within synthetic image detection. It incorporates a benchmark Convolutional Neural Network (CNN) model, specifically ResNet-50, to provide a comparative baseline for assessing the performance of Vision Transformers. Finally, the study contributes to the broader domain of AI-generated content detection by concentrating on diffusion-based generative models, which are becoming increasingly prominent in contemporary media.

The remainder of this report is structured as follows: Section 2 reviews relevant literature, beginning with the evolution of generative models from GANs to diffusion-based approaches, as well as looking at corresponding detection strategies. Section 3 outlines the research methodology, including dataset preparation, pre-processing, training strategy, and considerations for explainability. Section 4 details the design specifications of the selected models. Section 5 presents the implementation process, including software setup, training pipeline, and performance optimisations. Section 6 reports the experimental results and discusses the findings in the context of existing research, highlighting model performance, strengths, and limitations. Finally, Section 7 concludes the report by summarising the main contributions and suggesting directions for future research.

2 Related Work

Generative Adversarial Networks (GANs) have played a leading role in the field of image generation since their introduction by Goodfellow *et al.* (2014). These models improve image quality through adversarial training, involving two competing neural networks: a generator that produces synthetic images and a discriminator that evaluates their authenticity. However, as attempts to generate high-resolution outputs often led to issues such as mode collapse, early generative models were limited to producing low-resolution images, to overcome this, several advancements were introduced over the following years.

Progressive Growing of GANs (ProGAN) by Karras *et al.* (2018), which involved gradually increasing model complexity during training enabled the generation of higher-resolution images, laying the groundwork for StyleGAN (Karras *et al.*, 2019), which offered even greater control over image synthesis. Another important contribution came from Zhu *et al.* (2020) with the introduction of CycleGAN, enabling image-to-image translation across different domains without the need for paired training data by utilising unpaired image translation techniques.

While these generative methods made significant strides in image generation, their early limitations in controlling image content and maintaining diversity eventually led to the emergence of a new class of generative models known as diffusion models. A foundational example is the Denoising Diffusion Probabilistic Model (DDPM), introduced by Ho *et al.* (2020), which forms the basis for many modern generative systems, including Stable Diffusion and the diffusion-based image decoder used in DALL·E 2.

Unlike the adversarial approach used in GANs, diffusion models generate images through an iterative denoising process that progressively refines random noise into high-quality visuals. Several recent developments have extended the foundational diffusion framework. Notably, Rombach *et al.* (2022) introduced Latent Diffusion Models (LDMs), which underpin the widely adopted Stable Diffusion technology, utilising cross-attention mechanisms to guide image generation based on specific inputs or prompts. Instead of operating directly with pixel level data like conventional diffusion models, LDMs perform the denoising process within a learned latent space, allowing for more efficient training by reducing computational costs without compromising image quality. Similarly, DALL·E 2, developed by Ramesh *et al.* (2022), used the Contrastive Language–Image Pre-Training (CLIP) framework to enable text-conditional image generation by encoding textual prompts into the synthesis process, achieving strong performance across a range of image generation benchmarks as noted by Guarnera, Giudice, and Battiato (2023).

With the advancements in GANs and diffusion models over recent years the realism of synthetic image generation has advanced significantly, highlighting a need for improved detection techniques. The following section will review various detection approaches, starting with those designed for GAN-generated images and moving on to methods tailored to diffusion-based outputs, with a focus on comparing their effectiveness in identifying synthetic content.

2.1 Overview of Deepfake Detection Approaches

In computer vision, the term artifacts refer to visual irregularities or distortions that arise due to limitations in how generative models process images. As diffusion models have only recently emerged as a dominant approach in image synthesis, a lot of previous work in detecting AI-generated imagery has focused on outputs from GANs. Initial detection efforts commonly relied on manually engineered features, drawing on visual cues such as colour distribution, saturation, blending inconsistencies, and co-occurrence patterns to identify synthetic content (Wang *et al.*, 2023).

In addition to spatial artifacts, researchers have investigated a range of alternative detection methods. Frank *et al.* (2020) explored frequency-based approaches, noting that the up-sampling steps commonly used in GAN architectures often introduce distinctive patterns in the frequency domain, where image data is represented in terms of signal frequencies rather than pixel intensities. Building on this, Wang *et al.* (2020) demonstrated that convolutional neural networks trained on JPEG-compressed and blurred ProGAN images could generalise to other GAN-generated content. However, more recent studies by Corvi *et al.* (2023) and Ricker *et al.* (2024) identified notable differences between the frequency characteristics of images produced by GANs and those generated by diffusion models. Corvi *et al.* (2023) reported that detectors trained solely on GAN outputs struggled to perform well when applied to diffusion-generated images. In cross-architecture evaluations, they observed that even when detectors achieved high average precision, this was often accompanied by low accuracy, an inconsistency also highlighted by Zheng *et al.* (2024).

With diffusion models only appearing in recent years, research into this area is still in its early stages. Corvi *et al.* (2023) explored whether diffusion-generated images could be distinguished by distinct forensic fingerprints, similar to GAN-generated content, but found only partial evidence to support this. Their study evaluated the performance of well-known detectors across various conditions, revealing significant differences between models due to the unique forensic characteristics associated with each. Consistent with earlier findings, the study underscored the ongoing challenge of generalisation, noting that detectors trained exclusively on GAN outputs performed poorly when applied to diffusion-generated images. They found that including diffusion examples during training improved detection accuracy for some models, yet noted suboptimal results for others, reinforcing the need for continued research in this area. While there has been an abundance of research into detectors for GAN-generated images, the detection of diffusion-generated content is still a developing area. Studies such as that of Corvi *et al.* (2023) indicate that the architectural differences between GANs and diffusion models contribute to the poor performance of existing detectors when applied to diffusion-generated images. The following section examines newly developed detection methods designed to address the unique characteristics of diffusion-based image synthesis.

One of the early promising methods for detecting diffusion-generated images is DE-Fake, introduced by Sha *et al.* (2023), using multimodal fusion with CLIP as the backbone network. While image-only classifiers showed strong performance in certain scenarios, the authors echoed findings from previous studies noting that these models struggled to generalise across different generative architectures. This was addressed by introducing a second detection

strategy, combining image data with corresponding textual prompts using BLIP, an image captioning model employed when natural prompts were unavailable. This hybrid approach outperformed the image-only model in terms of accuracy. However, a limitation of the study mentioned by Guarnera, Giudice, and Battiato (2024) is that evaluations were conducted under idealised conditions, raising concerns about real-world applicability.

Building on the findings of Corvi *et al.* (2023), Wang *et al.* (2023) also observed that existing GAN detectors experienced substantial drops in performance when applied to diffusion-generated images, with accuracy falling below 60%, also finding similar to Corvi *et al.* (2023) that while there was some improvement when including diffusion generated images in the training process, detectors continued to perform poorly when tested on unseen diffusion architectures. To address this generalisation limitation, Wang *et al.* proposed the Diffusion Reconstruction Error (DIRE) method, which utilises a pre-trained diffusion model to evaluate reconstruction discrepancies. Their results showed that real images produced larger reconstruction errors compared to diffusion-generated ones, which could be more accurately reconstructed by the model. Using their newly introduced DiffusionForensics dataset, they demonstrated that DIRE outperformed earlier detection methods and achieved strong generalisation across a variety of diffusion models. However, this improved performance comes at a cost, DIRE is computationally intensive compared to lightweight classifiers. Moreover, because the method relies on a specific pre-trained diffusion model, it may still face challenges when applied to entirely novel or significantly different architectures.

While Wang *et al.* addressed generalisation through reconstruction-based techniques, other researchers have explored adaptive learning strategies. Epstein *et al.* (2023) proposed an online learning framework that continuously updates detection models with samples from newly released generators, reflecting real-world conditions where generative models evolve rapidly. Their results indicated that classifiers trained with this incremental method showed improved generalisation to unseen models. However, they also observed that significant architectural changes between models could lead to considerable drops in detection accuracy.

Bernabeu-Pérez *et al.* (2024) introduced SuSy, a patch-based architecture built on ResNet, which achieved strong performance when detecting diffusion-generated images from known generators. Nonetheless, similar to earlier studies, SuSy experienced a marked decline in accuracy when applied to previously unseen diffusion models, reinforcing the broader issue of generalisation reported by others in the field such as Corvi *et al.* (2023), Wang *et al.* (2023) Epstein *et al.* (2023), and Sha *et al.* (2023).

Since the introduction of Vision Transformers by Dosovitskiy *et al.* (2020), there has been growing interest in their application to image analysis tasks, including the detection of AI-generated content. Traditionally, Convolutional Neural Networks have dominated this field due to their efficiency and interpretability. This is exemplified by the use of CNN architectures in studies such as Epstein *et al.* (2023) who employed ResNet-50, Bernabeu-Pérez *et al.* (2024), and Corvi *et al.* (2023), all of which employed CNN-based detectors to analyse diffusion-generated content. ViTs, which rely on attention mechanisms to capture broader patterns and relationship across the entire image, have been proposed as a promising alternative. Baraheem and Nguyen (2023) evaluated a ViT-based model (MaxViT) alongside traditional CNNs such as ResNet-50 and EfficientNet on GAN-generated images. While the ViT model demonstrated competitive performance, it did not outperform the top-performing CNNs in their experiments.

However, their study was limited to GAN-generated content and examined only a single ViT variant.

To date, there has been a limited amount of research evaluating the performance of Vision Transformers on diffusion-generated images, revealing a significant gap in the literature. While generalisation challenges remain a common theme across existing studies, the primary aim of the current study is not to address this issue directly. Instead, this study focuses on assessing the effectiveness of a Vision Transformer architecture in detecting diffusion-generated content, using a benchmark dataset curated by Bernabeu-Pérez *et al.* (2024). By doing so, it aims to contribute to the growing body of work exploring ViTs in synthetic image detection, specifically within the context of diffusion-based models.

3 Research Methodology

This section covers the methodological framework including dataset handling, pre-processing, model selection, training procedures, evaluation strategies, and explainability tools used in this study.

3.1 Dataset and Pre-Processing

The dataset selected for this study is the SuSy dataset (Bernabeu-Pérez *et al.*, 2024), comprising of a mixture between real and synthetic images generated by various diffusion-based models, including DALL·E 3, MidJourney, and Stable Diffusion. The dataset was obtained through the Hugging Face datasets library[1] and is publicly available. No personally identifiable information (PII) was accessed or processed. Although some images depict human figures, all data is used in accordance with the terms and ethical guidelines provided by the dataset's creators.

For the purpose of reproducibility, the official dataset splits were used without modification. The dataset contains a total of 25,561 images, 5,435 authentic, and 20,126 synthetic images split into 14,451 for training, 5,555 for validation, and 5,555 for testing with a noticeable class imbalance in all data splits highlighted in Table 1. From Table 1, it can be seen that there is a mix of JPEG and PNG images, with all authentic (real) images in JPEG format, while the majority of synthetic images were stored as PNGs.

In the original dataset, each image source was assigned its own class, totalling 6 class labels, 5 synthetic and 1 authentic. However, as the current study is a binary classification task, the labels have been consolidated and relabelled as follows, synthetic images as class 0 and authentic images as class 1, allowing for a direct investigation into the model's ability to distinguish between real and AI-generated content. A breakdown of the original and relabelled distribution is shown in Table 1.

[1] Available at: https://huggingface.co/datasets/HPAI-BSC/SuSy-Dataset

Table 1: Original class distribution and new binary mapping for the SuSy dataset. Adapted from Bernabeu-Pérez *et al.* (2024). Note: Type R = Real, S = Synthetic

Source Dataset	Generator	Year	Image Format	Original Label	Type	New Label	Train	Validation	Test
COCO	—	2017	JPG	0	R	1	2,967	1,234	1,234
dalle3-images	DALL·E 3	2023	JPG	1	S	0	987	330	330
diffusiondb	Stable Diffusion 1.X	2022	PNG	2	S	0	2,967	1,234	1,234
SDXL realisticSDXL	SDXL	2023	PNG	3	S	0	2,967	1,234	1,234
mj-tti	MidJourney v1/2	2022	PNG	4	S	0	2,718	906	906
mj-images	MidJourney v5/6	2023	JPG	5	S	0	1,845	617	617

As both the ResNet and ViT models were pretrained on ImageNet (Deng *et al.*, 2009)[2], all images were resized to 224x224 pixels to meet the input dimensions expected by the models. To improve stability during fine-tuning and to ensure compatibility with the distribution learned during pretraining, normalisation was performed using the standard ImageNet statistics: mean = [0.485, 0.456, 0.406], and standard deviation = [0.229, 0.224, 0.225].

As external patch extraction is not required for ViTs to capture local or global context. This study adopts a strategy of full-image resizing, differing from the original SuSy authors who extracted multiple high-contrast patches from each image to increase training diversity for their CNN model. This strategy was chosen to simplify the training pipeline while also ensuring the images were compatible with the pretrained ViTs, which internally divide input images into fixed-size non-overlapping patches (e.g., 16x16) as part of their embedding process. This use of full images also allows for a direct comparison between the selected ResNet and ViT models under similar conditions.

As previous work such as that of Shorten & Khoshgoftaar (2019) has shown that common data augmentation strategies improve CNN and transformer generalisation in image classification tasks. Data augmentation techniques including random resized cropping, horizontal flipping, brightness/contrast adjustments, gamma correction, gaussian blur, and JPEG compression were applied to simulate real-world conditions with augmentations being applied probabilistically to introduce variability without distorting semantic content, partially matching the SuSy authors setup.

It was noted by Bernabeu-Pérez *et al.* (2024) that shortcut learning can occur when image compression artefacts inadvertently correlate with labels. For this reason, in addition to the JPEG compression performed during data augmentation, JPEG format normalisation was applied by creating an additional version of the dataset in which all images were converted to a JPEG format to ensure that both the authentic and synthetic images had similar compression characteristics, described in section 7.

[2] ImageNet dataset available at: https://www.image-net.org

3.2 Training Strategy

Although initial training was conducted on the original mixed format dataset, the JPEG normalised dataset was used in the subsequent and final training runs for both ViT and ResNet. Both Vit and ResNet were trained for a maximum of 20 epochs, mirroring the configuration set by the SuSy authors and with considerations for training time and computational resources. To prevent overfitting, early stopping was set up with a patience of five epochs, with the best performing model retained for final evaluation based on validation loss.

While the original SuSy implementation used the Adam optimiser, this study implements AdamW, a decoupled weight decay variant of Adam based on its improved regularisation and training stability in transformer-based architectures (Loshchilov & Hutter, 2019). The selected learning rate of 1e-4 and weight decay of 0.01 are retained from the SuSy setup and align with the empirically supported range for fine-tuning pretrained ViT models, as reported in Dosovitskiy *et al.* (2021) with lower learning rates typically being used for fine-tuning to avoid large updates that might overwrite pretrained weights, allowing the model to gradually adapt to the task.

In order to reduce the risk of model bias toward the majority class (synthetic), inverse class frequency weights were applied to the cross-entropy loss function. This was done to encourage the model to learn a more balanced decision boundary by increasing the penalty for misclassifying authentic images. This method was selected over other techniques such as oversampling or synthetic augmentation to preserve the natural image distributions and avoid potential overfitting.

Although binary cross-entropy loss is commonly used for binary classification tasks, it requires sigmoid activation and works with floating-point labels, which are less compatible with the transformer model's raw output format. For this reason, a logit-based cross-entropy loss was selected to work alongside the class weights to better handle class imbalance and maintain compatibility with the model's output format. As logit-based cross-entropy loss expects raw logits and integer class labels, it avoids the need to manually apply activation functions or convert labels to one-hot encoding.

3.3 Evaluation Metrics

Standard metrics for binary classification: accuracy, precision, recall, F1-score, and AUC were used to assess model performance, with evaluation performed solely on the provided test split. The metrics are defined as follows, where TP is true positives, TN true negatives, FP false positives, and FN false negatives.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \qquad Precision = \frac{TP}{TP + FP}$$

$$Specificity = \frac{TN}{TN + FP} \qquad (Recall)\ Sensitivity = \frac{TP}{TP + FN}$$

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

AUC is calculated as the area under the plot of the true positive rate (TPR) against the false positive rate (FPR) across various classification thresholds.

3.4 Model Explainability

Explainability techniques such as Grad-CAM for ResNet (Selvaraju *et al.*, 2017), and attention rollout for ViT (Abnar and Zuidema, 2020) were applied to a sample of test images to better understand model behaviour, providing a visual understanding of which features or regions each model prioritised during inference. Grad-CAM visualises the spatial regions that are most influential to the model's classification decisions. While attention rollout highlights the image patches that received the highest cumulative attention by aggregating weights across all layers.

4 Design Specification

The primary model evaluated in this study is ViT-B/16, a Vision Transformer architecture introduced by Dosovitskiy *et al.* (2021). ViT-B/16 divides each input image into non-overlapping 16×16 patches, which are then flattened and linearly projected into fixed-length embeddings. A learnable class token is added to the beginning of this sequence, and positional embeddings are included to retain spatial information. The resulting sequence is passed through a transformer encoder comprising 12 layers, each with 768 hidden units which are neurons in the feedforward part of the transformer, and 12 self-attention heads. These attention heads allow the model to focus on different parts of the input image simultaneously and capture complex patterns across the entire image. This design enables ViT to model global context through multi-head self-attention, making it well suited for detecting subtle or spatially distributed artefacts in synthetic images (Wang *et al.*, 2024). The ViT-B/16 architecture, including the patch embedding process and transformer encoder are displayed in Figure 1.

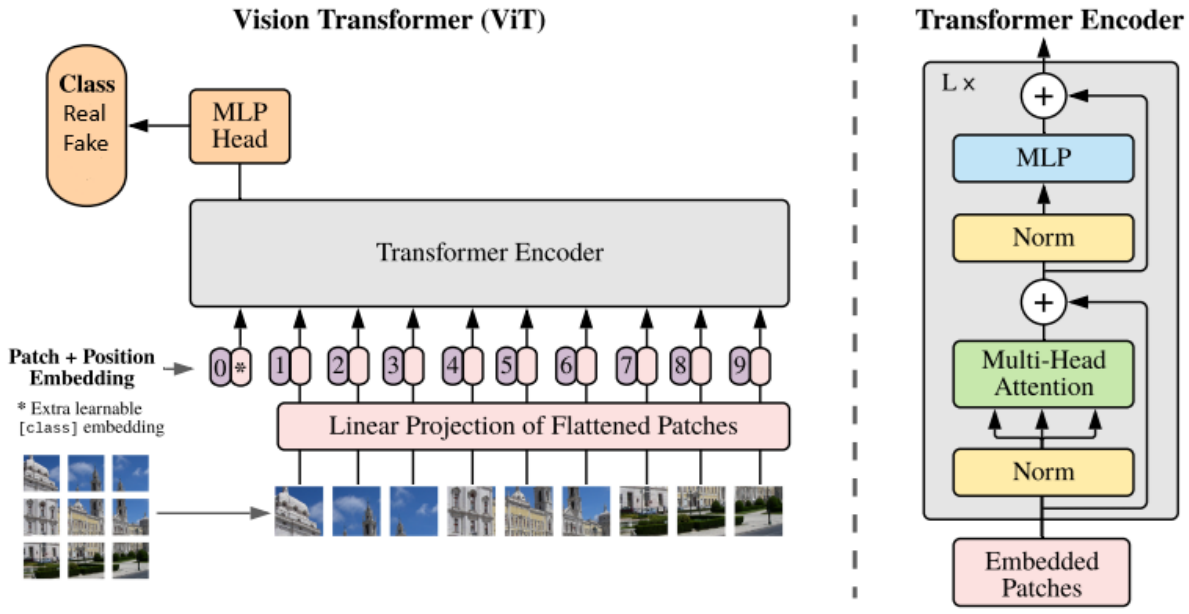

Figure 1: Simplified illustration of the Vision Transformer (ViT) architecture, showing patch embedding and Transformer encoding. Adapted from Dosovitskiy *et al.* (2020).

This study employs ResNet-50, a widely adopted CNN, as a baseline model to provide a point of comparison. Proposed by He *et al.* (2016), ResNet-50 consists of 50 layers organised into residual blocks that help mitigate the vanishing gradient problem, allowing for deeper networks to be trained effectively. It is known for its residual learning framework and effectiveness in hierarchical feature extraction and has demonstrated strong performance in image classification and deepfake detection tasks (Epstein *et al.*, 2023; Corvi *et al.*, 2023). Figure 2 illustrates the ResNet-50 architecture, highlighting its residual block structure and hierarchical feature extraction pathway.

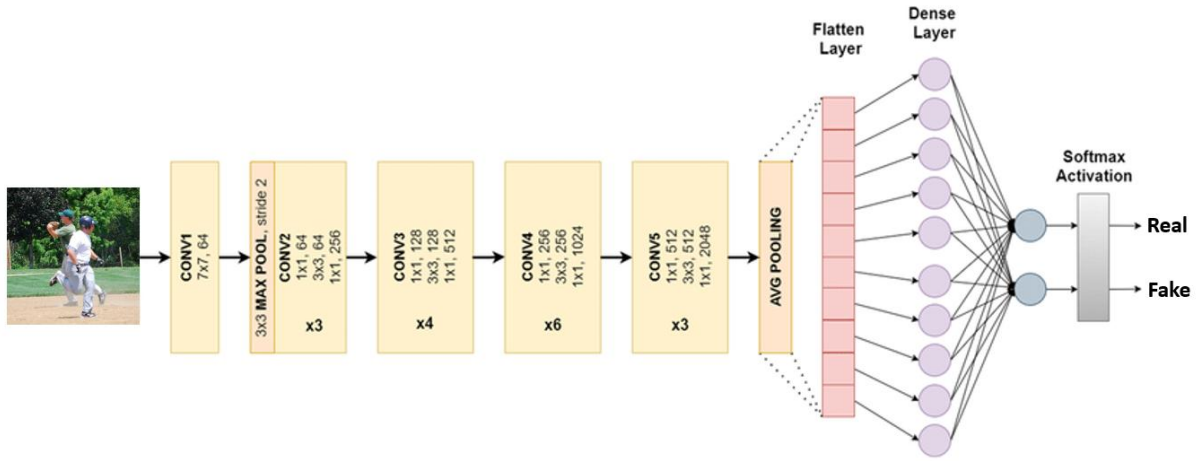

Figure 2: ResNet-50 architecture adapted for binary classification. Adapted from Anis *et al.* (2021).

ViT-B/16 was selected as the primary model due to its strong performance on ImageNet when pretrained and fine-tuned, as demonstrated by Dosovitskiy *et al.* (2021). It has also shown

promising results in recent studies on synthetic image detection (Wang *et al.*, 2024). While ResNet-50, as a strong CNN baseline, allows meaningful comparison between convolutional and transformer-based approaches.

5 Implementation

A Jupyter notebook environment was setup for initial model development, allowing prototyping, debugging and inspection of intermediate outputs. Following this, the workflow was migrated to modular Python scripts once the pipeline was validated and training procedures stabilised, to support more structured experimentation and allow for faster training runs. This section outlines the key implementation components, the system environment, and the rationale behind selected design decisions.

All training and evaluation were done using the PyTorch deep learning framework. The Vision Transformer architecture was accessed through the timm library, while ResNet-50 was sourced from torchvision.models. Data augmentation was carried out using the Albumentations library to ensure consistent pre-processing across training and evaluation stages.

The following software versions and hardware configurations were used throughout the experiments: Python 3.12.7, PyTorch 2.7.1 with CUDA 11.8, Torchvision 0.22.1, and timm version 1.0.15. The experiments were conducted on a machine equipped with an NVIDIA GeForce RTX 4070 GPU and an AMD Ryzen 5 7600X processor. Automatic Mixed Precision (AMP) training was enabled via torch.cuda.amp, reducing memory consumption and enabling larger batch sizes without degrading performance.

To allow the use of multi-threaded data loading and improve maintainability, the project was modularised into a series of Python scripts. Retrieval of the SuSy dataset from Hugging Face is handled by the module data_loader.py. Pre-processing and format conversion from PNG to JPEG are managed by convert_dataset.py, which uses the Python Pillow library to convert PNGs to RGB and save them as JPEG at 95% quality, applied in parallel across each dataset split. A custom PyTorch-compatible dataset class is defined in dataset_utils.py, supporting Albumentations3 based transformations specified in transforms_utils.py.

Model training is handled by train_format_test.py, which implements fine-tuning procedures for ViT and ResNet-50 and includes optional JPEG conversion for format analysis. Evaluation between the models is conducted using evaluate_models.py, which loads saved model checkpoints, computes classification metrics, and stores predictions for further analysis.

Due to multiprocessing limitations in the Jupyter notebook environment, initial training was conducted using a batch size of 32 and zero data loader worker threads. Training was moved to standalone Python scripts after pipeline validation, enabling the use of four data loading worker threads for multi-threaded data loading leading to improved training speed. To reduce memory usage and support larger batch sizes AMP (Automatic Mixed Precision) was implemented via torch.cuda.amp allowing the batch size to be increased to 64. The best-performing model checkpoint based on validation loss was saved for final evaluation. Separate checkpoints were maintained for each architecture (ViT and ResNet-50) to support

[3] Albumentations is an image augmentation library for computer vision tasks.

comparative analysis. This final configuration: batch size of 64, four worker threads, learning rate of 1e-4, weight decay of 0.01, AMP enabled, and early stopping was applied to both models.

6 Evaluation

The results of the ViT model alongside the baseline ResNet model are presented in this section, performance metrics, confusion matrices, and visualisation tools are used to assess and interpret each model.

The classification performance of each model on the test set is shown in Table 2 below. Metrics include accuracy, precision, recall, F1-score, and area under the curve (AUC). ViT mixed dataset refers to the initial ViT model trained on the original mixed-format dataset containing both JPEG and PNG images, before JPEG format normalisation and is included here for reference.

Table 2: Test performance of all evaluated models.

Model	Accuracy	Precision	Recall	F1-Score	AUC
ViT mixed dataset	0.9579	0.9019	0.9092	0.9056	0.9890
ViT JPEG only	0.9620	0.8974	0.9360	0.9163	0.9898
ResNet JPEG only	0.9694	0.9229	0.9408	0.9318	0.9930

As can be seen in Table 2 the ViT model performed well in all metrics tested, demonstrating its effectiveness in detecting diffusion generated images. Its high recall indicates a strong ability to identify real content, while the F1-score and AUC reflect a balanced and reliable overall performance. Although the ResNet model exhibited a higher performance across all metrics, the differences were modest. Both models showed high classification capability, with AUC values close to 1, indicating strong discrimination between real and synthetic classes. To assess convergence and training behaviour, training and validation loss values were logged during training. Both models displayed a similar smooth decline in training loss and stable training dynamics with early stopping triggered appropriately.

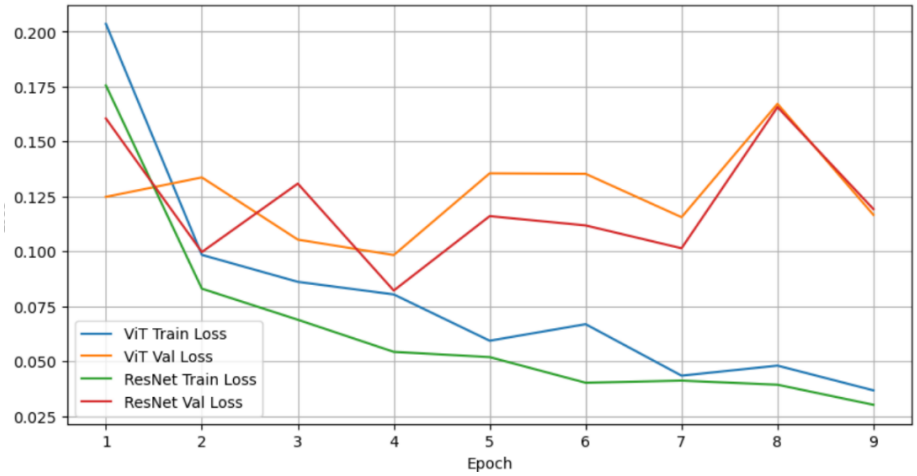

Figure 3: Training and Validation Loss Curves for ViT and ResNet models.

The confusion matrices in Figure 4 provide a breakdown of true positives, true negatives, false positives, and false negatives for each model.

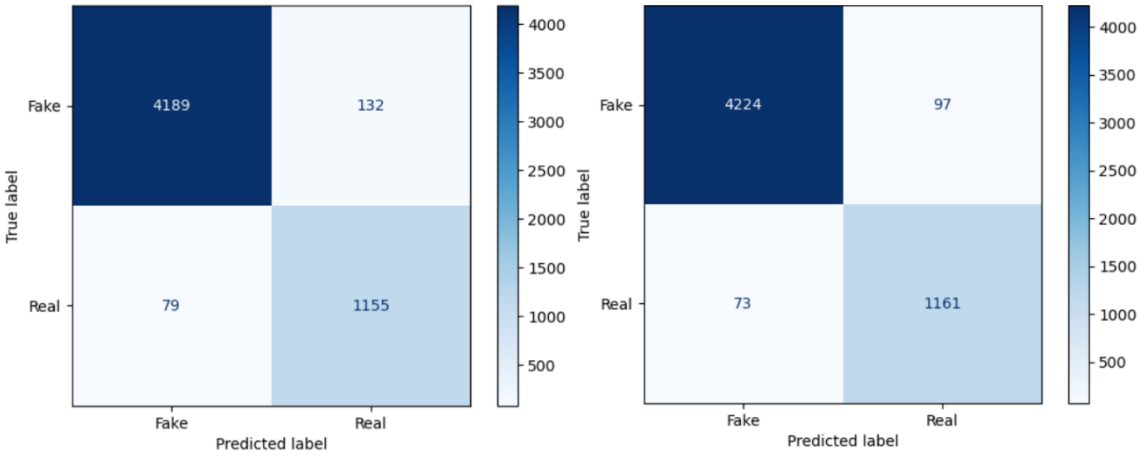

Figure 4: Confusion matrices for ViT (left) and ResNet (right).

The ViT model correctly identified a high number of both real and synthetic images, reflecting strong classification capability. While ResNet-50 achieved slightly fewer misclassifications overall, the difference was relatively small and results indicate that both models had high sensitivity and specificity, with the ResNet model edging the ViT model in performance across both classes, aligning with its marginally higher precision and recall scores shown in Table 2.

6.1 Discussion

This study investigated the effectiveness of Vision Transformers in distinguishing real from diffusion-generated images, using ResNet-50 as a benchmark CNN for comparison. Both models were fine-tuned on a JPEG format-normalised version of the SuSy dataset and evaluated on a held-out test set.

It was initially hypothesized that ViT might perform competitively with ResNet. While CNNs typically benefit from stronger inductive biases in smaller dataset settings, it was thought that the ImageNet pretraining may help to narrow the performance gap for the ViT, allowing it to remain competitive despite its usual reliance on large scale data. Examining the results from both models it can be seen that there was high classification performance from both the ViT and ResNet model. While ViT performed competitively the ResNet model slightly outperformed ViT across all metrics.

These findings align with previous findings of Baraheem and Nguyen (2023) who tested a ViT based model alongside traditional CNNs including ResNet-50 on synthetic image detection on GAN-generated images albeit this study used diffusion-generated images. The performance of the CNN aligns with prior literature, where ResNet variants have demonstrated strong performance in synthetic image detections tasks (Bernabeu-Pérez *et al.,* 2024; Epstein *et al.,* 2023; Corvi *et al.,* 2023) and its performance may stem from its strong inductive biases and capacity to generalise effectively, even on this smaller sized dataset.

Previous work from Touvren *et al.* (2021) has shown that ViTs can perform well in low-data settings when trained with strategies such as knowledge distillation and heavy augmentation, while this study did not employ these techniques, future work could potentially improve ViT performance further by incorporating these techniques.

Both models showed no evidence of overfitting, showing stable training behaviour with consistent decreases in training and validation loss. The loss curves shown in Figure 3 suggest clean convergence and quick adaptation to the task, likely in part to their pretraining. This combination of pretraining and task simplicity may help explain the high AUC and F1 scores observed even after limited epochs.

Explainability visualsations were utilised to assess how each model arrived at its predictions, with an example from each model shown in Figure 5. The ResNet-50 Grad-CAM concentrated heavily on the leftmost dog and nearby table area, indicating a focus on specific, content-rich regions. In contrast, the ViT attention rollout displayed a more even spread of focus across the scene, with a noticeable hotspot on the central dog's head. This broader, more distributed attention reflects the ViT's ability to capture global context and long-range relationships within the image, providing insights into the different decision-making processes of the two architectures.

Figure 5: Grad-CAM (left) and ViT Attention Rollout (right) for synthetic image examples.

ResNet-50, with its convolutional filters and strong local inductive biases, excels at detecting fine-grained textures or compression artefacts that may indicate synthetic content, while the Vision Transformer's global self-attention mechanism allows it to take in the whole scene at once, capturing long-range dependencies and subtle global inconsistencies. These capabilities make ViT a compelling alternative for diffusion-generated image detection, where artefacts can emerge at both local and global levels.

However, when interpreting the findings of this study, it is important to consider several experimental limitations that may have affected the robustness and generalisability of the results. The study was conducted using a single training run per model configuration. While this provides a clear snapshot of the model behaviour under fixed conditions, it does not account for variability caused by randomness in how training begins or progresses. To assess

consistency and reduce variance in performance estimates, results would be averaged over multiple seeds, however, due to the time required to train each model, repeated runs were not feasible within the scope of the project. As such, although this study was not designed as a formal comparison, no conclusions about statistical reliability can be drawn when examining the performance of the ViT to the baseline CNN model and insights into their comparative performance should be interpreted with caution.

Although the purpose of the study was to evaluate the performance of ViT on synthetic image detection, the granularity of the analysis was limited due to the grouping of multiple generative sources into a single 'synthetic' class. Synthetic images were drawn from various diffusion models (e.g., DALLE-3, SD1.X, SDXL, and Midjourney) each of these potentially exhibiting distinct artefacts. However, because of the small size of each individual subset, they were aggregated into one class for training stability. As a result, it does not reveal whether the ViT generalises better to specific generation methods. A more detailed study with sufficient data per model could evaluate model performance across individual generative techniques. This follows the approach of Bernabeu-Pérez *et al.* (2024) which tested their model across multiple datasets and generators to assess generalisability. Such an investigation was outside the scope of this study but could be valuable for future research.

As part of the pre-processing pipeline, all images were converted to JPEG format, this was carried out to decrease training time, standardise model input, and mitigate potential format induced shortcut learning. While this was successful in that regard, it may have introduced its own distribution shift. Grommelt *et al.* (2024) demonstrated that biases can be introduced into detection tasks through differences in JPEG compression and image resolution between real and synthetic images. In particular, detectors may learn to distinguish between compression artefacts rather than genuine generative features. Although format conversion was applied uniformly, the synthetic images may have responded differently to compression artefacts than the original COCO JPEGs, potentially influencing classifier behaviour. However, with the similar performance between the initial ViT trained on the mixed dataset and the final ViT trained on JPEG-only, compression bias does not seem to be a major concern for this study.

While this study uses the same training dataset as Bernabeu-Pérez *et al.* (2024) for real and synthetic image classification, its scope and methodology differ significantly. Bernabeu-Pérez *et al.* (2024) evaluated a wide range of CNN-based models (e.g., EfficientNet, ResNeXt) across multiple datasets and generative models in a patch level detection setting. In contrast, this study focuses on a binary classification task using only the single dataset, not testing model generalisability. Additionally, this work explores architectures not considered by Bernabeu-Pérez *et al.* (2024), specifically, Vision Transformers, to examine how transformer-based models compare to traditional CNNs in this domain.
While the evaluation setups are not directly comparable, this study provides complementary insights by introducing transformer-based architectures to the task, an area still in the early stages of exploration within the broader research community.

7 Conclusion and Future Work

This study set out to investigate the research question: How effectively can Vision Transformers classify real images from diffusion-generated images?

To explore this, a Vision Transformer (ViT) was trained on a binary classification task involving real images from the COCO dataset and synthetic images generated by various diffusion-based models. To contextualise ViT's performance, a ResNet-50 model was also trained under identical conditions and used as a CNN baseline

The results showed that both models achieved strong classification performance, with the ViT performing competitively with ResNet for synthetic image detection although slightly lower across all metrics. These results suggest that pretrained vision models, including ViTs, can be effectively adapted to synthetic image detection tasks even with limited training data.

In line with these findings, the objectives of the study were successfully met: both models were implemented and evaluated using standard performance metrics, and their internal mechanisms were explored using visualisation tools in the discussion.

In real-world applications such as content moderation, media forensics, or deepfake detection, the ability to fine-tune models like ViT or ResNet could offer a low-cost and scalable approach to identifying diffusion-generated content.

Nonetheless, the study has several limitations. Each model was trained and evaluated using a single run, without multiple seeds or variance estimation. The synthetic class was treated as a single group despite containing images from several different generative models, limiting both the granularity and generalisability of the findings. Additionally, JPEG compression while applied uniformly may have introduced biases, though the consistent performance across both versions of the ViT suggests this was not a major issue.

Future work should address these limitations by training with multiple seeds to assess performance variability, evaluating model performance across individual generator types, and testing generalisability across datasets. There is also clear potential for further exploring transformer-based models in synthetic image detection, particularly under more realistic deployment conditions or adversarial settings.

In summary, this study contributes to the growing field of synthetic image detection by investigating the use of Vision Transformers for detecting diffusion-generated images. While limited in scope, the findings show that ViTs when properly fine-tuned represent a viable and promising architecture for future research and practical applications in synthetic image detection.

References

Abnar, S. and Zuidema, W. (2020) 'Quantifying attention flow in Transformers', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4190–4197. doi: https://doi.org/10.18653/v1/2020.acl-main.385

Anis, S., Xuan, T., Chuah, J.H., Usman, J., Qian, P. and Lai, K.W. (2021) 'A comparative study of multiple neural network for detection of COVID-19 on chest X-ray', *EURASIP Journal on Advances in Signal Processing*, 2021(1). doi: https://doi.org/10.1186/s13634-021-00755-1

Balasubramaniam, S., Chirchi, V., Kadry, S., Agoramoorthy, M., Senthilvel, G.P., Kumar, S.K. and Sivakumar, T.A. (2024) 'The road ahead: Emerging trends, unresolved issues, and concluding remarks in generative AI—A comprehensive review', *International Journal of Intelligent Systems*, 38, Article ID 4013195. doi: https://doi.org/10.1155/2024/4013195

Baraheem, S.S. and Nguyen, T.V. (2023) 'AI vs. AI: Can AI detect AI-generated images?', *J. Imaging*, 9(10), p.199. doi: https://doi.org/10.3390/jimaging9100199

Bernabeu-Pérez, M., Bartrina-Rapesta, M., Bruna, J. and Giro-i-Nieto, X. (2024) 'SuSy: A patch-level classifier for detecting synthetic images from diffusion models', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, pp. 3798–3807. doi: https://doi.org/10.48550/arXiv.2409.14128

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K. and Verdoliva, L. (2023) 'On the detection of synthetic images generated by diffusion models', in *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. doi: https://doi.org/10.1109/ICASSP49357.2023.10095167

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009) 'ImageNet: A large-scale hierarchical image database', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. doi: https://doi.org/10.1109/CVPR.2009.5206848

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2021) 'An image is worth 16×16 words: Transformers for image recognition at scale', in *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*. Available at: https://openreview.net/forum?id=YicbFdNTTy [Accessed 2 February 2025].

Epstein, D.C., Jain, I., Wang, O. and Zhang, R. (2023) 'Online detection of AI-generated images', in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023)*, pp. 382–392. doi: https://doi.org/10.1109/ICCVW60793.2023.00045

Essa, E. (2024) ‘Feature fusion Vision Transformers using MLP-Mixer for enhanced deepfake detection’, *Neurocomputing*, 598, 128128. doi: https://doi.org/10.1016/j.neucom.2024.128128

Frank, J.C., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D. and Holz, T. (2020) ‘Leveraging frequency analysis for deepfake image recognition’, *arXiv preprint* arXiv:2003.08685. doi: https://doi.org/10.48550/arXiv.2003.08685

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) ‘Generative adversarial nets’, in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 27, pp. 2672–2680. doi: https://doi.org/10.48550/arXiv.1406.2661

Grommelt, P., Weiss, L., Pfreundt, F.J. and Keuper, J. (2024) ‘Fake or JPEG? Revealing common biases in generated image detection datasets’, *arXiv preprint* arXiv:2403.17608. doi: https://doi.org/10.48550/arXiv.2403.17608

Guarnera, L., Giudice, O. and Battiato, S. (2024) ‘Mastering deepfake detection: A cutting-edge approach to distinguish GAN and diffusion-model images’, *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(11), Article 343, pp. 1–24. doi: https://doi.org/10.1145/3652027

He, K., Zhang, X., Ren, S. and Sun, J. (2016) ‘Deep residual learning for image recognition’, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. doi: https://doi.org/10.1109/CVPR.2016.90

Ho, J., Jain, A. and Abbeel, P. (2020) ‘Denoising diffusion probabilistic models’, *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 33, pp. 6840–6851. doi: https://doi.org/10.48550/arXiv.2006.11239

Karras, T., Aila, T., Laine, S. and Lehtinen, J. (2018) ‘Progressive growing of GANs for improved quality, stability, and variation’, *arXiv preprint* arXiv:1710.10196. doi: https://doi.org/10.48550/arXiv.1710.10196

Karras, T., Laine, S. and Aila, T. (2019) ‘A style-based generator architecture for generative adversarial networks’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pp. 4401–4410. doi: https://doi.org/10.1109/CVPR.2019.00453

Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X. and Ouyang, W. (2023) ‘Seeing is not always believing: Benchmarking human and model perception of AI-generated images’, *arXiv preprint* arXiv:2304.13023. doi: https://doi.org/10.48550/arXiv.2304.13023 [Accessed 28 January 2025].

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M. (2022) ‘Hierarchical text-conditional image generation with CLIP latents’, *arXiv preprint* arXiv:2204.06125. doi: https://doi.org/10.48550/arXiv.2204.06125

Ricker, J., Damm, S., Holz, T. and Fischer, A. (2022) 'Towards the detection of diffusion model deepfakes', *arXiv preprint* arXiv:2210.14571. doi: https://doi.org/10.48550/arXiv.2210.14571

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022) 'High-resolution image synthesis with latent diffusion models', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695. doi: https://doi.org/10.48550/arXiv.2112.10752

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) 'Grad-CAM: Visual explanations from deep networks via gradient-based localization', in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. doi: https://doi.org/10.1109/ICCV.2017.74

Sha, Z., Li, Z., Yu, N. and Zhang, Y. (2023) 'De-Fake: Detection and attribution of fake images generated by text-to-image generation models', in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS 2023)*, pp. 3418–3432. ACM. doi: https://doi.org/10.1145/3576915.3616588

Shorten, C. and Khoshgoftaar, T.M. (2019) 'A survey on image data augmentation for deep learning', *Journal of Big Data*, 6(1), p.60. doi: https://doi.org/10.1186/s40537-019-0197-0

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jégou, H. (2021) 'Training data-efficient image transformers & distillation through attention', in *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, PMLR 139, pp. 10347–10357. Available at: https://proceedings.mlr.press/v139/touvron21a.html [Accessed 27 July 2025].

Wang, S.Y., Wang, O., Zhang, R., Owens, A. and Efros, A.A. (2020) 'CNN-generated images are surprisingly easy to spot... for now', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, pp. 8692–8701. doi: https://doi.org/10.48550/arXiv.1912.11035

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H. and Li, H. (2023) 'DIRE for diffusion-generated image detection', *arXiv preprint* arXiv:2303.09295. doi: https://doi.org/10.48550/arXiv.2303.09295

Zheng, C., Lin, C., Zhao, Z., Wang, H., Guo, X., Liu, S. and Shen, C. (2024) 'Breaking semantic artifacts for generalized AI-generated image detection', in *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*. Available at: https://proceedings.neurips.cc/paper_files/paper/2024/hash/6dddcff5b115b40c998a08fbd1cea4d7-Abstract-Conference.html [Accessed 2 February 2025].

Zhu, J-Y., Park, T., Isola, P. and Efros, A.A. (2020) 'Unpaired image-to-image translation using cycle-consistent adversarial networks', in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251. doi: https://doi.org/10.48550/arXiv.1703.10593