



Università degli Studi di Napoli Federico II
Ph.D. Program in
Information Technology and Electrical Engineering
XXXVIII Cycle

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

From Images to Videos: Analysis and Detection of AI-Generated Media

by

RICCARDO CORVI

Advisor: Prof. Luisa Verdoliva



SCUOLA POLITECNICA E DELLE SCIENZE DI BASE

DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELLE TECNOLOGIE DELL'INFORMAZIONE

*To my mother, for all the support given to me
throughout the years.*

FROM IMAGES TO VIDEOS: ANALYSIS AND DETECTION OF AI-GENERATED MEDIA

Ph.D. Thesis presented
for the fulfillment of the Degree of Doctor of Philosophy
in Information Technology and Electrical Engineering
by

RICCARDO CORVI

December 2025



Approved as to style and content by

Luisa Verdoliva

Prof. Luisa Verdoliva, Advisor

Università degli Studi di Napoli Federico II

Ph.D. Program in Information Technology and Electrical Engineering

XXXVIII cycle - Chairman: Prof. Stefano Russo



<http://itee.dieti.unina.it>

Candidate's declaration

I hereby declare that this thesis submitted to obtain the academic degree of Philosophiæ Doctor (Ph.D.) in Information Technology and Electrical Engineering is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.

Parts of this dissertation have been published in international journals and/or conference articles (see list of the author's publications at the end of the thesis).

Napoli, December 10, 2025

A handwritten signature in black ink, reading "Riccardo Corvi". The signature is written in a cursive style with a horizontal line underneath.

Riccardo Corvi

Abstract

Synthetic media generation has improved enormously in the previous years and is now available to everyday users. Powered by large language models, these new generators allow users to create incredibly detailed content starting from a simple text description. These are incredible resources for all types of creative works. However, they provide extraordinarily powerful tools for malicious actors to spread disinformation. As such, Multimedia Forensics is more relevant than ever with the need for general and robust detectors to distinguish real from generated content. This thesis aims first to conduct a systematic study of a large number of generators to discover the most relevant characteristics for forensics applications. This initial analysis has shown that all generators leave artifacts in the generated content and that there are some major discrepancies in the mid-high frequencies between real and synthetic images. Then, the power of large pre-trained Vision-Language Model has been exploited to design a lightweight synthetic image detector based on CLIP. Such models have also been used to develop a robust strategy for the task of fully synthetic video detection. To this end, a novel forensic-oriented data augmentation strategy based on the wavelet decomposition is introduced, in which specific frequency-related bands are replaced to drive the model to exploit more relevant forensic cues. In fact, a well-designed forensic classifier should focus on identifying intrinsic low-level artifacts introduced by a generative architecture rather than relying on high-level semantic flaws that characterize a specific model. This approach improves the generalizability of the detectors, without the need for complex algorithms. Despite its simplicity, the method achieves a significant accuracy improvement over SoTA detectors and obtains excellent results even on very recent generative models.

Keywords: Image Forensics, Deepfakes, Synthetic Image Detection, Synthetic Video Detection

Sintesi in lingua italiana

La generazione di media sintetici è migliorata significativamente negli ultimi anni, ed è oramai disponibile per gli utenti di tutti i giorni. I generatori più recenti permettono di creare contenuti dettagliati a partire da una semplice descrizione testuale. Da un lato questi strumenti sono straordinari per qualsiasi applicazione creativa. Tuttavia, permettono ad agenti malevoli enormi potenzialità per diffondere disinformazione. Di conseguenza, il "Multimedia Forensics" è diventato sempre più importante con la necessità di detector che siano robusti e possano distinguere i contenuti reali dai generati. Nella tesi viene condotta una analisi di svariati generatori per identificare le caratteristiche forensi più rilevanti per distinguere tra immagini reali e sintetiche. Questa analisi ha mostrato che tutti i generatori lasciano artefatti nelle immagini e che le immagini reali e sintetiche presentano delle discrepanze alle frequenze medio-alte. È stata poi sfruttata la forza di un grande Vision-Language Model preaddestrato per progettare un semplice detector di immagini sintetiche basato su CLIP. Questi modelli sono stati poi utilizzati per sviluppare una strategia robusta per la detection di video sintetici. È stata sviluppata una nuova data augmentation per scopi forensi basata sulla decomposizione wavelet, in cui specifiche bande sono sostituite per spingere il modello a concentrarsi sugli artefatti di basso livello. Infatti, un classificatore forense ben progettato dovrebbe concentrarsi sulle tracce intrinseche di basso livello introdotte dal generatore e non su difetti di alto livello dello specifico modello. Questo approccio migliora la generalizzazione del modello senza la necessità di complessi algoritmi. Nonostante la semplicità, il metodo ottiene significativi miglioramenti rispetto i SoTA detector e ottiene ottimi risultati anche su generatori recenti.

Parole chiave: Deepfake, Rilevamento di Immagini Sintetiche, Rilevamento di Video Sintetici, Analisi forense di contenuti multimediali

Acknowledgements

The author's work has been supported by a PhD scholarship funded by the Defense Advanced Research Projects Agency (DARPA) under the SemaFor program. The project is called DISCOVER (a Data-driven Integrated approach for Semantic inCONSistencies VERification), agreement number FA8750-20-2-1004. This work has been also supported by NVIDIA.

I would like to express my deepest gratitude towards my advisor, Professor Luisa Verdoliva, for the unwavering support and mentorship throughout my research activities. Her guidance has played a crucial role in my academic and personal growth.

I am deeply grateful to Professor Cozzolino for his infallible support. His expertise and willingness to address every doubt has been invaluable throughout my PhD.

To all my colleagues in the GRIP research group, thank you for sharing this passion and experience with me.

I extend my sincere thanks to the the AI-Mediated Reality and Interaction (AMRI) Research group at NVIDIA, especially to my supervisors, Shalini De Mello and Koki Nagano, for their mentorship during my internship. Their guidance helped me navigate challenges and they have provided me with the opportunity to experience working in a large company.

To my friends, both the old ones and the new ones made along the way, thank you for always standing by my side and helping me bear the weight of this journey. I want to especially thank Pasquale, with whom I have shared a lifetime, and Giampiero, for being there every day.

To my family, thank you for your support and encouragement. Your love and comfort have been my source of strength and have helped me through the most challenging times. I want to especially thank my brother for always being there when I needed him the most.

Finally, to my mother, your strength and love have made me in the person I am today. I will never be able to repay you for everything you have given me. Thank you.

Contents

Abstract	i
Sintesi in lingua italiana	iii
Acknowledgements	vi
List of Acronyms	xi
List of Figures	xvi
List of Tables	xix
1 Introduction	1
1.1 Contributions	3
1.2 Outline of the Thesis	5
2 Detection of synthetic content	7
2.1 Synthetic Images	8
2.1.1 Generative Adversarial Networks	9
2.1.2 Diffusion Models	11
2.1.3 Synthetic Image Detection	12
2.1.4 Synthetic Image Attribution	21
2.2 Detection of AI-generated videos	27
3 Artifact analysis of synthetic content	29
3.1 Forensic Analysis	32

3.1.1	Architecture related traces	34
3.1.2	Training dataset bias	36
3.1.3	Fidelity to natural images	38
3.2	Conclusions	41
4	Leveraging Vision Language Models for the detection of synthetic images	43
4.1	Datasets and metrics	46
4.2	CLIP for synthetic image detection	48
4.2.1	Influence of the classifier and layer	49
4.2.2	Influence of the reference set size	49
4.2.3	Influence of the reference set content	51
4.2.4	Influence of pre-training	52
4.3	Comparison with the state-of-the-art	53
4.4	Beyond low-level forensic traces	60
4.5	Few-shot analysis in the wild	63
4.6	Discussion	65
5	Analysis and detection of synthetic videos	67
5.1	Dataset	70
5.2	Proposed Method	71
5.2.1	Video forensic artifacts	71
5.2.2	Training strategy	76
5.3	Results	79
5.3.1	Experimental setup	79
5.3.2	Comparison with the State-of-The-Art	81
5.4	Ablation study	85
5.5	Conclusions	90
6	Conclusions	95

Bibliography	97
Author's publications	115

List of Acronyms

The following acronyms are used throughout the thesis.

GAN	Generative Adversarial Network
DM	Diffusion Model
VQ-GAN	Vector Quantized Generative Adversarial Network
VLM	Vision Language Model
CLIP	Contrastive Language-Image Pre-Training
VAE	Variational Autoencoder
AUC	Area Under the Curve
NLL	Negative Log-Likelihood
ECE	Expected Calibration Error
AP	Average Precision
SoTA	State of The Art
Pd@5	Probability of Detection at a 5% False Alarm Rate

List of Figures

2.1	Examples of synthetic images and artificial fingerprints of the following models: Latent Diffusion, Stable Diffusion, Midjourney v5, DALL·E Mini, DALL·E 2, DALL·E 3. . . .	9
2.2	Pipeline of a Generative Adversarial Network	10
2.3	Pipeline of a Diffusion Model	11
2.4	Taxonomy of synthetic image detection methods.	13
2.5	Synthetic image detection results in terms of AUC with and without post-processing (PP) of the state of the art methods.	19
2.6	Synthetic image detection results in terms of AUC for the three best state of the art methods over various contents.	19
2.7	Synthetic image detection results in terms of AUC on images from X generated by DALL·E 3, Midjourney and Firefly.	20
2.8	Scores distribution provided by the method [60] (trained on ProGAN) for real images, and for synthetic images from ProGAN and Firefly.	21
2.9	Synthetic image attribution results of state of the art methods in both closed-set and open-set scenarios.	24
2.10	Robustness analysis in both closed-set and open-set scenarios of state of the art methods for attribution	25

2.11	Synthetic image attribution results in the more challenging scenario where also the real class is included.	26
3.1	Various examples of synthetic images from both GAN and Diffusion Models alongside their artificial fingerprints. . . .	30
3.2	Examples of real and synthetic images drawn from popular datasets.	31
3.3	Power spectra (top) and autocorrelation functions (bottom) of noise residuals of three slightly different latent diffusion models.	34
3.4	Effect of post-processing operations on the autocorrelation of noise residuals.	36
3.5	Autocorrelation of noise residuals for real and synthetic images.	37
3.6	Radial spectrum power density. Synthetic images are compared with the real images used for training the correspondent model.	38
3.7	Angular spectrum power density (top) and the square-root of Fisher's discriminant ratio applied at each direction (bottom).	40
4.1	Area Under ROC Curve (AUC %) on unseen synthetic generators (x -axis) and on post-processed data (y -axis) of our proposal and other state of the art methods	44
4.2	Examples of synthetic images from generators used in our experiments.	45
4.3	Average AUC for our approach considering different classifiers and different features on the original dataset (top) and after random compression and resizing (bottom).	50
4.4	Performance of the CLIP-based detector as a function of the number of real and synthetic images in the reference set. . .	50

4.5	Performance of the CLIP-based detector as a function of the pre-training.	53
4.6	Robustness analysis in terms of AUC carried out on the Stable Diff. 2, SDXL, DALL·E 3, DALL·E 2, Midjourney, and Firefly generation models of the SynthBuster dataset [8].	58
4.7	A synthetic image generated by Stable Diffusion XL [111] and its $4\times$ decimated version.	60
4.8	Scatter plots of scores provided by Corvi et al.[31] (y -axis) on $2400+2400$ images sampled from the 18 datasets of Tab.4.1.	63
5.1	Examples of video from our proposed test dataset generated with recent state of the art generators.	68
5.2	Schema of our proposed training-time data augmentation for synthetic video detection.	69
5.3	A real video from the dataset proposed in [26] and synthetic videos generated using the associated prompt.	72
5.4	Spatial and temporal-spatial power spectra of OpenSora-Plan videos before and after compression, compared to those computed from a real (compressed) video.	74
5.5	For each generator (Allegro, CogVideo X1.5, NOVA), we show the close up of its spatial power spectrum $S_{yx}(u, v)$ (left), and the close up of its temporal-spatial power spectra $S_{yt}(u, w)$ (right), both before compression (top) and after compression (bottom),	75
5.6	Average distance in the frequency domain as defined in eq.5.5, before compression (left) and after compression (right). . . .	76
5.7	For each real video, four corresponding fake versions are generated: one with all fake subbands and others with selective replacements.	77

5.8	Comparison with SoTA methods on 16 generative models across different evaluation metrics.	80
5.9	Some examples of videos from our dataset that we created with recent state of the art generators.	92
5.10	Some examples of videos from our dataset that we created with recent state of the art generators.	93

List of Tables

- 2.1 List of the AI generative models analyzed in the experiments
to evaluate the state of the art. 18

- 3.1 Main properties of the generative models analyzed. 32

- 4.1 Image generators used in our experiments: GAN-based, DM-
based, and commercial tools. 47
- 4.2 AUC performance of the CLIP-based detector for various
real/fake data, with and w/o augmentation (resizing/com-
pression). 52
- 4.3 List of methods used for the comparisons. 54
- 4.4 Comparison with SoTA methods in terms of AUC on images
generated by Diffusion Models. 54
- 4.5 Comparison with SoTA methods in terms of AUC on images
generated with GANs and commercial tools. 55
- 4.6 Comparison with SoTA methods in terms of AUC and Ac-
curacy. 56
- 4.7 Comparison with SoTA methods (AUC) on images from Dif-
fusion Models post-processed by random cropping, resizing
and compression. 57

4.8	Comparison with SoTA methods (AUC) on images from GANs and Commercial tools post-processed by random cropping, resizing and compression.	58
4.9	AUC/Accuracy in the presence of post-processing.	59
4.10	Results in terms of AUC without attacks (first group of columns) and after attacks on synthetic images (second group) and real images (third group).	61
4.11	AUC/Accuracy results for Corvi et al. [31], proposed method (10k and 10k+ versions), and their fusion over the three families of generators: GAN, Diffusion, Commercial Tools. .	62
4.12	Results in terms of AUC and accuracy of proposal and SoTA methods on real and synthetic images download from X. . .	64
5.1	Comparison in terms of balanced accuracy with SoTA methods trained on their original datasets.	81
5.2	Comparison in terms of balanced accuracy with SoTA methods re-trained using Pyramid Flow.	83
5.3	Comparison in terms of AUC with SoTA methods both trained on their original datasets and re-trained using Pyramid Flow.	83
5.4	Comparison in terms of Pd@5 with SoTA methods both trained on their original datasets and re-trained using Pyramid Flow.	84
5.5	Comparison in terms of NLL with SoTA methods both trained on their original datasets and re-trained using Pyramid Flow. .	84
5.6	Comparison in terms of ECE with SoTA methods both trained on their original datasets and re-trained using Pyramid Flow.	85
5.7	Influence of the augmentation.	86
5.8	Influence of the backbone.	87

5.9 Performance by varying L (number of wavelet decomposition levels). 87

5.10 Performance by varying compression quality (CRF) and the number of wavelet decomposition levels (L). 88

5.11 Performance by varying the type of generator and the size of the dataset during training. 89

5.12 Performance by varying the number of generators during training in terms of AUC and balanced Accuracy. 90

Chapter 1

Introduction

Man is condemned to be free; because once thrown into the world, he is responsible for everything he does.

Jean-Paul Sartre

Over the past decade, there has been tremendous progress in the creation of synthetic media, primarily due to the development of powerful methods based on generative adversarial networks (GAN). The past few years, methods based on diffusion models (DM) have gained the spotlight. Originally, they were able to create only a very limited number of contents with each generator capable of creating only one subject, with many approaches focusing on faces. Another class of generators, instead, aimed at converting images between two types for example from summer to winter and vice versa. Although impressive, these early models were limited in their usability due to the lack of user control. However, these tools, which were once limited, have evolved significantly, allowing for incredible control in the generation process. In addition to providing surprising levels of detail, they enable the creation of impressively realistic visual content starting from a simple textual description. Powered by the advancements in large language models, they now provide the possibility of generating with ease and with fine-grained control, allowing to determine the style and subject of what the user wants to generate. Commercial tools such as Midjourney, Adobe Firefly, and Flux have made them widely accessible by providing easy to use UIs and ease of access. This has allowed for infinite

creative possibilities, stimulating users to explore their imagination freely, and share their creations online on social media. They are not only meant for individuals to use, but are nowadays widely utilized in many industries from marketing and video-game production to individuals for their day to day activities, like creating visuals for their presentations. However, these incredibly powerful technologies have also given to malicious actors new options when spreading disinformation online. The dangers that these tools can pose in the wrong hands are clear, the power to create fully synthetic images or videos can be used for propaganda, to influence the public opinion or elections, and damaging reputations. Especially, combined with the widespread influence of social media, the reach of these networks spreading fake news is tremendous. This threat is further amplified by the general carelessness of users who do not question the truthfulness of the content. In fields like journalism, politics, and the judiciary, verifying the authenticity of the content becomes of paramount importance. Various governments around the world have started to address the serious issue posed by these threats, on one hand by passing new laws to combat these dangers, on the other by founding research projects aimed at developing tools capable of detecting such attacks. One such initiative funded by the European Union is the Verification Assisted by Artificial Intelligence (Vera.ai) project. Its aim is to build trustworthy solutions based on Artificial Intelligence to fight disinformation. Alongside these public initiatives, various companies have started to develop their own tools to protect their users. The Coalition for Content Provenance and Authenticity (C2PA) was founded as a way to provide a common standard for publishers, creators and consumers to determine the origin of and the editing history of the data. With members such as Adobe, Amazon, Google, Meta, Microsoft, OpenAI and Sony, the objective of the standard is to track both the real and the synthetic content. While not aimed directly at detecting the synthetic content, these initiative can aid the fact-checkers in determining the truth behind a content or aid them in finding the necessary proof that the content is fake. With the advent of such powerful tools, multimedia forensics and related fields have seen a drastic increase in activity to develop new tools to face these new threats. Among these, many have been proposed for the detection of synthetic images and more recently for the detection of synthetic videos. Although these images are nowadays indistinguishable from real

ones, the generators leave imperceptible traces, invisible to the naked eye but that can be used to detect whether the content is real or synthetic. These traces, which are specific to each generator, are tied to their architecture and are related to the up-sampling step of the model. While many advancements have been done so far, the SoTA detectors are still struggling in the wild. These deficiencies are tied mainly to two key factors:

- i) limited generalization;
- ii) limited robustness;

One major issue is the speed with which modern generators are being developed. Despite the fact that all synthetic content contains some traces, these depend on the specific generator, and can vary a lot from generator to generator. As such, it is quite difficult to develop techniques which, despite being trained on one or a limited number of generators, can learn features capable of generalizing to many types of generators including those that are yet to come. Furthermore, the traces left by the generators used by the detectors, are quite susceptible to image perturbations which are commonly done on images posted on social networks, such as JPEG compression, blurring and resizing. Although not performed with malicious intention, these operations can hamper the traces making the detection more challenging.

1.1 Contributions

The aim of this thesis is to tackle these challenges and develop robust and general methods for the detection of synthetic content. A crucial step in developing increasingly effective detectors is understanding which image features can help distinguish fake images from real ones. While some work have analyzed the generation traces present in the synthetic images working in the Fourier Domain, these works were only focused on artifacts caused by GAN models, such as those due to the up-sampling steps used to achieve the desired spatial resolution [44, 165]. Little attention was paid to the more recent architectures like the Diffusion Models (DM) and the autoregressive models based on transformers. Hence, one main contribution has been the analysis of such traces in the frequency domain for

these recent models. More specifically, we conducted a study of the radial and angular spectra that showed how synthetic generators are not able to reproduce the spectra of real images especially at the middle frequencies along the diagonal directions.

Another contribution has been an extensive evaluation of SoTA methods designed for the detection of GAN images for new types of generators. To evaluate their performance in a more realistic scenario, we also consider what happens after they have undergone post-processing operations. We further extended this analysis to the attribution task both in a closed set and in an open set scenario. This analysis highlighted the major challenges in the field and identified the areas where further research is needed.

We then tackled the issue of generalization. Generalizing to Out of Distribution data is one of the major issues of deep-learning and a topic of intense research in recent years. Large pretrained vision-language models have brought a number of new and exciting solutions due to their capability of being excellent zero-shot and few-shot learners across many different type of applications. As such, alongside other methods which have proposed this idea, we explored the potential of CLIP for the detection of synthetic images and conducted an extensive experimental evaluation in a real world scenario involving a large number of generative models. The developed approach shows impressive generalization and very good robustness, resulting in strong performance improvements when compared with the state of the art. We obtain these results by leveraging the descriptive power of the CLIP features without relying on intensive training or domain-specific data. To train the detector, we used a small number of paired real/fake images coming from just one generator where the caption of the real images are used to create the synthetic ones. Despite only a single method being used to generate the training data, the performance is equally good across many different data sources.

To further push the field, in the period spent at the Centre for Research and Technology Hellas (CERTH), we organized the MediaEval 2025 challenge as part of the Vera.AI project. The challenge was comprised of two tasks. The first focused on evaluating the submissions in an in-the-wild scenario, evaluating them using also a dataset of real and fake images downloaded from social media. The second task focused on the objective of detecting and localizing forgeries done through text-guided artificial in-

telligence methods. The test dataset contained both spliced images, where the manipulation is put back in the original image, and fully regenerated images, where the image is recreated with the only difference being the desired manipulations. The latter scenario is particularly challenging, since the image is fully synthetic and the aim is to find the manipulation which is only semantically local.

For the final contribution, we worked on synthetic video generation. Few approaches have been proposed to tackle this problem, but they often exhibit poor generalization, which limits their use in the wild. To this end, we developed a new data augmentation which proposes a forensic-oriented strategy based on wavelet decomposition. We propose to replace specific frequency-related bands to push the model to rely more on the cues that we have found previously and that are more relevant. Our proposed training framework improves the generalization over most of the state-of-the-art methods that propose complex algorithms and large datasets with multiple generators. Our method obtains a very high accuracy improvement over state-of-the-art detectors despite its simplicity. Notably, the method ranked first in the SAFE Synthetic Video Detection Challenge 2025 [124]. This challenge simulates a realistic scenario where no information about the test models are known in advance and videos vary in resolution, bitrate, frame rate, and length. This confirms the advantage of our training data and augmentation strategy to ensure good generalization ability.

1.2 Outline of the Thesis

The rest of the thesis is structured as follows.

Chapter 2 introduces the topics and challenges related to synthetic image detection and synthetic video detection.

Chapter 3 presents our analysis of the artifacts left in synthetic images and highlights their difference from real images.

Chapter 4 describes our approach to leverage a large vision language model (CLIP) for the detection of synthetic images.

Chapter 5 introduces the artifacts present in synthetic videos and our approach for their detection.

Chapter 6 summarizes the main contributions of this thesis.

Chapter 2

Detection of synthetic content

*It's the job that's never started as
takes longest to finish.*

J.R.R. Tolkien

Generative AI offers numerous opportunities for many kinds of industries, from entertainment, to healthcare, to finance and manufacturing. However, it can be used for all kinds of illicit purposes, to provide evidence to support disinformation campaigns and political propaganda. Even more so that these goals can now be pursued faster and at a larger scale than before with minimal to no human intervention producing results that are incredibly realistic and that are aligned perfectly with any desired narrative. This poses an incredibly serious threat to our society and enforces the need of a growing focus on tools capable of distinguishing synthetic images from natural ones. In this context, two slightly different objectives can be pursued:

- i. detection gives a global score that evaluates the probability that the image is synthetic
- ii. attribution goes a step further and aims to trace the specific generative model used to synthesize the image. By providing more specific information about the generation process, attribution validates the detection output and improves its interpretability.

However, while many of the state-of-the-art methods work well when tested

on the content as it is when generated, the most common way of sharing images and videos is through social media. These sites do routine post-processing which, despite not having malicious intentions, can make it much harder to detect whether something is synthetic or real. The first generative topic that has been tackled is the creation of images. While the first generator of synthetic images has been proposed in 2014 by Goodfellow et al. [59], it's only in the past years with the advent of Diffusion Models (DM) that the quality has reached the point of producing images indistinguishable from real making the detection of synthetic images of paramount importance. These high quality generators have then been adapted to generate multiple images that show the evolution through time pushing the research of the generation of synthetic videos to the forefront. Unlike for images, in just a few years the generation of synthetic videos has made incredible strides leading to generators capable of producing incredibly realistic videos. As such, the topic of their detection has been getting an increasingly pressing issue. However, while for images many approaches have been proposed, due to the novelty of these generators, there is a lack of methods proposed for the detection of synthetic videos.

In this Chapter, we are going to go through the methods for both the generation and detection. Firstly, we are going to tackle the issue of generation of synthetic images. We later delve into the detection and attribution of synthetic images. On top of describing these methods, we are also going to present the results of the state of the art methods on a dataset of synthetic images both as they are when generated and in a more challenging scenario after post-processing. Finally, we are going to close this chapter by discussing the methods for the generation of synthetic videos and finish with discussing the SoTA approaches for the detection.

2.1 Synthetic Images

Throughout the years, many powerful generative approaches have been proposed, such as variational autoencoders, energy-based models, normalizing flows, autoregressive models, generative adversarial networks, and diffusion models. In particular, we will focus on the last two type of generators due to their ability of generating high quality images and for their easy support for text-image synthesis. In fact, natural language-based im-

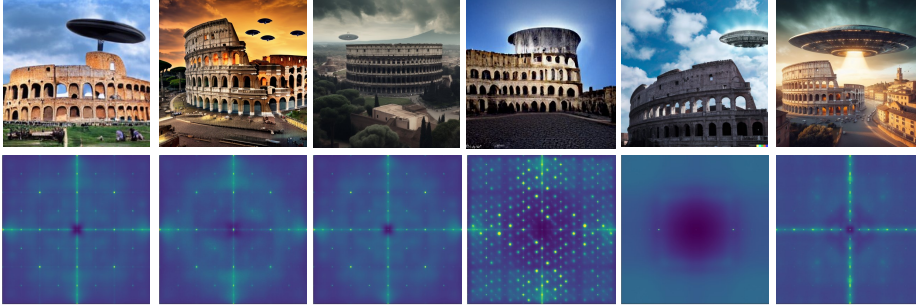


Figure 2.1. Top: examples of synthetic images, generated using (from left to right) Latent Diffusion, Stable Diffusion, Midjourney v5, DALL-E Mini, DALL-E 2, DALL-E 3. The prompt used for their generation is the following: *a photo of the Rome Colosseum with a UFO over it, detailed, 8k*. Bottom: Average Power Spectra of the artificial fingerprints for each of such model. Forensic artifacts are clearly visible as spectral peaks in the Fourier domain, stronger or weaker based on the specific model. We can observe that the first three images share very similar artifacts while the fingerprints of the three releases of DALL-E differ greatly from one another, testifying to very different generative architectures. We will delve deeper on these artifacts in Chapter 3

age generation provides incredible level of flexibility and control over the generation process leading to new and more advanced applications. We are going to first describe the methods for the generation and later the methods for detection and attribution.

2.1.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs), introduced in [59], leverage the adversarial game between two networks, a generator creates a synthetic image and a discriminator is tasked with distinguishing it from a natural image as shown in Figure 2.2. These networks are trained jointly with a min-max game. As the discriminator improves, so does the generator, which learns to create more realistic samples over time. One of the first successful GAN-based methods is BigGAN, a class-conditional image generator [13]. ProGAN [73] has brought further improvement by producing high-quality images using a fast and stable image generator with a progressive procedure that increases the quality of the images over time.

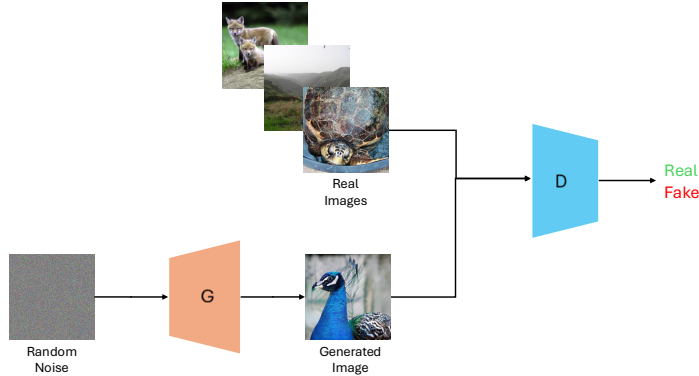


Figure 2.2. Pipeline of a Generative Adversarial Network. The generator (G) creates synthetic images that the detector (D) tries to distinguish from real samples.

With further improvements it has led to the StyleGAN[75, 76, 74] family of generators, where the convolutional kernels of the generator are controlled by a latent code that allows for the control of the synthesis process. Noticeable results have been obtained when introducing 3D information in the synthesis with E3GD, a method capable of generating multi-view-consistent renderings and detailed geometries of synthetic faces. Recently, research has been shifting the focus to text-based image synthesis, and several GAN-based methods have been proposed for this purpose. Two proposed approaches are StyleGAN-T[126] and GALIP[138], both have been trained on both images and text using as the underlying language model CLIP (Contrastive Language-Image Pre-Training) [115] and are capable of generating high-quality images controlling them through text. GigaGAN[72] brought further improvement, it was the first GAN-based method to be trained on a billion real-world images, it is capable of synthesizing high-resolution images very quickly and it supports latent interpolation and stylization.

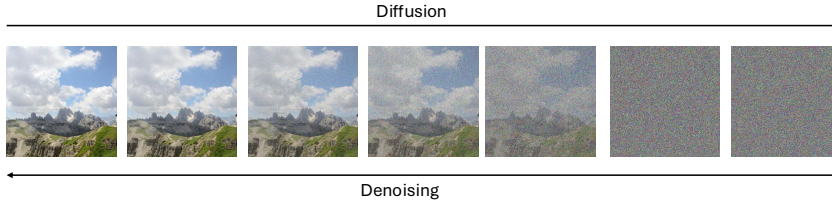


Figure 2.3. Pipeline of a Diffusion Model. During the first phase the model is trained by adding stronger and stronger noise to an image and teaching the model to denoise it. Once trained, the model can start from a pure gaussian noise and produce an image.

2.1.2 Diffusion Models

Diffusion models have two interconnected stochastic processes at their core. The pipeline can be seen in Figure 2.3, and some examples are visible in Figure 2.1. The forward process transforms natural images into random noise by adding Gaussian noise in small steps. The samples generated in a single step of the forward process are then used to train a neural network to invert that step, removing the noise added from the input sample. Such networks are chained to perform the backward process, gradually converting input Gaussian noise into synthetic images. Nowadays, Diffusion Models (DMs) have surpassed the generation quality of GAN and other approaches [65, 40]. Not only that, but they also provide a easier and more stable training than GANs, since they do not suffer from model collapse, although they are much more time consuming in terms of both training and inference. But even more important it is their flexibility, DMs provide excellent support for text to image synthesis, enabling the generation of complex images based on diverse and arbitrary text descriptions. This has caused a revolution in how complex generative artificial intelligence tasks are tackled, and, as a result, many different architectures for text-image synthesis have been proposed. The first models leverage a U-Net and its variations as backbone, like GLIDE[103] and DALL-E 2[117], which uses a text encoder based on CLIP to condition the generation process on a natural description. Ediff-I[7] further improves on it adopting multi-

ple U-Net models specialized for different synthesis stages. However, this approach is quite onerous. In order to reduce computational costs, Latent DM[120] combines a diffusion model with a variational autoencoder, where the former operates in the low-dimensional space of the latter and creates the vector of the desired image that gets reconstructed by the decoder. A notable generator of this type is Stable Diffusion[122], it is part of an open-source project and is trained on the 5.85 billion images from the LAION dataset [128]. Instructpix2pix[14] is a fast solution that can generalize to real images and user-written instructions to modify part of the image. It combines the knowledge of GPT-3 and Stable Diffusion. A further improvement was made by Stable Diffusion XL [111] that leverages a three times larger UNet backbone to generate very high resolution images. In Diffusion Transformers (DiTs)[110], instead, the usual U-Net backbone is replaced by a transformer. On the backbone of the latter, further improvements have been made to scale this type of solution. Stable diffusion 3 [49] has been a further improvement of the open-source project that has switched to the transformer approach as well. Another final great improvement has been introduced by DALL-E 3[105], while still using the U-Net approach, it shows the importance of the captions used during training and it introduces an appropriate recaptioner to improve the quality.

2.1.3 Synthetic Image Detection

Early methods developed to distinguish if an image is real or synthetic leveraged CNN-based architectures trained on large amounts of data. While these approaches work extraordinarily well when tested on data which is aligned with the training set, they exhibit a significant performance drop when there is a mismatch between test and train. In particular, the two major problems are the lack of robustness and the limited generalization capabilities. Robustness is critical to resist the impairments caused by the routine post processing that the images go through when uploaded on social media, like re-compression and re-sizing, that weaken the subtle traces exploited by the detectors. On the other hand, generalization is essential for the method to detect images coming from generators not seen during training, especially those that are yet to come. In the following, we will review the main strategies proposed to handle such issues. It is important to highlight that most papers described in this Section focus

on GANs, but in recent years there has been an ever increasing attention to DMs. However, methods conceived originally for GAN image detection usually turn out to work equally well on more recent AI-generative approaches. A taxonomy of all the methods is presented in Figure 2.4. On top of describing these methods, we will also evaluate the performance of some of the best methods on images generated with GANs and Diffusion Models. These results work as the base of the following research conducted.

Forensic artifacts

The images generated by synthetic architectures often used to show visual artifacts, like unnatural colors or incorrect perspectives and shadows. Usually, these types of semantic inconsistencies are called semantic artifacts. They were far more present in the earlier methods. However, the most recent ones, especially the most recent Diffusion Models, can generate images which are visually perfect with no indication of them being synthetic. This has led to great concern among the end users. Without these traces, it is much harder to determine at a glance if an image is synthetic or not. However, despite the images looking perfectly natural, it is still possible to distinguish them from the real ones using the correct methodologies. The general basis is that each image has various distinctive artifacts that are related to the acquisition or generation process that can be leveraged to trace it back to its origin. For physical devices this is notorious, since both hardware and software leave precious forensics traces. However, a

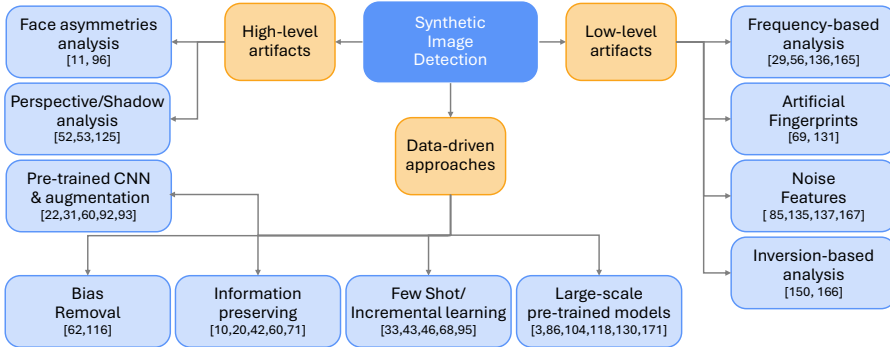


Figure 2.4. Taxonomy of synthetic image detection methods.

similar process happens also for synthetic images. AI-based generative models use complex processing chains involving a large number of specific processes, including filtering, pooling, downsampling and upsampling. All such processes leave peculiar marks in the images which may be exploited to accomplish multiple forensic tasks, from source identification to forgery detection [25]. These are imperceptible traces, called also low-level artifacts, that can be exposed only by means of statistical analyses. We will discuss it further in Chapter 3.

Data-driven methods

The first possible strategy to produce a robust detector is to leverage deep CNN architectures [93] and training them using suitable augmentation. [148] showed that it is possible to achieve good robustness even if the network is trained on a single generative architecture (ProGAN[73]) through simple augmentations like compression and blurring. One of the key aspects of the proposal is also the diversity of the training set, they ensure that by using images from 20 different categories. These insights have been further confirmed by subsequent papers [92, 60] which have shown that it is of key importance to improve the generalization ability of the detector. Another relevant element is the adoption of a large dataset for model pre-training. [60] shows that extreme augmentation, while providing only marginal improvements to the robustness, leads to a significant improvement to the generalization on unseen models.

Furthermore, another relevant principle, that applies just as well to the detection of both GAN and DM images, is to avoid any loss of information. This applies both during training and test and to all layers of the network architecture but especially to those closest to the input. In [20] this is achieved by using a patch-based classifier and avoiding image resizing, in order not to erase the subtle traces left by the generation process. To preserve the invisible forensics cues, in [60] it is explicitly suggested to: *i*) train the network on randomly cropped patches; *ii*) make the final decision on the whole image by means of some fusion strategy; *iii*) avoid any downsampling in the first layers of the network. In [71] patch-based analysis is further enhanced by combining it with global spatial information extracted from the whole image.

To improve generalization, [22] has the objective of singling out trans-

ferable features usable to design universal detectors. One of the most important transferable features is shown to be color when used in a suitable data augmentation scheme. Another avenue to pursue generalization is to use few-shot or incremental learning strategies, as done in [33, 95, 43, 68]. These approaches need the availability of a few example samples from the new architecture in order to function. However, such data cannot be taken for granted in the most challenging scenarios. In order to explore this approach, a recent work [46] has investigated whether a detector could perform correctly in a simulated online framework. The detector is regularly retrained by preserving the temporal order by release date with which the synthetic generators have been made available. As shown by the results, the generalization is good as long as the new models are closer to the architecture of the old ones.

Furthermore, [31] conducted an investigative analysis with several CNN-based methods for the detection of Diffusion Model images. The results show that one of the critical aspects of the detectors to work across different generators is calibration and that some form of fusion strategy could help. In [42] it is shown that in order to ensure generalization across many different categories, only higher-quality images should be included in the training dataset. While much has been done to improve the performance of the detectors, their explainability has been explored only recently with a few works that leverage Gradient Class Activation Mapping to interpret the results [10].

More recently, works are considering backbone architectures which are different from CNNs, in particular transformers and vision-language models in [104, 118, 3, 171, 79]. [104] is particularly of interest as the first to demonstrate the ability to generalize of the features of the CLIP-ViT model pre-trained on internet-scale image-text pairs. They use the fixed feature space of the model to perform classification, and, despite being trained on the images generated with ProGAN created by [148], the detector is able to generalize and obtain very good performance even on images generated by DMs. A different strategy was pursued in [130, 86] that suggests using during training also the prompt associated with the image. [118, 3, 171] made available to the community large datasets of synthetic images to foster research on synthetic image detection for generative AI. One more recent avenue currently being explored in research is to focus on

the images used during training. [116, 62] have highlighted the importance of removing as many biases as possible from the training data. In order to do so, they propose to generate the synthetic data from the real images. In [116] they propose to use the autoencoder of a Latent Diffusion Model to have images that are indistinguishable from the original but now contain the artifacts of the generation process. [62] goes one step further by also performing diffusion steps to leverage even subtler inconsistencies at the lower frequencies and augmenting the original dataset with inpainted versions of real images.

Methods exploiting forensic cues

In this section, we describe methods that more explicitly rely on some specific forensic traces both low-level and high-level cues.

Low-level artifacts As said previously, AI-generated images show clear traces of their origin as peaks in the Fourier domain. These artifacts are due to the up-sampling operations commonly used in the synthesis network [44]. However, as we will show in Chapter 3, even when such peaks are absent, synthetic images have significant differences from real images in the medium-high frequencies. [165] is the first work to explore the use of these artifacts in the Fourier domain for the purpose of detecting synthetic images. Firstly, a pipeline is introduced to simulate these types of artifacts and then use them to train a spectrum based classifier. There are also other methods pursuing the same idea. In [56] the synthetic images are obtained through an adversarial autoencoder. While [69] proposes to simulate the fingerprints from real images by leveraging various types of generative models.

These low-level artifacts in the high-frequency traces can be also exposed by removing the scene content and extracting the noise residual [137]. In these noise patterns, it is possible to exploit, for the detection of synthetic images, the difference in inter-pixel dependencies between real and synthetic images [167, 135]. These noise patterns can be extracted in various ways, it can be learned as done in [85, 131] or estimated during the inverse diffusion process [166]. The inversion process can be used in other useful ways for the detection of synthetic images as shown in [150]. The main intuition is that DM-generated images can be accurately reconstructed by a DM while real images cannot. Therefore, it is possible to

carry out the detection by measuring the error between an input image and its reconstructed version. We will discuss in further detail these artifacts and their nature in Chapter 3.

High-level artifacts While they are less prevalent, there are some methods that focus in exploiting semantic errors like asymmetries in faces, wrong perspective and odd shadows. Some of them are focused on faces [96, 11] while other try to tackle generic scenes [52, 53, 125]. In particular, [125] a classifier is proposed that looks at the perspective field, lines and the relation between the objects and their shadow, based on the observation that inconsistencies in these elements are systematic in Diffusion Models.

2.1.3.1 Experimental evaluation

We conducted an experimental analysis of some of the methods described in the previous section, to evaluate the state of the field with the advent of Diffusion Models. We perform a comparison of the methods proposed for synthetic image detection in the literature for which code and model were publicly available: Wang2020 [148], PatchFor. [20], Grag2021 [60], Liu2022 [85], Corvi2023 [31], LGrad [137], Ojha2023 [104], and DIRE [150]. For each method we use the model already trained by the authors as proposed in their original paper. For our evaluation we used both GAN- and DM-based image generators, in Tab. 2.1 we highlight the model used.

The test set comprises 1,000 synthetic images for each of the generators listed in Table 2.1. To avoid biases induced by different contents, we evaluated the performance on the synthetic images of each generator comparing with real images with the same content. In particular, we used 5,000 real images, 1,000 per each dataset: LSUN, FFHQ, ImageNet, COCO and LAION. For conciseness, we aggregate the results for all GAN-based generated images and all DM ones.

We also consider a post-processed (PP) version of the datasets to simulate a realistic scenario on the web. Firstly, an image crop is taken of a random size that can vary in a range from $\frac{5}{8}$ to the full image size. The image is then resized to 200×200 pixels, and finally it is JPEG compressed with a random quality factor between 65 and 100.

In Fig. 2.5 results are shown in terms of Area Under the ROC Curve (AUC). We can observe that DM images are harder to detect than GAN images. This likely stems from most detectors being trained on the Pro-

AI generative models		o	s	f	v
[73]	ProGAN	✓			
[13]	BigGAN	✓			
[28]	StarGAN			✓	
[108]	GauGAN		✓		
[76]	StyleGAN2	✓		✓	
[74]	StyleGAN3			✓	
[21]	EG3D			✓	
[151]	Diffusion-GAN	✓			
[126]	StyleGAN-T		✓		
[138]	GALIP		✓		
[72]	GigaGAN	✓			✓
[65]	DDPM			✓	
[40]	ADM	✓			
[133]	Score-SDE			✓	
[103]	GLIDE		✓		
[120]	Latent Diff.	✓	✓	✓	
[117]	DALL·E 2				✓
[122]	Stable Diff.				✓
[7]	eDiff-I		✓		
[110]	DiT	✓			
[14]	InstructPix2Pix				✓
[78]	DeepFloyd IF				✓
[111]	SDXL				✓
[105]	DALL·E 3				✓

Table 2.1. List of the AI generative models analyzed in these experiments. Each model generates images with different content, as specified in the following: generic objects (o) by training on ImageNet/LSUN, faces (f) by training on FFHQ/CelebA, scenes (s) by training on COCO, various (v) includes objects, faces and scenes not coming from the previous datasets.

GAN dataset introduced in [148], whereas Corvi2023 [31] is trained on images generated with Latent Diffusion. The best generalization overall is provided by Ojha2023[104], that, as described, relies on the features of the CLIP backbone. However all performance figures worsen in the presence of a significant training-test mismatch (Fig. 2.5, bottom) and some detectors exhibit an AUC very close to 50%, especially with DM images. Interestingly, the performance does not vary across different types of content as

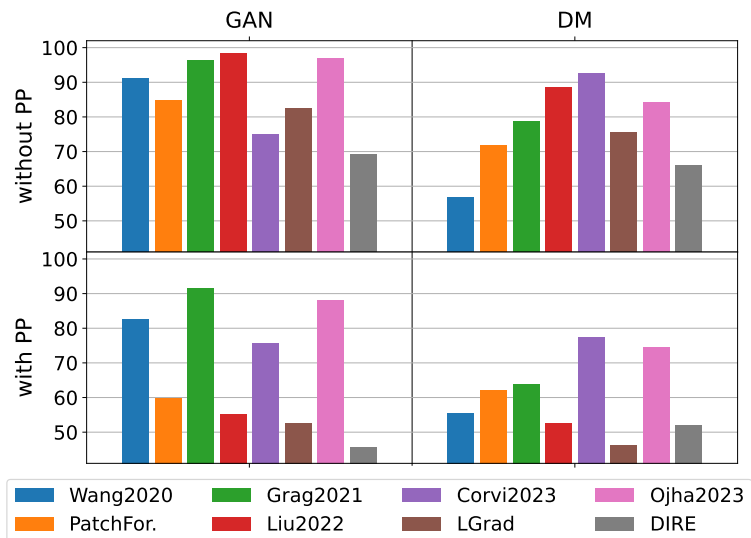


Figure 2.5. Synthetic image detection results in terms of AUC with and without post-processing (PP).

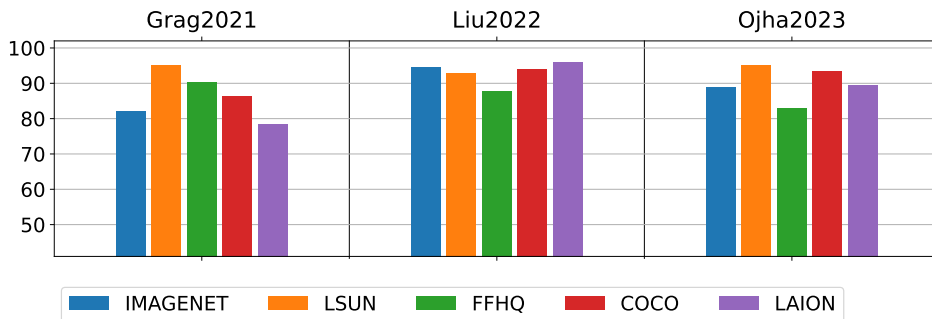


Figure 2.6. Synthetic image detection results in terms of AUC for three methods over various contents.

can be seen in Fig. 2.6. For example, both Ojha2023[104] and Liu2022[85] show very good performance on scenes or faces even though they are not trained for these categories.

In the wild

While the post processing experiments done could give us an idea of the

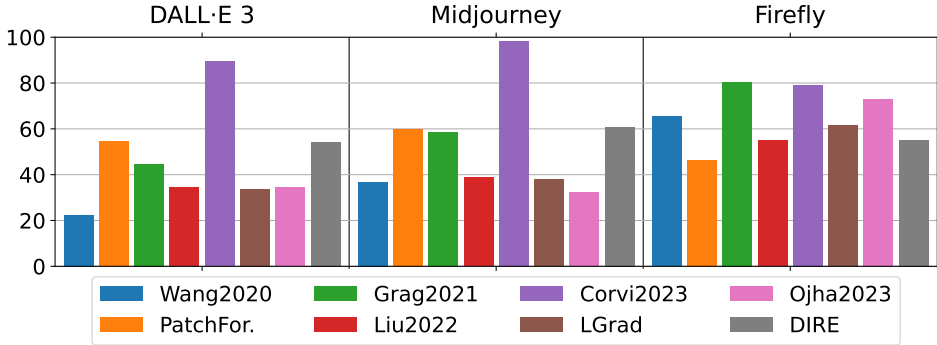


Figure 2.7. Synthetic image detection results in terms of AUC on images from X generated by DALL-E 3, Midjourney and Firefly.

performance, they are not a perfect representation of the operations images go through when uploaded only. As such, to evaluate the performance in a more challenging and realistic situation, we downloaded a total of 2,000 images from a well known social network (X), both real and generated from Midjourney v5, DALL-E 3 and Firefly (according to their associated hashtag). In 2.7 we show the results in terms of AUC. On these images, the methods perform quite differently from before. CNN-based architectures perform much better than the other solutions. In particular, Corvi2023[31] has a near perfect performance on both Midjourney v5 and DALL-E 3. This can be easily understood by looking at the figure Fig. 2.1, we can see that the forensic artifacts present in Midjourney 5 are very similar to those present in the Latent Diffusion images that the method has used in its training.

Calibration

So far, we presented all the results in terms of AUC. However, in a practical scenario we need to set a threshold to determine if the image under analysis is real or fake. This is far from a trivial problem: a fixed threshold is hardly appropriate in all situations, especially if training and test data are misaligned, and could lead to disappointing results. One example of this issue can be seen in Fig. 2.8, we can see that the distribution of the scores of the method Grag2021[60] trained on ProGAN. The distribution of ProGAN test images is very well separated from the real one when

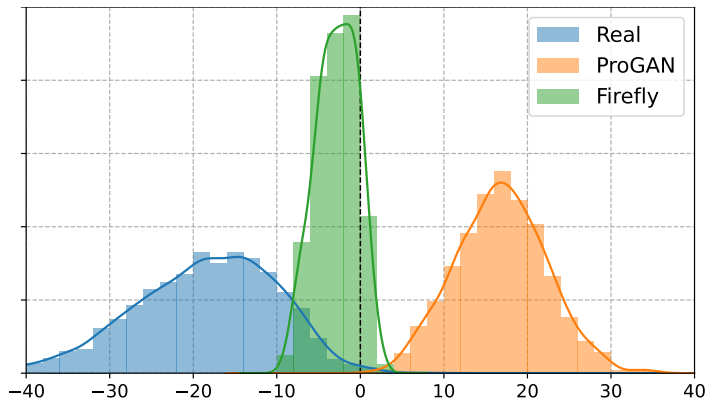


Figure 2.8. Scores distribution provided by the method [60] (trained on ProGAN) for real images, and for synthetic images from ProGAN and Firefly.

choosing a zero threshold. However, the same threshold would not work if we try to distinguish Firefly generated images, which were not included in training and would require a lower threshold to be used in order to detect them.

2.1.4 Synthetic Image Attribution

Image attribution goes beyond detection and focuses on inferring the provenance of the image by tracing it back to the generative model that created it. If the "real" class is among those considered in the process, then the attribution includes automatically the detection.

Artificial fingerprints

The field of synthetic image attribution in the early research focused primarily on adapting successful techniques and methods developed for conventional multimedia forensics to this new domain. Device and model fingerprints have proven to be of particular value for a wide range of forensic tasks and have been widely used especially for source identification. Chen et al. [25] where the first to introduce the concept of device fingerprint, they demonstrated that imperfections inherent in the camera sensors generated a unique and stable pattern in each captured image, a fingerprint by any means, called photo-response non-uniformity (PRNU) pattern. It

is possible, as said previously, to extract such a fingerprint from synthetic images, these represent a valuable tool for image attribution. To obtain it, the procedure is to first remove the semantic content of the scene through a denoiser. The resulting residual images are then averaged. As the number of averaged images grows, a weak but stable quasi-periodic pattern can be observed [94].

Such fingerprints allow different models to be discriminated but can also provide information on the different datasets used for training. In fact, the same architecture, trained in different conditions, gives rise to slightly different fingerprints that allow fine-grained model authentication. Overall, the observed fingerprint turns out to depend on the specific architecture, the training dataset and also the training hyperparameters (e.g. learning rate, optimizer, training iterations) [158]. A learned fingerprint is also extracted in [41] through a hierarchical Bayesian approach. It does not depend on the initial seed, but is shown to change based on the training dataset and is robust to benign image transformations, such as compression, blurring and additive noise. A slightly different perspective is taken in [4], where a fingerprint estimation network is proposed to capture the unique patterns used to predict the network architecture and loss function that characterize a generative model. It is worth underlining that, in any case, these fingerprints are due to subtle traces present in the image, and therefore lay in the medium/high frequency bands. Therefore, while exhibiting a certain robustness, they can be vulnerable to adversarial attacks or can even be replaced by alien fingerprints extracted from real cameras [32].

Attribution via inversion

Attribution can also be pursued by image inversion, since a generative model is not able to perfectly synthesize an image coming from another model nor it can perfectly reproduce a real image [75]. The idea is to know a set of possible generators, both architectures and weights, and then compute the input data of the generator that allows to produce an image as similar as possible to the image under test. The likely source is the generator that ensures the minimum reconstruction error [2]. A robustness analysis and the extension to an open set scenario can be respectively found in [161] and [64].

Attribution as a classification problem

The source attribution problem can be simply considered as an N -ary classification, where the N classes can be associated for example with different architectures. In [16], in order to learn an attribution model that is robust to different categories and possible perturbations, it has proposed a mix-up representation training strategy at the feature level. Interestingly, the approach can effectively handle both detection and attribution through a compound loss that takes into account the hierarchical nature of the problem. The image can be classified as real or fake and only in the latter case should the problem of attribution be considered. Instead in [153] a patchwise contrastive learning strategy is pursued with pre-training on image transformation classification that have been found to be similar to the generator components.

Open-set methods

When dealing with synthetic image attribution, several methods assume to work in a closed-set scenario, where test images were generated by a limited set of architectures, known in advance, whose samples were present in the training set. However, this scenario is far from realistic, given that new generation methods are proposed every day. In fact, images of new architectures may not even be available. In this condition, it is very important to be able to identify an out-of-distribution sample, recognizing that it does not belong to any of the classes observed in the training process. This more realistic scenario is known as open set recognition. A first open-set approach was proposed in [58] with an iterative algorithm that discovers new classes and re-trains the network using them as pseudo-labels. This method uses a fixed set of labeled images and then performs attribution on a set of unlabeled data through a clustering procedure. More recent works rely on classification with a rejection option, as in [147] based on a hybrid ViT+ResNet50 architecture, or in [51] where a metric-learning based embedding is developed to measure the similarity between the source generators of synthetic images. Finally, in [154] a progressive open space solution is proposed. The idea is to simulate unknown classes through a set of augmentation models, based on lightweight networks that can model the traces of a variety of unknown models.

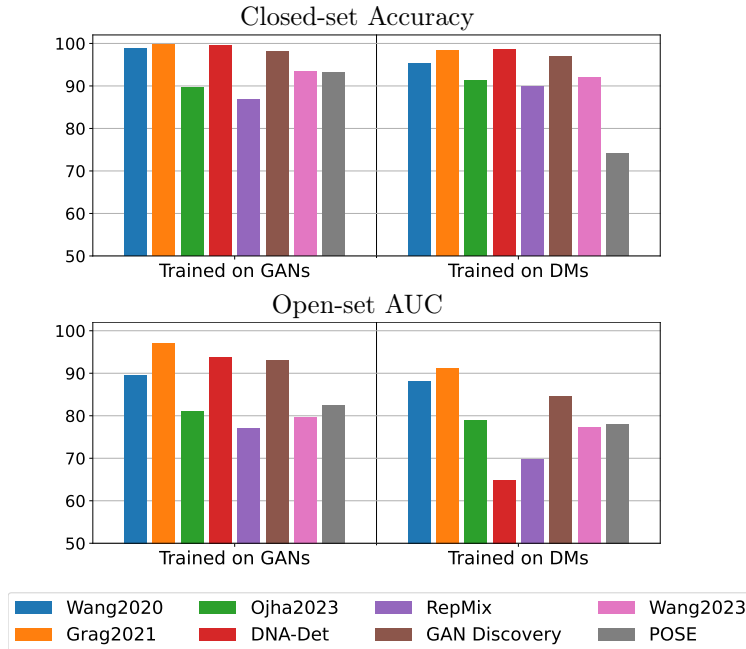


Figure 2.9. Synthetic image attribution results in both closed-set and open-set scenarios, considering two separate training sets (GANs and DMs).

2.1.4.1 Closed-set vs Open-set analysis

In this Section we carry out a comparison of some of the methods proposed in the literature for which code was publicly available on-line: RepMix [16], DNA-Det [153], GAN Discovery [58], Wang2023 [147], POSE [154]. All methods were re-trained using exactly the same data. Along with the methods specifically proposed for attribution, we also consider some of the best detection methods analyzed in the previous section re-trained in a multi-class configuration [148, 60, 104]. We consider both closed-set and open-set scenarios and carry out attribution at architecture-level. We evaluate the performance of closed-set attribution in terms of accuracy, a widely utilized metric for multi-classification problems. For the open-set scenario, we measure the capacity to distinguish between in-distribution and out-of-distribution samples in terms of AUC, as is typically done in

the literature [147, 51, 154]. More specifically, We analyze different situations as described below.

GAN-based vs DM-based

To analyze the possible different performances among GANs and DMs, we consider two separate sets of 7 synthetic generators, one is GAN-based (ProGAN, BigGAN, StarGAN, GauGAN, StyleGAN2, StyleGAN3, Diffusion-GAN) the other is DM-based (ADM, DDPM, Score-SDE, GLIDE, Latent Diffusion, Stable Diffusion, DALL·E 2). These two sets comprise 2,500 images per architecture, for a total of 17,500 images each. Then we carry out an experiment in a closed set scenario (Fig. 2.9, top) and in an open set scenario (Fig. 2.9, bottom) where we limit the new unknown classes to come from 3 GAN-based generators (EG3D, GALIP, GigaGAN) and 3 DM-based generators for (DiT, DeepFloyd-IF, SDXL), respectively. In both scenarios, we consider 500 images per generator under test. We

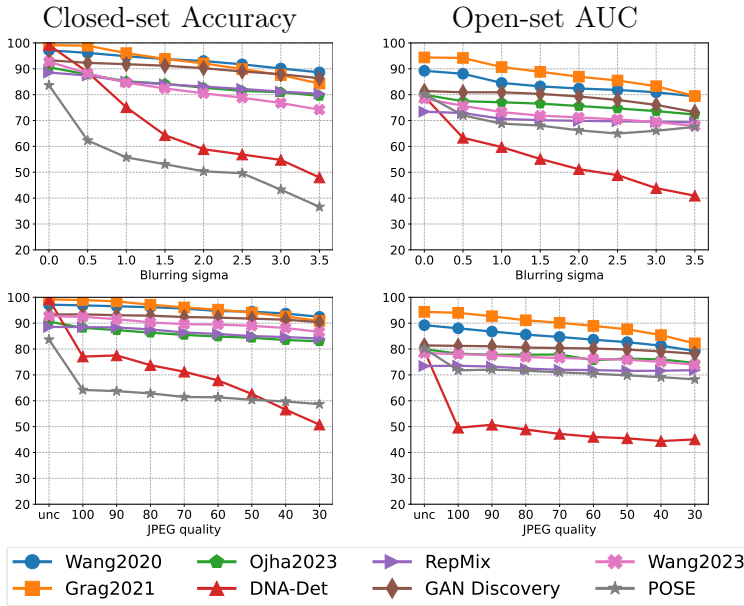


Figure 2.10. Robustness analysis in both closed-set and open-set scenarios, by varying Gaussian blurring (top) and JPEG compression quality (bottom).

can observe that in the closed-set scenario the performance is very good both for GANs and DMs. It worsens for open-set recognition, especially for DMs, where AUC can hardly reach 90%. Interestingly, detection-based methods perform very well in this new setting, coherently with recent findings that the ability of a classifier to make a reject-option decision is very much correlated with its accuracy on the closed-set classes[144].

Robustness analysis

In the same scenario presented above, we also evaluate robustness to JPEG compression and blurring in order to check the sensitivity of the methods to possible benign perturbations. Note that in this case, suitable augmentation is included in all the approaches as also suggested in the original papers. Results are presented in Fig. 2.10 and show that for most of the analyzed methods the decrease in performance is quite limited.

Detection and attribution

In this last experiment we consider a more challenging scenario where, beyond mixing GANs and DMs architectures, also real images are present. Note that in the literature, real images are separated based on their dataset of provenance. However, this may introduce biases that in turn help achieving a good performance and, eventually, lead to unfair comparisons. To

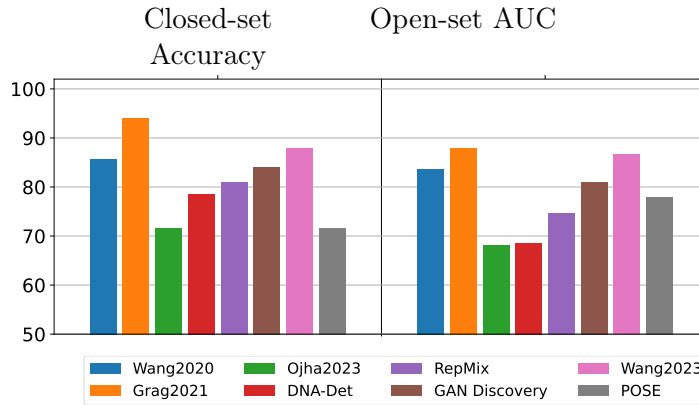


Figure 2.11. Synthetic image attribution results in the more challenging scenario where also the real class is included.

be completely fair, we consider a single class that includes all real images, no matter what the original dataset. In this situation, a very similar behavior is observed in the closed-set and open-set scenarios (Fig. 2.11). All methods suffer only a small performance loss in the open-set scenario. Instead, a substantial loss is observed with respect to the situation where only synthetic images are considered (Fig. 2.9), especially in the closed-set scenario. It is also worth noting that the probability of correctly classifying the real class and the fake ones is balanced, with 94.8% and 93.9% respectively, for the best approach.

2.2 Detection of AI-generated videos

Synthetic video generation has advanced rapidly in recent years, driven largely by powerful tools like diffusion-based models. Early methods like Text2Video-Zero [77], Modelscope [146], and Hotshot-XL [102] adapted image generators for video synthesis, but struggled with challenges such as temporal coherence and motion consistency. Models—including Mochi-1 [139], Allegro [168], Opensora-Plan [82], and CogVideoX [155]—introduce dedicated architectures leveraging 3D causal autoencoders and 3D transformers, enabling better modeling of spatio-temporal relationships. These models also achieve stronger semantic alignment with input text and produce videos with higher quality and improved temporal consistency [66]. Recently, autoregressive models, such as Loong [149] and NOVA [38], have shown impressive performance for video generation. Although they are based on a completely different generation paradigm, they still exhibit anomalies similar to diffusion models that can be exploited by the forensic classifier. We will dive deeper into this anomalies in Chapter 5.

Initial approaches for detecting fully generated videos focused on human motion cues [12], while others adapted image-based forensic detectors using few-shot learning [143]. More recent methods exploit both spatial and temporal artifacts. Bai et al. [5] propose a two-branch CNN to capture spatial anomalies and optical flow inconsistencies, while Chang et al. [23] use three 3D CNNs that target appearance, motion, and geometry. Liu et al. [87] take a different path, feeding RGB frames and reconstruction errors into a CNN+LSTM model, based on the idea that diffusion models can better reconstruct frames synthetic of synthetic videos than from real

ones, a principle previously applied to images [150]. Other methods leverage vision-language foundation models like CLIP [91], X-CLIP [24], and LLaVa [132] to capture both spatial and temporal inconsistencies. A more recent approach [80] also introduces a loss function to encourage attention over diverse spatial regions, improving detection beyond facial areas.

Chapter 3

Artifact analysis of synthetic content

*Thoughts without content are empty,
intuitions without concepts are blind.*

Immanuel Kant

Detecting fake images has become a major goal of computer vision. This need is becoming more and more pressing with the continuous improvement of synthesis methods based on Generative Adversarial Networks (GAN), and even more with the appearance of powerful methods based on Diffusion Models (DM).

While this technology offers tremendous value to artists, game designers, and creative individuals, it also comes with risks. The potential security and ethical issues that can arise from the misuse of these images cannot be ignored. Research on synthetic image detection has received a huge boost in recent years [93, 97], especially since the first realistic images generated by Generative Adversarial Networks [170, 73] appeared. Some works focus mainly on the inability of such generators to perfectly reproduce the high-level semantic features of natural images.

However, the generation methods have been improving through time, showing fewer and fewer synthesis artifacts. Therefore, other approaches focus on low-level features that are related to the generative model architecture and can be exposed by removing the high-level semantic content

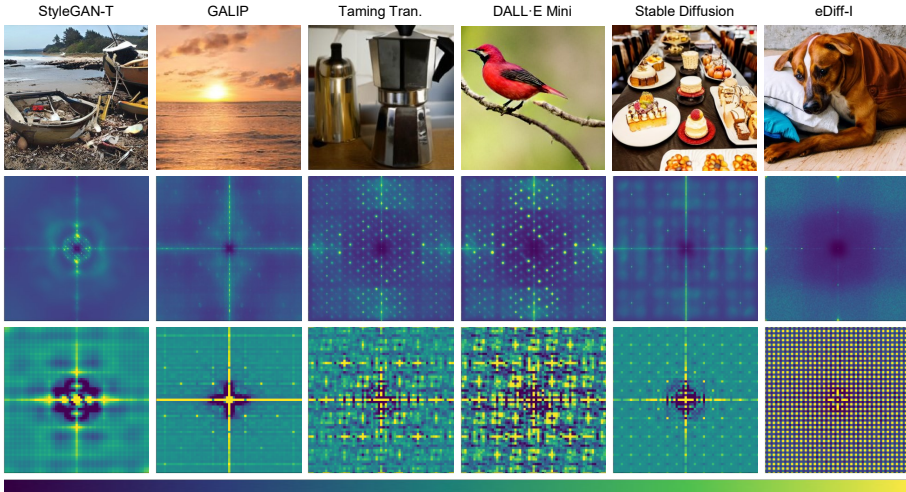


Figure 3.1. Examples of synthetic images (top) generated by state-of-the-art models: StyleGAN-T [126], GALIP [138], Taming Transformers [50], DALL-E Mini [37], Stable Diffusion [121], eDiff-I [7]. Each model leaves peculiar traces in the generated images, which can be regarded as a sort of artificial fingerprint and used for forensic analyses. They are typically visible in the frequency domain as spectral peaks in the power spectra (middle), or in the spatial domain as anomalous regular patterns in the autocorrelation (bottom). It is easy to observe that models based on the same architecture, like Taming Transformers and DALL-E Mini, give rise to similar artifacts.

of the captured scene [94, 158]. In fact, the complex processing pipeline needed to generate synthetic images inevitably leads to the introduction of digital artifacts that characterize the specific architecture and significantly differ from those typical of traditional acquisition devices. Some works have already analyzed the traces present in the synthetic images working in the Fourier domain. The analysis focused mostly on artifacts caused by GAN models, such as those due to the up-sampling steps used to achieve the desired spatial resolution [165, 45]. However to date, little attention has been given to more recent and promising architectures, such as those based on auto-encoders and diffusion models [31, 130, 118, 104].

We carry out a systematic investigation of the traces left not only by established GAN-based generators but also by several recent VQ-based and DM-based generators. To this end, we consider second-order statistics

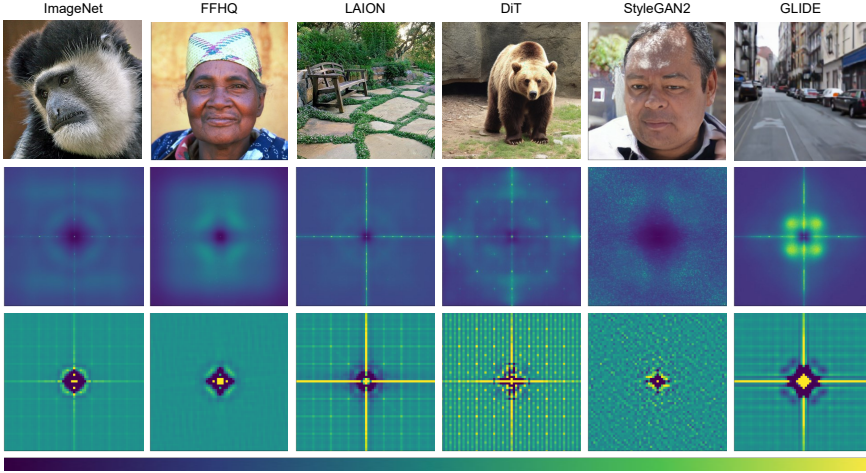


Figure 3.2. Examples of real and synthetic images drawn from popular datasets. Real datasets (left): ImageNet [39], FFHQ [75], LAION [128]. Synthetic datasets (right): DiT [110], StyleGAN2 [76], GLIDE [103]. For each dataset, we also show the average power spectra (middle) and autocorrelation (bottom) of noise residuals. It is possible to observe that real images embed some traces of JPEG artifacts while all generated images exhibit anomalous periodic patterns in the spatial domain.

both in the spatial domain and in the frequency domain. For the spatial domain, we use the average autocorrelation function while its transform, the average power spectrum, is used for frequency-domain analyses together with some derived quantities: the radial and angular spectra. Examples of power spectra and (central crop) autocorrelations are shown in Fig. 3.1 and Fig. 3.2. In Table 3.1 we can find all the synthetic image generators analyzed in this chapter. As a result of our analyses we contribute the following main observations:

1. not only the GAN models but also the DM and VQ-GAN (Vector Quantized Generative Adversarial Networks) models give rise to strong peaks in the Fourier domain and exhibit anomalous regular patterns in the autocorrelation;
2. when the dataset used to train the model lacks sufficient variety, its biases can be transferred to the generated images;

	Method	Backbone	Dataset	Conditioning
GAN	[13] BigGAN	ResNet-style	ImageNet	Class
	[76] StyleGAN2	StyleGAN like	FFHQ	Uncond.
	[74] StyleGAN3	StyleGAN like	FFHQ	Uncond.
	[126] StyleGAN-T	StyleGAN like	CC12M, CC, YFCC100m, Redcaps, LAION-aesthetic-6+	Text
	[138] GALIP	Transf. based	CUB, MS-COCO, CC3M, CC12M	Text
VQ	[50] Taming Transf.	VQGAN + Transf.	ImageNet, FFHQ	Class, Uncond.
	[37] DALL-E Mini	VQGAN + Transf.	CC, YFCC100M, CC12M	Text
DM	[133] Score-SDE	U-Net	FFHQ	Uncond.
	[40] ADM	U-Net	ImageNet, LSUN	Class, Uncond.
	[103] GLIDE	U-Net	CC, YFCC100M, Wikipedia text-image pairs	Text
	[117] DALL-E 2	U-Net	Unknown	Text
	[7] Ediff-I	Mul. U-Nets	Unknown	Text
LDM	[120] Latent Diff.	VAE + U-Net	ImageNet, LSUN	Class, Uncond.
	[121] Stable Diff.	VAE + U-Net	LAION	Text
	[110] DiT	VAE + ViT	ImageNet	Class

Table 3.1. Main properties of the generative models analyzed.

- synthetic and real images exhibit significant discrepancies in the mid-high frequency signal content, observable in their radial and angular spectral power distributions.

3.1 Forensic Analysis

In order to develop increasingly effective detectors, a crucial step is understanding which image features can help distinguish fake images from real ones. Interestingly, a few carefully designed simple experiments provide us already with several insightful observations. We are going to go through some of these experimental results, with the aim of conveying the importance of these findings. Our analysis is carried out by relying on second-order statistics observed both in the spatial domain, the image autocorrelation function, and in the frequency domain, the image power spectrum¹.

For all sources of interest we consider a set of $I = 1000$ images, either real or generated under controlled conditions. For a single image $x_i(m, n)$ the autocorrelation function is

$$R_{x_i}(\Delta m, \Delta n) = \langle x_i(m, n) x_i(m + \Delta m, n + \Delta n) \rangle \quad (3.1)$$

where $\langle \cdot \rangle$ indicates spatial average. The average autocorrelation of the

¹In the literature, “power” is sometimes replaced by “energy”, but it is only a matter of theoretical image modeling.

source should be obtained through statistical averaging, which we approximate by averaging on all I images

$$R_x(\Delta m, \Delta n) = \frac{1}{I} \sum_{i=1}^I R_{x_i}(\Delta m, \Delta n) \quad (3.2)$$

This function summarizes the statistical correlation between pixels separated by a $(\Delta m, \Delta n)$ displacement.

For frequency domain analyses, we start from the Fourier transform of the $M \times N$ image

$$\begin{aligned} X_i(k, l) &= \mathcal{F}[x_i(m, n)] = \\ &= \sum_{m=1}^M \sum_{n=1}^N x_i(m, n) e^{-j2\pi(\frac{k}{M}m + \frac{l}{N}n)} \end{aligned} \quad (3.3)$$

and obtain the source power spectrum, again, by averaging all individual power spectra

$$S_x(k, l) = \frac{1}{I} \sum_{i=1}^I S_{x_i}(k, l) = \frac{1}{I} \sum_{i=1}^I |X_i(k, l)|^2 \quad (3.4)$$

The power spectrum accounts for the fraction of the total image power concentrated at a given (horizontal, vertical) frequency pair $(\frac{k}{M}, \frac{l}{N})$. Note that these two functions are themselves related by a frequency transform

$$S_x(k, l) = \mathcal{F}[R_x(\Delta m, \Delta n)] \quad (3.5)$$

hence redundant, in theory. However, we study both of them since some phenomena are more visible in the spatial domain while other ones in the frequency domain. Nonetheless, the effects of interest are only visible in the noise residuals, obtained by removing the high-level semantic content from the image by means of a denoising filter.

$$r_i(m, n) = x_i(m, n) - \mathcal{D}(x_i(m, n); \sigma) \quad (3.6)$$

In our experiments, we use always the denoiser by Zhang *et al.* [163] with noise parameter $\sigma=1$.

All autocorrelations and spectra, associated to images or residuals, are normalized to the average power per pixel computed as the mean value of $S_x(k, l)$. Note that we show only the central 65×65 crop of the autocorrelation function.

3.1.1 Architecture related traces

Previous research has shown that images produced by GAN-based generators contain noticeable artifacts left by the processing pipeline. These can be exposed by extracting and averaging the noise residuals, to compute a model artificial fingerprint, and analyzing the results in the spatial and/or spectral domain. Images generated by diffusion models appear to show artifacts of a similar nature and, arguably, a similar origin. Fig. 3.3 shows the power spectra (top) and autocorrelation functions (bottom) of

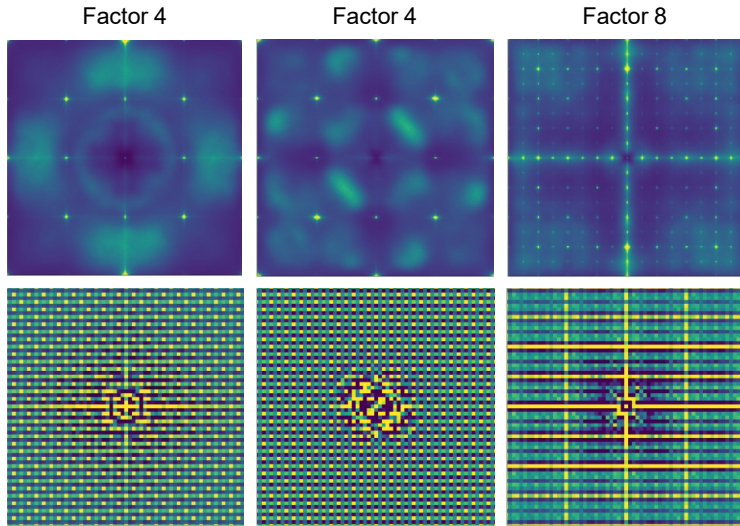


Figure 3.3. Power spectra (top) and autocorrelation functions (bottom) of noise residuals of three slightly different latent diffusion models. The upsampling factor between latent space and pixel space is 4 for the first two architectures and 8 for the third one. This single parameter appears to be responsible for the positioning of the peaks in the power spectra and the periodicity of the regular patterns in the autocorrelations.

the artificial fingerprints of three slightly different latent diffusion architectures. The first two models are characterized by an upsampling factor of 4 between latent space and pixel space and, not by chance, exhibit clear peaks in the spectrum at frequencies multiples of $1/4$, and strong regular patterns with periodicity 4 in the autocorrelation function. The minor differences between the two are mostly due to the different training set used, ImageNet vs. CelebHQ. On the contrary, the third model, trained on LSUN churches images, has an upsampling factor of 8. As expected, peaks in the power spectrum are now denser, at frequencies multiple of $1/8$, while they are sparser in the autocorrelation function. While these artifacts are clearly present in current generation architectures, and can be exploited for forensic needs, these analyses refer to images observed “in the lab”, as they leave the generator, with no further processing. However, this is not a realistic condition, it is enough to think of all the standard processing steps, especially compression and resizing, applied to images as soon as they are uploaded on a social network. While it is not their objective, these post-processing steps may significantly modify and hide the artifacts highlighted before and should be always taken into account when designing a forensic detector. Fig. 3.4 shows some interesting examples of the effects of various post processing operations. We consider images generated by Stable Diffusion and show, on the top-left corner, the autocorrelation function observed in the absence of any post-processing. A clear regular pattern is visible which suggests the artificial nature of these images and may even be used to trace back the individual generator used. In the other slots, we show the autocorrelation function after various types of post-processing, blurring, sharpening, resizing at three different scales, JPEG compression at three different quality factors. The autocorrelation function changes more or less markedly depending on the nature and intensity of the post-processing, but in most cases it can be hardly associated with the original one. In the presence of resize, we can see that the artifacts shifts quite significantly, while with strong JPEG compression, for example, the compression artifacts dominate the scene and hide completely the generation artifacts.

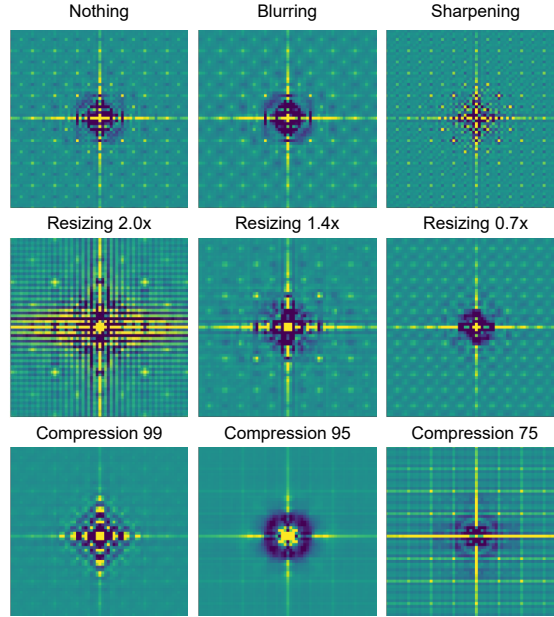


Figure 3.4. Effect of post-processing operations on the autocorrelation of noise residuals. We consider Stable Diffusion images and apply to them different operations to see how the autocorrelation changes. In some cases the original generation traces (top-left) are completely hidden by post-processing artifacts.

3.1.2 Training dataset bias

Moreover, what we have shown so far is not the only issue. During our analysis we highlighted one further problem that compromises the reliability of the high-frequency fingerprints. This is exemplified by the images of Fig. 3.5, where, in the bottom row, the noise residuals of the Diffusion Models exhibit an obvious regular pattern that one may believe is caused by their artificial origin. However, this is the very same pattern observed in the residuals of real images JPEG compressed at quality factor $Q=75$ (see Fig. 3.4). When analyzing the training pipeline of the two generation models presented, we noticed that both used as dataset entirely comprised of images that were JPEG compressed at $Q=75$. The model learned to include the compression artifacts in the generated images, that

is, it transferred dataset-related artifacts in the generated image, and has the potential to mislead fingerprint-based detectors. This example rings also a bell about the experimental protocol. In fact, most of the times we do not know the details of the dataset used to train a model, but such details may well impact the characteristics of the generated images. A fully reliable analysis must keep this aspect under control. Interestingly, this ability is also present in GANs models that are able to transfer the same fingerprints present in the training data to the generated images. [159] proposed to use this capability to protect synthetic images by having them produce the artificial fingerprint embedded in the training data.

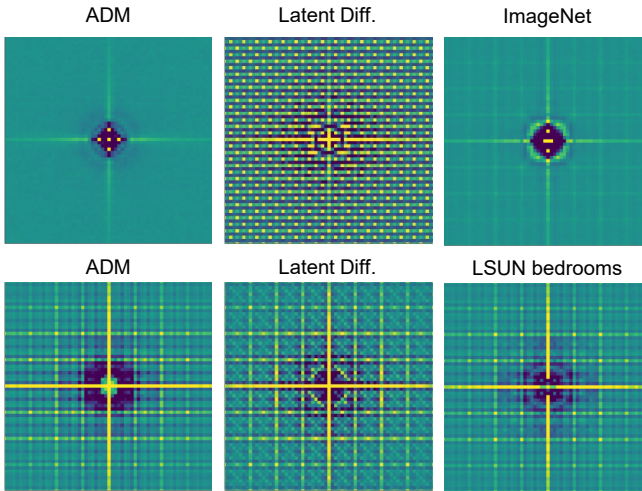


Figure 3.5. Autocorrelation of noise residuals for real and synthetic images. Top: two diffusion models (ADM and Latent) and the dataset used for their training (ImageNet). Bottom: two diffusion models (ADM and Latent) and the dataset used for their training (LSUN Bedroom). In the first case the JPEG traces (8×8 grid) are barely visible in the training dataset and they are not present in the generated images. Instead, in the second case they are clearly visible both in the training dataset and in the generated images.

3.1.3 Fidelity to natural images

Natural images are known to exhibit strong statistical regularities in the frequency domain. In particular, several studies [17, 140] observed that the average power spectrum decays with a $1/f^\alpha$ law, with $\alpha \simeq 2$. Moreover, it has been also recognized [134, 142] that the power spectrum is far from isotropic, with power concentrated especially at the horizontal and vertical orientations and much less at the others. These may be precious pieces of information towards the detection of synthetic images. Indeed, considering their good photorealism and high semantic quality, we expect their spectra to be very similar to those of real images, at low frequencies. However, it is certainly possible that significant deviations arise at mid-high frequencies, not because of generation artifacts but due to the inability to correctly replicate the statistics of natural images. This conjecture is indirectly supported also by the literature where there is wide agreement on the superior performance of dual-stream detectors, which exploit both high-frequency artifacts and original RGB images. With this approach, the detector has the opportunity to analyze not only the highest frequencies, with their subtle generation artifacts, but also the intermediate frequency, which carry information on the signal itself, with beneficial effects.

Analysis of radial spectra.

To test this conjecture, we now analyze the radial spectra of real and synthetic images. The radial spectrum of the i -th image is computed by averaging the values of its Fourier transform (normalized by its standard

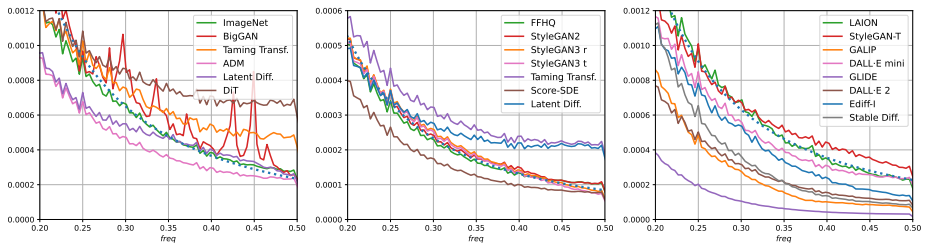


Figure 3.6. Radial spectrum power density. Synthetic images are compared with the real images used for training the correspondent model. Real images (green) fit very well the expected theoretical curve (dotted).

deviation) at distance ρ from the origin, irrespective of the angular direction

$$X_i(\rho) = \langle X_i(k, l) \rangle_{A(\rho)} \quad (3.7)$$

where $A(\rho)$ is the annulus between radii $\rho - \Delta\rho$ and $\rho + \Delta\rho$, and ρ goes from 0 to 0.5 in 128 discrete steps. Then, as usual, an average on $I = 1000$ images is taken to characterize the whole source. In our analyses we will discard low and very high frequencies and focus on the range $\rho \in [0.2 - 0.5]$.

Our goal here is to test whether the radial spectra of synthetic images follow the same “correct” law as those of real images. However, to avoid dataset-related biases, for a given real source (*e.g.*, imagenet), we study only the images that are generated using the very same source as training set. With this care, statistical difference can be attributed only to the generation engine. In Fig. 3.6, for a few widespread real sources, we compare radial spectra of real and corresponding generated images. The former (green) decay with a nearly perfect power law (dashed black). On the contrary, all radial spectra of synthetic sources deviate more or less markedly and in one or more frequency ranges from that law. Generally, but not always, real images seem to possess a richer content at the intermediate frequencies. Only for the FFHQ dataset, the radial spectra of some synthetic sources approximate pretty well that of the real source. However, the FFHQ dataset is composed by images that have been interpolated, that is low-pass filtered, a process that inevitably modifies the original spectral content and makes it indistinguishable from that of some synthetic images. In hindsight, this further supports our conjecture.

Analysis of angular spectra.

A similar analysis is carried out for the angular spectra, where now

$$X_i(\theta) = \langle X_i^{hp}(k, l) \rangle_{C(\theta)} \quad (3.8)$$

$C(\theta)$ is a double cone between angles $\theta - \Delta\theta$ and $\theta + \Delta\theta$, and θ goes from 0 to π in 16 discrete steps. Note that the non-informative low-pass content is filtered beforehand with cut-off frequency 0.1. The resulting plots are reported in Fig. 3.7(top) for the same real and synthetic sources as before. Mostly, the very same considerations already done for radial spectra apply. The only exceptions are observed again for the FFHQ dataset and for

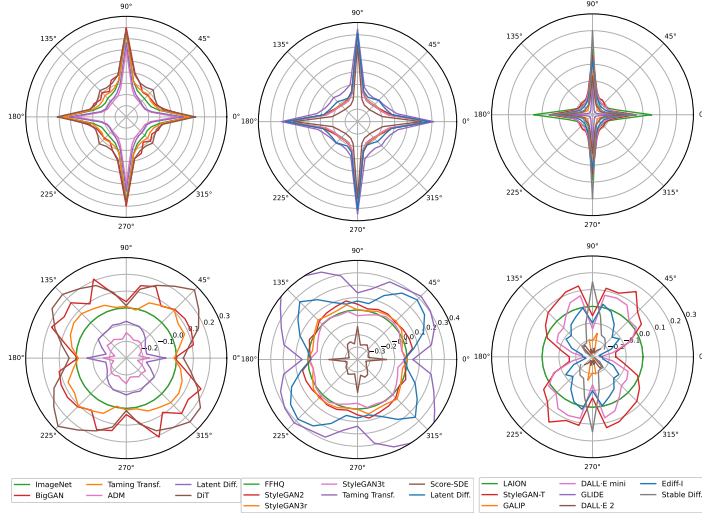


Figure 3.7. Angular spectrum power density (top) and the square-root of Fisher's discriminant ratio applied at each direction (bottom).

the same reasons as before. In this case, there is no theoretical reference pattern from the literature. We observe again significant deviations from the angular spectra of the real sources (green). Such deviations can be better appreciated in the plots of Fig. 3.7(bottom), showing the square root of the Fisher's discriminant ratios

$$F_s(\theta) = \frac{\mu_s(\theta) - \mu_0(\theta)}{\sqrt{\sigma_s^2(\theta) + \sigma_0^2(\theta)}} \quad (3.9)$$

where $\mu_s(\theta) = X_s(\theta)$ and $\mu_0(\theta) = X_0(\theta)$ are the mean angular spectra of synthetic (s) and real (0) images, respectively, and $\sigma_s^2(\theta), \sigma_0^2(\theta)$ the corresponding variances. In this case the real source is represented by the green unit circle. Finally, it is worth noting that the StyleGAN2 and StyleGAN3 models better fit the spectral distribution of real images. For StyleGAN3, this is justified by the development of GAN architectures that were able to avoid aliasing[74].

3.2 Conclusions

Generators of synthetic images are becoming more and more sophisticated. Casual viewers are already unable to tell real images from fake ones and even forensic tools struggle to cope with the rapid progress of new models. In this context, brute-force detectors, even trained on huge amounts of data, may not be up to the task, especially when asked to generalize to new generators. Therefore, it is important to single out and possibly exploit all features that distinguish real from fake data. We analyzed a large number of modern and popular generators looking for such features both in the spatial domain and in the frequency domain. In both domains we relied on second-order statistics, autocorrelation functions and power spectra. This investigation provides some interesting results. On one hand, even the most sophisticated architectures keep generating artifacts that can be exploited for detection. These may be more visible in the spatial or in the spectral domain and both should be always explored. In addition, generators do not seem able to mimic the spectral distribution of real images at the mid-high frequencies and radial and angular spectra may help expose such discrepancies. Finally, we noted that a strongly biased training set may teach the model to transfer the very same biases in the generated images, a possible problem for both generators and detectors.

Chapter 4

Leveraging Vision Language Models for the detection of synthetic images

For it is not enough to have a good mind; the main thing is to use it well.

Rene Descartes

One of the main challenges for current forensic detectors is the ability to generalize to new and unseen generative methods. Indeed, the in-distribution (ID) scenario, with perfectly aligned training and test sets, is rarely met in practice. Test images are often generated by new approaches, unseen architectures, or even known architectures re-trained under different conditions, all representing different flavors of the out-of-distribution (OOD) scenario. Additionally, most images are downloaded from social networks, where they undergo post-processing operations such as compression and resizing (sometimes multiple times), processes that tend to wash out the tiny traces so precious for detection. In [148], it is shown that the diversity of the training dataset and intense augmentation are crucial for generalization. In their proposal, a ResNet-50 detector is trained on a single but highly diverse dataset of about 360k ProGAN [73] images. The results obtained by their method on OOD images produced using other

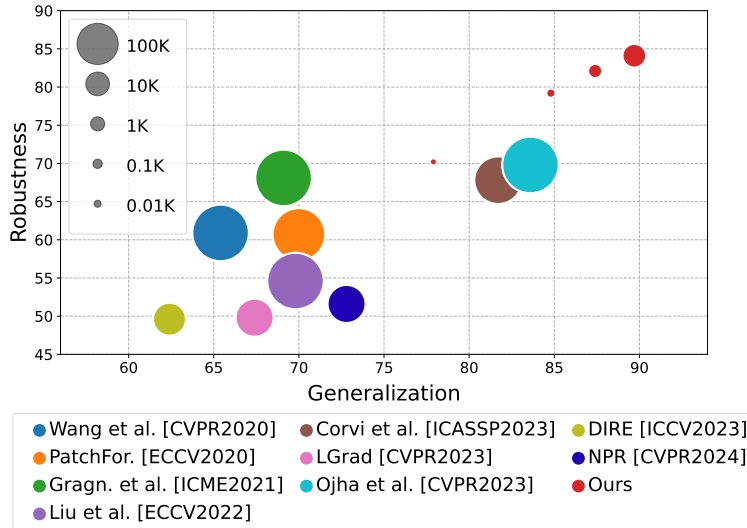


Figure 4.1. Area Under ROC Curve (AUC %) on unseen synthetic generators (x -axis) and on post-processed data (y -axis). The first number measures the generalization ability of the detector, the second measures its robustness to possible impairments. Circle area is proportional to training set size. Performance is measured over 18 different synthetic models. Our CLIP-based detector largely outperforms all SoTA methods with very limited training data.

GAN-based generators turned out to be surprisingly good. However, the authors themselves attribute this performance to the structural similarities between GAN-based generators. Indeed, results are not equally good on images generated by recent diffusion-based methods, which present somewhat different generation artifacts [31].

Generalization to OOD data is a major issue in deep learning and has been the subject of intense research in recent years. In this context, the advent of large pre-trained vision-language models has brought about a number of new solutions and exciting results. These models have been shown to be excellent zero-shot and few-shot learners in many diverse applications, such as image classification [115, 164], detection [48, 99] and segmentation [152, 169]. Recently, there have been attempts to exploit the power of Vision Language Models (VLMs) to detect synthetic images [130, 3, 104]. For example, [104] considers the same dataset, augmenta-

tion strategy and experimental protocol as in [148] but uses a pre-trained VLM, the Contrastive Language-Image Pre-Training (CLIP) [115], as a feature extractor rather than training a ResNet-50. The only part which is trained on the task-specific dataset is the classifier. Compared to [148], the performance improves significantly, especially on images generated by diffusion models, showing excellent generalization capability.

We explore in depth the potential of CLIP for image forensics and conduct an extensive experimental analysis in challenging real-world scenarios involving a large number of generative models. We find that CLIP-based methods show impressive generalization ability and very good robustness, resulting in large performance improvements compared to the state of the art. In order to achieve such results, we avoid any intensive training on domain-specific data, which could introduce unwanted biases and undermine the descriptive power of CLIP features. Instead, we rely on a small number of paired real/fake images with the same textual description and use their CLIP features to model the decision space. Synthetic images come from a single generator. However, the results are equally good across all different data sources.

Noticeably, the top performance is achieved with just 1,000 to 10,000



Figure 4.2. Examples of synthetic images from generators used in our experiments. From left to right, Top: GLIDE [103], Latent Diffusion [117], DALL-E 2 [117], Stable Diffusion 1.4. Bottom: Stable Diffusion 2.1 [122], Stable Diffusion XL [111], Adobe Firefly [54], DALL-E 3 [105].

paired images. Moreover, only minor performance decays are observed when this number reduces to 100 or even 10 (see Fig. 4.1). In summary, the key contributions of this chapter are as follows:

- We show that CLIP features achieve excellent generalization: by exploiting only a handful of examples, not even belonging to the generator under test, the performances are comparable to those of intensively trained solutions.
- We carry out a large set of experiments on diverse synthetic generators and in very challenging conditions, achieving the best performance on average.
- Our experiments make clear that the features extracted from CLIP are partially orthogonal to the low-level features used by previous methods.

4.1 Datasets and metrics

The main objective is to study the generalization capacity of conventional and CLIP-based methods, i.e. to evaluate how detectors behave on data never seen before, of different origins and possibly impaired by various forms of post-processing. Therefore, we consider a very large test dataset and take measures to mitigate any potential biases in the experimental validation process.

Tab. 4.1 lists all generation models used during test. They are grouped into three families based on the type of generator: GAN-based (ProGAN, StyleGAN2, StyleGAN3, StyleGAN-T, GigaGAN, BigGAN, EG3D, Diff.ProjectedImageNet, GALIP), DM-based (Score-SDE, ADM, GLIDE, eDiff-I, Latent and Stable Diffusion, DiT, DDPM, DeepFloyd-IF, Deepfloyd-IF II stage, Stable Diffusion XL) and commercial tools (DALL-E 2, DALL-E 3, Midjourney v5 and Adobe Firefly), which include some recent text-based methods available online, whose architecture is not always disclosed. The second column of the table specifies the generation modalities employed by these models: unconditional (u), conditional (c), and text-to-image (t). The last column reports the resolution of the images in the dataset. The central columns indicate which datasets of real images were considered to

Generator	modality	LSUN [156]	FFHQ [75]	ImageNet [39]	COCO [84]	LAION [128]	RAISE [35]	Resolution
ProGAN [73]	u	✓						256 ²
StyleGAN2 [76]	u	✓						256 ² -1024 ²
StyleGAN3 [74]	u		✓					1024 ²
StyleGAN-T [126]	t				✓			512 ²
GigaGAN [72]	c,t			✓		✓		256 ² , 512 ²
BigGAN [13]	c			✓				256 ² , 512 ²
EG3D [21]	u		✓					512 ²
Diff.ProjectedGAN [151]	u	✓						256 ²
GALIP [138]	t				✓			224 ²
Score-SDE [133]	u		✓					256 ²
ADM [40]	u,c	✓		✓				256 ²
GLIDE [103]	t				✓		✓	256 ²
eDiff-I [7]	t				✓			256 ² , 1024 ²
Latent Diff. [120]	u,c,t	✓	✓	✓	✓			256 ²
Stable Diff. [122]	t					✓	✓	256 ² -768 ²
DiT [110]	c			✓				256 ² , 512 ²
DDPM [65]	u		✓					256 ²
Deepfloyd-IF II stage [78]	t				✓			256 ²
DeepFloyd-IF [78]	t				✓			1024 ²
Stable Diff. XL [111]	t						✓	1024 ²
DALL·E 2 [117]	t						✓	1024 ²
DALL·E 3 [105]	t						✓	1024 ²
Midjourney V5 [98]	t						✓	1024 ² -1100 ²
Adobe Firefly [54]	t						✓	2032 ² -2048 ²

Table 4.1. Image generators used in our experiments: GAN-based, DM-based, and commercial tools. For unconditional (u) and conditional (c) models, the real images used in test come from the same datasets used to train the generator, while for text-to-image (t) models synthetic images have been generated using the prompt extracted from the real counterpart.

carry out the detection experiments. These have been carefully selected to avoid all foreseeable biases. For example, when testing synthetic images generated by ProGAN, we use as real counterparts images from the LSUN dataset on which the ProGAN model was trained.

The StyleGAN-T [126], eDiff-I [7], and GigaGAN [72] images were provided by the authors of the respective papers. Instead, we generated

the images ourselves for Score-SDE [133], Stable Diff. [122], DiT [110], and Deepfloyd-IF [78] using the pre-trained models available online. Additional generated images were sourced from two public datasets [148, 31]. For the text-driven commercial tools, we used a recently proposed dataset [8]. Pristine and synthetic images with similar semantic content are built by extracting textual descriptions from a real dataset (RAISE [35]) and using them as prompts for generating synthetic images. A few examples of such images are shown in Fig. 4.2.

To consider a more realistic scenario, we simulate images shared on social networks and subject to post-processing operations: random cropping that can vary in a range from $\frac{5}{8}$ to the full size of the image, resizing to 200×200 pixels, and JPEG compression with a random quality factor between 65 and 100.

To measure performance, we consider three metrics. The area under the receiver operating curve (AUC) and average precision (AP) are both threshold-independent. The Accuracy instead, given by the number of correct predictions divided by the number of tested images, depends on the decision threshold. We always use a fixed threshold of 0.5, to simulate a realistic scenario in which no prior information on the data under test is available to carry out calibration.

4.2 CLIP for synthetic image detection

We propose a simple procedure to distinguish real images from synthetic images based on features extracted from the image encoder of CLIP ViT L/14. The design of the detector consists of the following four steps:

1. collect N real images $\{R_1, \dots, R_N\}$ with the corresponding captions $\{t_1, \dots, t_N\}$;
2. use the captions to feed a text-driven image generator, $G(\cdot)$, so as to obtain N synthetic images $\{F_1, \dots, F_N\}$ with $F_i = G(t_i)$. Now we have N real / fake pairs that share the same textual description;
3. feed CLIP with the N real and N fake images and collect the corresponding feature vectors $\{\mathbf{r}_1, \dots, \mathbf{r}_N\}$ and $\{\mathbf{f}_1, \dots, \mathbf{f}_N\}$ extracted at the output of the next-to-last layer, with $\mathbf{r}_i = \text{CLIP}(R_i)$ and $\mathbf{f}_i = \text{CLIP}(F_i)$;

4. use these $N+N$ vectors to design a linear SVM classifier.

If the real images have low-quality associated captions, or no caption at all, these can be generated by a dedicated tool such as BLIP [81]. Real and fake images are coupled to avoid possible semantic biases, which proves beneficial when the number of images is very small, *e.g.*, $N = 10$. We chose to extract features from the next-to-last layer rather than the last layer on the basis of preliminary experiments.

4.2.1 Influence of the classifier and layer

We selected a SVM Classifier amongst a set of simple classifiers since it has provided us with the best performance. In this subsection, we are going to evaluate this choice against other classifiers: Logistic Regression (LR), Mahalanobis distance (MAH), Gaussian Naive Bayes classifier (GNB), Soft voting k-Nearest Neighbor (SNN) [100]. In Figure 4.3, we show the results in terms of the average AUC over the dataset described in Section 4.1 using both original images and their version impaired by common post-processing steps such as recompression and resizing. For all detectors, we consider the versions trained with 1000 and 10000 reference images per class, respectively referred to as 1k and 10k.

In all cases, the SVM classifier seems to ensure the best performance, followed closely by the Logistic Regression, while the other classifiers provide a less consistent performance. Furthermore, in the same image, we show that using features from the next-to-last layer is almost always preferable to using features from the last layer, and this happens always with the SVM classifier. These results motivated our design choices.

4.2.2 Influence of the reference set size

We start by investigating the role of the most important parameter of the proposed detector, the number of real / synthetic samples N in the reference set, which in this experiment includes COCO real images and Latent Diffusion synthetic images. Fig. 4.4 shows results for N ranging from 10 to 100k in log scale. Results are in terms of AUC, AP and Accuracy. When N is less than 10k, we perform several runs with different sets of images and report the average metrics, *e.g.*, for N equals 1k, we average on 10 runs.

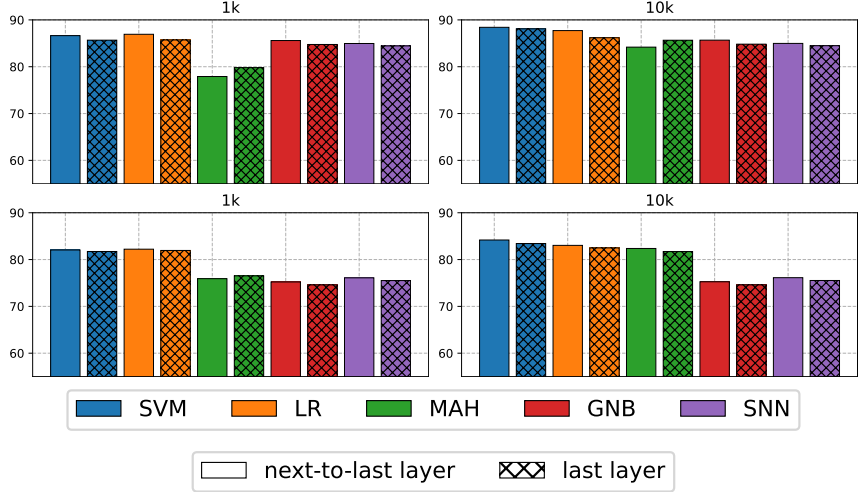


Figure 4.3. Average AUC for our approach considering different classifiers and different features on the original dataset (top) and after random compression and resizing (bottom). Left: 1k images, Right: 10k images.

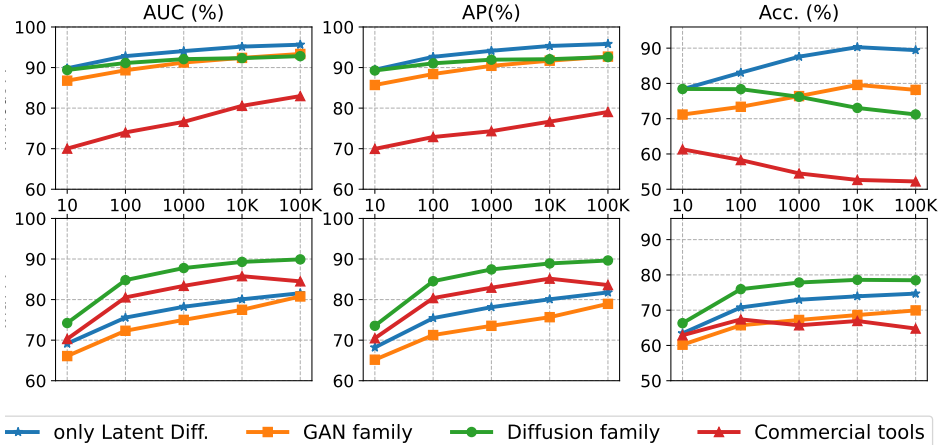


Figure 4.4. Performance of the CLIP-based detector as a function of the number of real and synthetic images in the reference set. We show AUC, AP and Accuracy on the original dataset (Top) and on post-processed images that simulate a realistic scenario (Bottom).

Let’s focus first on the top row, which shows results on images that were neither recompressed nor resized. In this simpler scenario, both AUC and AP values are extremely high, consistently above 85% for all known families of synthetic images (GAN, Diffusion) and significantly worse only for images generated by commercial tools. It is particularly remarkable that very good results are achieved with as few as 10+10 reference images, proving that this lightweight solution is fully viable. The performance improves with N , growing up to 5-10% at the 10k plateau. We note, in passing, that AUC and AP curves are very similar, so we drop the latter for brevity in further experiments.

In the rightmost figure, we see that, unlike AUC and AP, accuracy does not always improve as N increases. This is because AUC and AP are threshold-independent integral metrics. They tell us how well a method with a perfectly calibrated threshold might work. Instead, the precision depends critically on the selected threshold which, in the absence of any prior knowledge, is set at 0.5. For Latent, 0.5 is a good threshold and accuracy improves as N grows. Instead, in general, the optimal threshold moves further away from 0.5 as N increases and the precision decreases significantly. So, there is a trade-off between optimality and robustness in our truly OOD scenario, where no calibration data exists. On the other hand, if we had data from the target class, even just 10 images, we could use them to design an *ad hoc* classifier.

The bottom row of the figure shows results for images that have been compressed and/or resized. As expected, performance degrades slightly in this scenario, but continues to be very good, with AUC between 75% and 90% and accuracy between 65% and 80% at $N=10k$. These are excellent results, well beyond the current state of the art, as we will show later. In this case, using only 10+10 images does not seem advisable but the performance is already almost optimal at $N=100$.

4.2.3 Influence of the reference set content

We found that not only the quantity but also the quality of images in the reference set significantly impacts performance. This is intuitive with a very small reference set, like 10+10 images, where low quality or limited diversity could have catastrophic consequences. However, similar effects are observed even with a larger reference set. Tab. 4.2 shows the results in

Datasets	augm.	GAN family	Diffusion family	Commerc. tools	AVG
COCO + Latent		92.4	92.6	80.5	88.5
COCO + Latent	✓	89.3	91.8	87.0	89.4
COCO + ProGAN		93.6	90.0	65.6	83.1
COCO + ProGAN	✓	91.3	90.1	82.3	87.9
LSUN + Latent		88.7	82.7	65.6	79.0
LSUN + Latent	✓	87.9	79.9	80.3	82.7
LSUN + ProGAN		94.1	79.2	44.7	72.7
LSUN + ProGAN	✓	95.0	82.7	67.5	81.7

Table 4.2. AUC performance of the CLIP-based detector for various real/fake data, with and w/o augmentation (resizing/compression).

terms of AUC obtained using various combinations of real and synthetic images in the reference set, that is, $(\text{Real}, \text{Synth}) \in \{\text{COCO}, \text{LSUN}\} \times \{\text{Latent}, \text{ProGAN}\}$, with and without augmentation. In all cases, the reference set size is 10,000+10,000. The results are clearly influenced by the specific combination. Indeed, considerable degradation occurs when the LSUN dataset is used instead of COCO to draw the real samples and, to a lesser extent, when using ProGAN instead of Latent for the synthetic images.

We conjecture that both LSUN and ProGAN lack the diversity necessary to adequately describe the decision domain. Furthermore, unlike COCO, the LSUN dataset exhibits several biases: the images represent only a few well-defined categories, have all the same size, and most of them are compressed with the same quality factor. In Tab. 4.2, we see that data augmentation only marginally reduces the effects of the dataset used.

4.2.4 Influence of pre-training

We experimented also with various versions of CLIP ViT L/14 [67], trained on different datasets, observing significant changes in performance. Fig. 4.5 shows, the AUC, AP and Accuracy averaged on all families of models as a function of the pre-training dataset. Each curve refers to

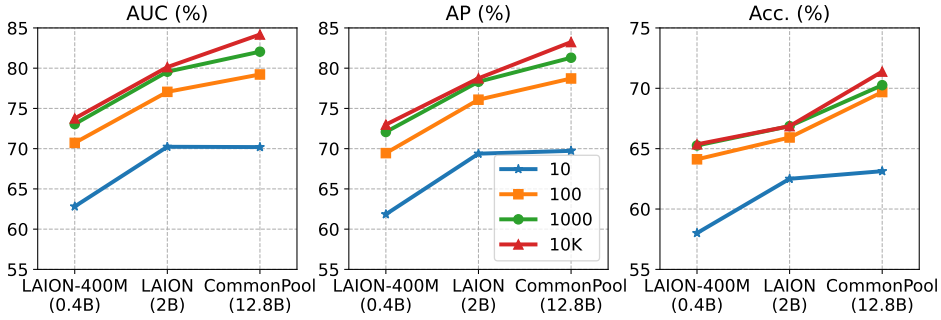


Figure 4.5. Performance of the CLIP-based detector as a function of the pre-training. We show AUC, AP and Accuracy on post-processed images for models pre-trained on LAION-400M (0.4B images), LAION (2B) and CommonPool (12.8B).

a different value of N and we consider only the case of post processed images for brevity. For both metrics and all values of N there is a steady increase in performance as increasingly larger datasets are used for pre-training, from LAION-400M (0.4B images) [128], to LAION (2B) [127], to CommonPool (12.8B) [57]. Overall, there is a ten-point improvement from the smallest to the largest dataset. This confirms the importance of pre-training a large VLM on the largest possible collection of different images.

4.3 Comparison with the state-of-the-art

In this Section we perform an extensive comparison with SoTA methods in different scenarios. We consider four versions of our approach, with 1,000+1,000 (1k) or 10,000+10,000 (10k) real+fake images in the reference set, and with (+) or without compressed/resized images for augmentation. To ensure a fair comparison, we only include SoTA methods with code and/or pre-trained models publicly available online. They are listed in Tab. 4.3.

Generalization analysis.

In the two Tables 4.4 and 4.5 we show the results in terms of AUC on 24 generative models. For each method, the results obtained on the same

Method	Real/Synth. Training	Size (k)	Aug.	Test Strategy
[148] Wang et al.	LSUN / ProGAN	360 / 360	✓	global pooling
[20] PatchFor.	CelebA,FF / various	84 / 272		resizing
[60] Grag. et al.	LSUN / ProGAN	360 / 360	✓	global pooling
[92] Mand. et al.	various / various	232 / 386	✓	patch aggregation
[85] Liu et al.	LSUN / ProGAN	360 / 360	✓	global pooling
[31] Corvi et al.	COCO,LSUN / Latent	180 / 180	✓	global pooling
[137] LGrad	LSUN / ProGAN	72 / 72	✓	resizing
[104] Ojha et al.	LSUN / ProGAN	360 / 360	✓	cropping
[150] DIRE-1	LSUN-Bed / ADM	40 / 40		resizing
[150] DIRE-2	LSUN-Bed / St.GAN	40 / 40		resizing
[135] NPR	LSUN / ProGAN	72 / 72		resizing
Ours 1k	COCO / Latent	1 / 1		resizing
Ours 1k+	COCO / Latent	1 / 1	✓	resizing
Ours 10k	COCO / Latent	10 / 10		resizing
Ours 10k+	COCO / Latent	10 / 10	✓	resizing

Table 4.3. List of methods. For each method, we report the datasets of real and synthetic images used for training, their sizes, whether or not augmentation is used, and the testing strategy.

Method	Diffusion family										Deepfloyd-IF II stage
	Score SDE	ADM	GLIDE	Latent Diff.	Stable Diff.	DeepFl. IF	Ediff-I	DiT	SDXL	DDPM	
Wang et al.	32.9	64.3	48.5	59.2	41.5	78.0	64.9	58.6	54.3	31.6	43.1
PatchFor.	83.3	74.8	96.2	78.1	62.4	62.7	78.7	83.1	68.4	98.4	83.4
Grag. et al.	92.1	74.7	62.8	91.9	52.5	69.9	69.6	65.3	58.0	49.7	71.2
Mand. et al.	99.8	56.5	40.5	70.0	36.8	47.2	65.0	59.1	27.0	99.9	49.2
Liu et al.	95.4	82.5	76.5	97.6	77.4	72.2	98.7	88.0	31.1	22.8	97.4
Corvi et al.	65.0	80.7	91.9	100.	100.	99.9	85.7	100.	100.	100.	68.5
LGrad	80.6	76.9	66.1	81.1	61.5	68.8	74.1	56.2	57.2	9.8	75.0
Ojha et al.	58.4	86.7	80.8	85.7	89.5	92.9	80.6	77.8	85.1	77.7	60.8
DIRE-1	44.3	75.7	71.4	68.7	39.4	98.9	99.1	99.6	47.1	23.1	99.4
DIRE-2	41.0	70.1	70.1	69.3	46.9	97.0	98.2	98.3	42.8	14.0	95.6
NPR	91.1	86.3	79.3	90.2	64.5	91.6	80.1	78.4	76.7	92.4	90.9
Ours 1k	89.1	81.1	99.9	94.1	87.6	96.5	98.5	94.1	87.8	95.1	99.7
Ours 1k+	84.3	75.2	99.6	81.6	89.8	98.0	99.1	92.5	88.9	94.9	99.7
Ours 10k	89.0	81.4	99.9	94.2	90.7	97.0	98.7	95.0	87.4	93.9	99.8
Ours 10k+	89.2	79.9	99.7	84.7	91.3	97.9	99.4	94.0	90.1	98.1	99.8

Table 4.4. Comparison with SoTA methods in terms of AUC on images generated by Diffusion Models. For our approach we show four variants: 1k and 10k indicate the number of real and fake training images, + indicates augmentation (compression/resizing). Results on the dataset used for training are in light gray, while bold underlines the best performance for each dataset with a margin of 1%.

Method	GAN family									Commercial tools			
	Pro GAN	Style GAN2	Style GAN3	Style GANT	Giga GAN	BigGAN	EG3D	Diff. Proj. GAN	GALIP	DALL-E2	DALL-E3	Midj.	Adobe Fref.
Wang et al.	100.	96.5	98.5	98.9	66.6	92.7	94.4	89.8	89.7	64.8	10.9	40.2	84.8
PatchFor.	92.3	84.5	91.8	91.2	64.7	85.5	69.8	92.6	98.2	41.9	52.7	57.8	49.4
Grag. et al.	100.	99.8	97.5	98.8	82.8	98.7	98.9	100.	96.2	58.3	2.4	43.1	63.5
Mand. et al.	96.2	93.8	100.	92.6	61.8	92.2	100.	64.5	77.6	14.5	14.7	24.3	36.7
Liu et al.	100.	99.8	98.4	98.5	98.2	94.7	99.0	99.5	94.3	70.4	0.2	40.7	11.8
Corvi et al.	79.4	73.7	50.0	97.1	63.4	83.4	25.2	96.6	87.7	69.4	60.8	100.	98.0
LGrad	100.	91.2	83.8	81.8	82.2	77.2	68.8	99.5	56.7	58.6	37.9	56.3	40.6
Ojha et al.	100.	93.9	92.3	98.2	96.0	99.6	92.6	97.4	98.6	95.2	36.4	66.2	97.5
DIRE-1	50.6	56.9	47.8	99.9	74.1	99.8	50.1	49.8	100.	44.7	47.6	51.0	57.4
DIRE-2	54.2	52.5	43.0	99.6	76.0	98.6	46.1	50.2	99.3	41.0	49.6	47.8	43.0
NPR	100.	85.6	77.0	96.4	88.7	86.8	53.3	100.	90.7	39.5	48.7	77.0	32.1
Ours 1k	98.9	90.5	85.5	100.	81.3	96.9	87.0	99.3	100.	89.0	70.0	73.0	74.4
Ours 1k+	91.4	80.9	84.0	99.8	74.7	92.0	76.4	89.6	99.9	83.6	93.6	78.7	85.1
Ours 10k	99.8	91.8	86.8	100.	83.6	98.2	87.7	99.7	100.	89.2	77.6	75.3	80.1
Ours 10k+	93.4	87.1	87.6	99.9	78.5	93.1	79.9	92.5	100.	86.3	92.9	81.7	87.2

Table 4.5. Comparison with SoTA methods in terms of AUC on images generated with GANs and commercial tools. For our approach we show four variants: 1k and 10k indicate the number of real and fake training images, + indicates augmentation (compression/resizing). Results on the dataset used for training are in light gray, while bold underlines the best performance for each dataset with a margin of 1%. The last column shows the average over all datasets.

dataset used for training are in light gray, since the ID scenario is of little interest for our analysis. The items in bold, instead, highlight the best OOD performance for each dataset, considering a margin of 1%. We can observe that methods trained on one GAN dataset generally perform well on other datasets of the same family but fare much worse on DM datasets. Then, the roles change for methods trained on DM datasets. This is unsurprising, as generators of the same family share architectural details that leave similar traces on the generated images. No SoTA method performs uniformly well on all datasets. In contrast, the proposed lightweight CLIP-based detector consistently delivers strong performance. The version with 10k+10k reference images, without augmentation, outperforms the best competitor by +6.5% in terms of average AUC. For GAN-based generators, the proposed CLIP-based detectors keep providing the best performance, generally much better than SoTA methods, except for the Liu method which works almost at the same level. However, this as well as many other methods, when evaluated on images coming from commercial tools, show a catastrophic loss in performance. This is the most realistic and interesting scenario, with images of unknown origin and no prior

Method	Families of Generators			Average AUC/Acc
	GAN AUC/Acc	Diffusion AUC/Acc	Comm. Tools AUC/Acc	
Wang et al.	91.9 / 64.7	52.4 / 50.3	50.2 / 49.8	64.8 / 54.9
PatchFor.	85.6 / 56.2	79.0 / 50.3	50.5 / 50.0	71.7 / 52.2
Grag. et al.	97.0 / 87.8	68.9 / 56.3	41.8 / 43.6	69.2 / 62.6
Mand. et al.	86.5 / 77.8	59.2 / 57.3	22.6 / 32.9	56.1 / 56.0
Liu et al.	98.0 / 86.6	76.3 / 69.3	30.8 / 42.9	68.4 / 66.2
Corvi et al.	73.0 / 53.8	90.2 / 74.3	82.1 / 62.8	81.7 / 63.6
LGrad	82.4 / 72.9	64.3 / 56.6	48.3 / 52.1	65.0 / 60.5
Ojha et al.	96.5 / 85.6	79.6 / 62.6	73.8 / 68.3	83.3 / 72.2
DIRE-1	69.9 / 68.4	69.7 / 72.8	50.2 / 49.9	63.3 / 63.7
DIRE-2	68.8 / 62.9	67.6 / 65.6	45.3 / 49.9	60.6 / 59.5
NPR	86.5 / 78.9	83.8 / 68.3	49.3 / 50.1	73.2 / 65.8
Ours 1k	93.3 / 79.0	93.0 / 78.5	76.6 / 54.5	87.6 / 70.7
Ours 1k+	87.6 / 75.9	91.2 / 82.1	85.3 / 72.6	88.0 / 76.8
Ours 10k	94.2 / 81.4	93.4 / 73.4	80.5 / 52.6	89.4 / 69.1
Ours 10k+	90.2 / 76.0	93.1 / 79.9	87.0 / 67.3	90.1 / 74.4

Table 4.6. Comparison with SoTA methods in terms of AUC and Accuracy. We report the mean AUC and Accuracy for each family of generators and the global average.

information on the possible generation process. In this situation, most methods provide unreliable decisions. Corvi’s method works surprisingly well on some datasets, perhaps generated by diffusion models similar to Latent. However, only the proposed CLIP-based approach provides good stable performance across all cases. In particular, the version with resized and compressed images in the reference set prove particularly strong on images generated with the commercial tools. We can observe that SoTA methods present a larger variability in terms of performance, with very good results on some generators and very bad on others. Instead, our method can provide consistently good performance over all the data.

Furthermore, we show the result aggregated by family (GAN, Diffusion, Commercial tools) not only in terms of AUC but also of balanced accuracy computed with a fixed threshold of 0.5. Indeed in a practical scenario a threshold must be set to discriminate between real and fake images, which is not a trivial task as can be seen in Table 4.6. In fact, in the absence of an adequate calibration procedure, a significant drop between AUC and accuracy is often observed.

Method	Diffusion family										
	Score SDE	ADM	GLIDE	Latent Diff.	Stable Diff.	DeepFl. IF	Ediff-I	DiT	SDXL	DDPM	Deepfloyd-IF II stage
Wang et al.	68.0	65.0	60.6	67.1	55.2	50.3	48.3	55.1	64.5	69.6	47.9
PatchFor.	69.0	66.2	83.3	58.8	48.3	61.3	65.0	68.1	63.3	82.9	69.7
Grag. et al.	77.1	77.1	79.8	84.0	53.5	50.6	55.6	66.7	66.6	74.5	54.3
Mand. et al.	64.0	54.8	42.6	52.9	39.4	55.7	54.5	49.8	42.2	61.1	49.0
Liu et al.	62.6	51.1	58.6	50.7	58.6	50.9	64.2	53.9	56.0	58.3	64.4
Corvi et al.	70.4	79.0	93.5	99.3	69.9	60.7	72.1	89.2	61.8	74.7	70.4
LGrad	45.9	49.2	42.3	50.4	54.8	40.7	46.4	49.4	53.2	56.0	38.7
Ojha et al.	79.3	83.7	83.3	81.8	75.0	59.9	68.7	70.1	61.8	90.8	59.1
DIRE-1	44.1	48.0	44.7	46.1	47.0	66.2	62.8	53.2	47.1	38.5	64.6
DIRE-2	44.3	45.1	40.2	45.9	56.4	70.7	72.2	53.0	42.8	42.2	66.3
NPR	44.7	46.9	47.2	47.7	55.4	49.6	54.6	50.9	52.8	54.5	51.3
Ours 1k	74.7	78.1	97.2	77.1	77.6	80.1	86.6	77.5	76.5	76.1	90.0
Ours 1k+	84.0	74.3	99.6	78.2	83.7	94.5	97.1	88.9	89.6	96.4	99.5
Ours 10k	75.8	78.4	97.7	77.8	78.1	81.2	87.1	77.2	76.4	70.9	91.3
Ours 10k+	85.4	77.3	99.7	80.2	85.8	95.9	98.2	91.1	89.9	96.1	99.7

Table 4.7. Comparison with SoTA methods (AUC) on images from Diffusion Models post-processed by random cropping, resizing and compression.

Robustness to perturbations.

The observed trend with respect to unknown models is further accentuated when the images undergo post-processing, as shown in Tab. 4.7 and in Tab. 4.8. These impairments attenuate forensic traces, to the point that most SoTA methods become essentially useless, performing no better than random chance, particularly on unknown commercial models. CLIP-based detectors, instead, keep providing a good performance.

We further explore the robustness in more detail in Figure 4.6 as a function of JPEG Quality Factor, ranging from 100 to 60, and resizing scale, going from 125% to 25%. As the generators evolve through time, to do new compression standard get proposed. So in addition, we consider also another compression method, WebP, which is gaining popularity on social networks but has been rarely considered in experimental analyses. Images are from the SynthBuster dataset [8], generated by Stable Diffusion 2, SDXL, DALL·E 2, DALL·E 3, Midjourney, and Firefly. We show results (AUC) only for the SoTA methods that proved most competitive in terms of robustness, Corvi and Ojha. For our method we consider again the versions with 1k or 10k real and fake reference images, and the corresponding versions, 1k+ and 10k+, with the same number of images including random compression and resizing. Ojha suffers some performance losses, up

Method	GAN family									Commercial tools			
	Pro GAN	Style GAN2	Style GAN3	Style GANT	Giga GAN	BigGAN	EG3D	Diff. Proj. GAN	GALP	DALL·E2	DALL·E3	Midj.	Adobe Firef.
Wang et al.	100.	86.6	88.4	61.7	59.2	82.5	84.7	80.8	92.2	46.2	27.7	46.7	55.9
PatchFor.	57.9	51.8	57.0	50.6	53.8	58.7	60.3	55.6	72.3	64.3	63.3	59.0	65.1
Grag. et al.	100.	95.4	90.9	94.4	64.4	97.6	92.7	99.2	99.1	55.2	25.1	48.5	60.2
Mand. et al.	81.1	79.3	87.2	49.1	49.3	71.4	85.8	60.3	78.1	47.9	42.3	35.2	53.4
Liu et al.	64.3	55.1	50.1	57.3	45.4	57.7	51.6	62.9	65.2	44.4	61.8	52.6	53.1
Corvi et al.	77.5	74.7	69.4	82.1	66.6	77.3	64.5	93.3	87.4	65.9	32.4	51.9	58.1
LGrad	56.3	58.3	49.8	52.3	43.5	53.5	53.2	56.9	57.0	41.8	53.5	50.4	51.8
Ojha et al.	99.8	75.5	75.4	91.1	88.5	94.8	79.3	84.7	95.7	63.2	41.7	40.6	52.9
DIRE-1	48.4	42.5	39.1	53.5	54.3	47.2	35.2	35.2	47.4	44.6	47.6	51.0	57.4
DIRE-2	49.3	41.6	38.6	53.8	55.0	44.6	33.6	37.5	46.2	40.9	49.7	47.8	43.0
NPR	54.5	48.5	41.9	54.0	44.8	52.6	40.2	48.8	57.8	50.0	67.5	50.8	55.5
Ours 1k	85.0	64.0	66.6	90.2	75.2	84.6	64.8	73.4	98.3	77.9	77.4	63.1	70.5
Ours 1k+	78.7	62.5	68.4	97.5	67.9	89.6	60.5	69.6	99.9	81.2	90.9	77.6	83.7
Ours 10k	85.7	65.5	68.1	90.5	74.7	84.1	64.3	73.7	98.1	78.2	76.4	65.0	72.2
Ours 10k+	82.8	67.4	70.7	98.4	71.9	92.4	63.1	74.1	99.9	83.8	90.1	79.4	85.5

Table 4.8. Comparison with SoTA methods (AUC) on images from GANs and Commercial tools post-processed by random cropping, resizing and compression.

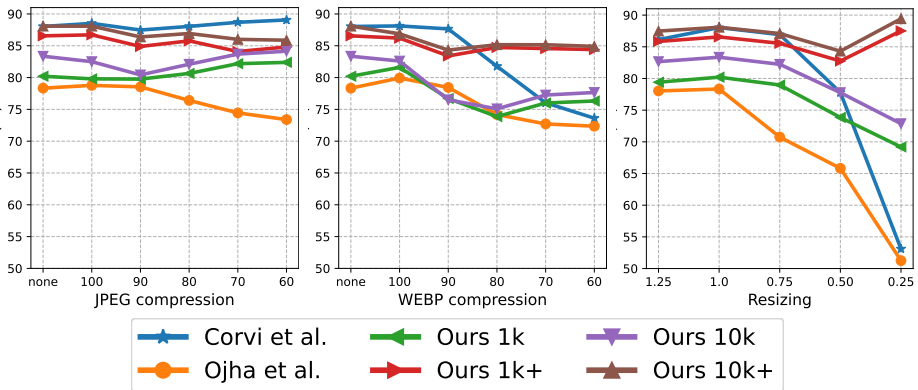


Figure 4.6. Robustness analysis in terms of AUC carried out on the Stable Diff. 2, SDXL, DALL·E 3, DALL·E 2, Midjourney, and Firefly generation models of the SynthBuster dataset [8].

to 10 points, in the presence of compression, be it JPEG or WebP, and a much more rapid decline with resizing, to the point of becoming useless in the presence of a 25% rescaling. Corvi was trained with strong augmentation and, in fact, is not affected at all by JPEG compression, while it presents acceptable losses with strong WebP compression (not seen in training), and strong resizing. The most remarkable outcome of this ex-

Method	Families of Generators			Average AUC/Acc
	GAN AUC/Acc	Diffusion AUC/Acc	Comm. Tools AUC/Acc	
Wang et al.	81.8 / 58.9	59.2 / 50.3	44.1 / 49.9	61.7 / 53.0
PatchFor.	57.6 / 50.1	66.9 / 50.3	62.9 / 50.1	62.5 / 50.2
Grag. et al.	92.6 / 68.2	67.3 / 50.6	47.3 / 50.0	69.0 / 56.3
Mand. et al.	71.3 / 55.6	51.4 / 51.5	44.7 / 49.6	55.8 / 52.2
Liu et al.	56.6 / 51.8	57.2 / 51.0	53.0 / 50.7	55.6 / 51.2
Corvi et al.	77.0 / 55.5	76.5 / 60.8	52.1 / 50.1	68.5 / 55.5
LGrad	53.4 / 51.0	47.9 / 50.3	49.4 / 50.4	50.2 / 50.6
Ojha et al.	87.2 / 74.2	74.0 / 60.6	49.6 / 50.6	70.2 / 61.8
DIRE-1	44.7 / 50.0	51.1 / 50.6	50.1 / 49.9	48.7 / 50.2
DIRE-2	44.5 / 48.3	52.7 / 52.7	45.3 / 49.9	47.5 / 50.3
NPR	49.2 / 50.2	50.5 / 49.8	56.0 / 50.1	51.9 / 50.0
Ours 1k	78.0 / 71.1	81.0 / 71.7	72.2 / 60.7	77.1 / 67.8
Ours 1k+	77.2 / 70.0	89.6 / 80.2	83.4 / 65.7	83.4 / 72.0
Ours 10k	78.3 / 65.2	81.1 / 63.8	73.0 / 52.4	77.4 / 60.5
Ours 10k+	80.1 / 71.8	90.8 / 81.1	84.7 / 66.4	85.2 / 73.1

Table 4.9. AUC/Accuracy in the presence of post-processing. All images have been randomly cropped, resized and compressed to simulate a realistic scenario on the web.

periment, however, is the robustness of the CLIP-based detectors. Both the 1k+ and 10k+ versions, those with augmentation, are basically insensitive to compression, no matter JPEG or WebP, and resizing. The versions without augmentation suffer some loss of performance but not nearly as dramatic as for the reference methods, and remain effective in all conditions. We then evaluate the performance on post processed images in terms of both Accuracy and AUC. Our approach provides good results even in this challenging context (Tab. 4.9) where most state of the art methods are equivalent to flipping coins. The inconsistency in performance on images without post processing, shown previously, is even more noticeable on data that have undergone random compression and resizing. In fact, for most of the competitors there is a dramatic performance loss, sometimes to a random guess level. Some other methods, such as, Ohja

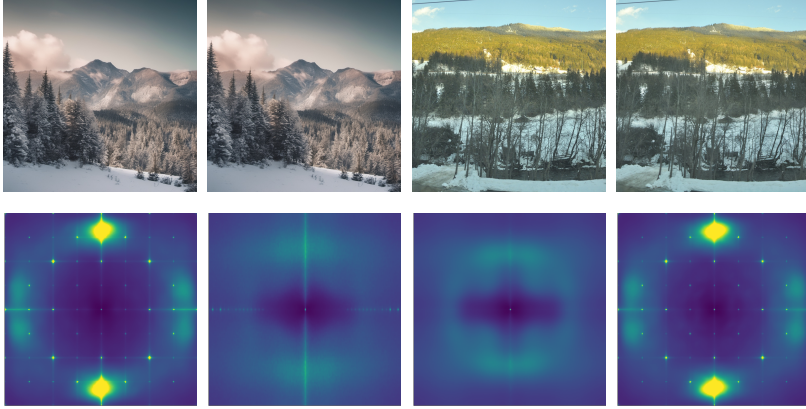


Figure 4.7. Top (from left to right): a synthetic image generated by Stable Diffusion XL [111] and its $4\times$ decimated version; a real image and the corresponding image processed by the autoencoder of Stable Diffusion XL. Bottom: Fourier spectra of the noise residuals for images shown on the top. A suitable decimation removes Fourier peaks in synthetic images, while passing a real image through an autoencoder creates new peaks.

et al., preserve a good performance and the same happens for ours. Of course, in this more challenging scenario our best variant is the one that includes some form of augmentation in the reference data but even the variant trained on the original data provides satisfying results.

4.4 Beyond low-level forensic traces

As previously mentioned, conventional detectors often experience significant performance degradation when images are resized or compressed, as the subtle forensic traces they depend on are significantly reduced. In contrast, the proposed CLIP-based detector keeps working well under these conditions, suggesting that it relies on higher-level semantic features. To further investigate this hypothesis, we are now going to show some targeted experiments.

Removing/inserting low-level traces in images.

We take synthetic images generated by Stable Diffusion 1.4, Stable Diffu-

Method				downsample by 1/4			add low-level traces		
	St.Diff.	St.Diff.	SDXL	St.Diff.	St.Diff.	SDXL	St.Diff.	St.Diff.	SDXL
	1.4	2.0		1.4	2.0		1.4	2.0	
Corvi et al.	100.	100.	100.	41.6	46.2	45.6	79.3	80.4	67.6
Ours 1k	88.8	87.0	87.8	85.8	82.2	79.3	89.3	82.5	80.8
Ours 1k+	90.2	89.4	88.9	89.8	91.1	90.0	90.8	86.1	83.2
Ours 10k	93.8	90.6	87.4	86.8	84.4	80.5	93.7	86.2	80.6
Ours 10k+	94.0	90.4	90.1	95.9	92.6	91.6	94.1	86.7	84.0

Table 4.10. Results in terms of AUC without attacks (first group of columns) and after attacks on synthetic images (second group) and real images (third group). Corvi et al.[31], based on low-level forensic traces, is severely affected by the attacks while the proposed method keeps working well in all conditions.

sion 2.0 [122], Stable Diffusion XL [111]) and downsample them by 1/4. Fig. 4.7 shows, on the left, the high- and low-resolution images (top) along with their spectra (bottom) for Stable XL. After reducing the resolution, the peaks in the Fourier domain completely disappeared, because they were caused by the aliasing effect of the oversampling filters, which can be removed by appropriate decimation. In a complementary experiment, we take real images from the RAISE dataset and pass them through the autoencoders used by Stable Diffusion XL.

In Table 4.10 we quantify the effects of such attacks on 1,000 real and 1,000 synthetic images, reporting the AUC for our method and the best competitor before the attack (first group of columns) after removing traces from synthetic images (second group) and after adding them to real images (third group). Before the attack, all detectors work quite well, with the method proposed by Corvi et al., based on low-level features, reaching perfect discrimination. After the attacks, however, the performance of Corvi et al. degrades dramatically, while the proposed method suffers only a very limited loss. In summary, the CLIP-based detector appears to withstand various types of image degradations easily, both innocent and malicious. The latter case is particularly relevant: attacks can be carried out with the very aim of making everything appear as synthetic, misleading current forensic detectors based on low-level traces, and undermining public trust in forensic analyses.

Method	Families of Generators			Average AUC/Acc
	GAN AUC/Acc	Diffusion AUC/Acc	Comm. Tools AUC/Acc	
Corvi et al.	73.0 / 53.8	90.2 / 74.3	82.1 / 62.8	81.7 / 63.6
Ours 10k	94.2 / 81.4	93.4 / 73.4	80.5 / 52.6	89.4 / 69.1
Ours 10k+	90.2 / 76.0	93.1 / 79.9	87.0 / 67.3	90.1 / 74.4
Ours fusion	94.4 / 81.9	97.5 / 88.7	88.2 / 65.3	93.4 / 78.6
Ours fusion+	90.9 / 76.3	97.4 / 90.0	92.3 / 77.2	93.5 / 81.2
Corvi et al.	77.0 / 55.5	76.5 / 60.8	52.1 / 50.1	68.5 / 55.5
Ours 10k	78.3 / 65.2	81.1 / 63.8	73.0 / 52.4	77.4 / 60.5
Ours 10k+	80.1 / 71.8	90.8 / 81.1	84.7 / 66.4	85.2 / 73.1
Ours fusion	80.4 / 69.1	84.7 / 71.0	72.7 / 52.6	79.3 / 64.3
Ours fusion+	82.1 / 73.6	92.5 / 82.3	84.5 / 66.6	86.4 / 74.2

Table 4.11. AUC/Accuracy results for Corvi et al. [31], proposed method (10k and 10k+ versions), and their fusion over the three families of generators: GAN, Diffusion, Commercial Tools. Top: original images; bottom: compressed/resized images.

Fusion.

Based on the above results, the proposed CLIP-based detector appears not to rely on the same low-level traces exploited by most of the current detectors. This is further supported by the scatter plot of Fig.4.8 where each point gives the scores of the CLIP-based and Corvi detectors for a given image. The two sets of scores are almost orthogonal, as if they depended on uncorrelated features, a property that paves the way for appropriate fusion strategies. We implemented a simple decision rule where the image is declared real only if both detectors agree on this choice. Tab.4.11 reports AUC and Accuracy results for Corvi, the proposed method (10k and 10k+ versions), and their fusion. Results are given as averages over the three families of generators: GAN, Diffusion and Commercial Tools. The fusion ensures a further boost in performance over the proposed CLIP-based detector, both in terms of AUC (+3.4%) and Accuracy (+6.8%). A smaller improvement is observed when images are resized/compressed, arguably because low-level traces are more compromised in this case.

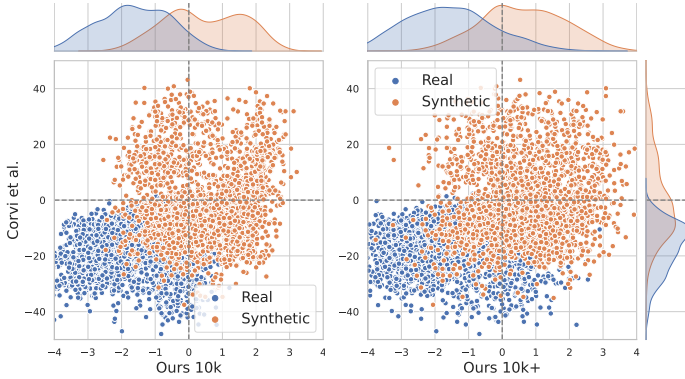


Figure 4.8. Scatter plots of scores provided by Corvi et al.[31] (y -axis) and proposed method without and with augmentation (x -axis) on 2400+2400 images sampled from the 18 datasets of Tab.4.1.

4.5 Few-shot analysis in the wild

In this section we present an experiment in a few-shot scenario. The idea is to explore the ability of our detector to work with very limited data and adapt to a situation where only a few real and synthetic examples are available. We downloaded 500 real images and 1,500 synthetic images from X from three different generators DALL·E 3, Midjourney and Firefly. To understand which generator was used to create a specific image, we relied on tags and annotations present on the social network.

In our few-shot scenario, we take 10 real images and 10 generated images as examples from a specific model and test on all the others images (note that we present average results on 1,000 runs). Results are reported in Tab. 4.12 in terms of AUC and accuracy. The availability of just 10+10 images of the target data provides an impressive performance boost with respect to the same method trained on a dataset much larger but not aligned with the test data (real images come from COCO, synthetic images from Latent diffusion models). It is important to underline that this is a realistic scenario, in which one is called to decide on images generated by a new method and a few sample images are available as a support. We believe that this can represent an interesting direction for the application in the wild or to easily adapt a detector to more challenging situations

where some prior information is available.

To complete our analysis on this dataset we include in Tab. 4.12 the comparison with SoTA methods and our original proposal. We can observe that the performance of all the methods degrade which highlights the increased difficulty to handle a realistic scenario over the web. The best performance can be obtained by using our fusion approach that take the best of the low-level and high-level features and is able to achieve on average more than 90% in terms of AUC which is comparable with the few-shot analysis where some prior knowledge is available.

AUC/Acc	Social network: X			
Method	DALL-E3	Midjourney	FireFly	Average
Wang et al.	22.5 / 50.0	36.7 / 50.0	65.3 / 50.0	41.5 / 50.0
PatchFor.	54.5 / 50.0	59.9 / 50.0	46.5 / 50.0	53.6 / 50.0
Grag. et al.	44.7 / 49.6	58.4 / 49.7	80.5 / 49.9	61.2 / 49.7
Mand. et al.	42.9 / 49.3	53.7 / 50.6	52.7 / 50.1	49.8 / 50.0
Liu et al.	34.5 / 49.7	38.7 / 49.7	54.9 / 49.9	42.7 / 49.8
Corvi et al.	89.3 / 86.8	98.3 / 90.8	79.1 / 52.1	88.9 / 76.6
LGrad	33.7 / 39.0	37.9 / 41.4	61.4 / 58.4	44.3 / 46.3
Ojha et al.	34.4 / 45.6	32.2 / 46.2	73.0 / 65.7	46.5 / 52.5
DIRE-1	54.3 / 52.0	60.7 / 54.3	55.1 / 51.1	56.7 / 52.5
DIRE-2	60.0 / 55.0	58.6 / 53.9	47.7 / 46.5	55.4 / 51.8
NPR	43.8 / 41.7	46.2 / 44.9	65.4 / 62.0	51.8 / 49.5
Ours 1k	69.4 / 50.9	70.2 / 53.3	69.3 / 51.4	69.6 / 51.9
Ours 1k+	82.1 / 73.3	82.8 / 74.5	83.8 / 74.9	82.9 / 74.3
Ours 10k	65.8 / 50.0	69.3 / 51.7	70.7 / 50.9	68.6 / 50.9
Ours 10k+	78.7 / 63.7	82.2 / 69.6	82.7 / 69.6	81.2 / 67.6
Ours fusion	90.0 / 85.9	97.5 / 93.8	77.8 / 57.0	88.4 / 78.9
Ours fusion+	94.7 / 87.3	97.9 / 92.1	85.4 / 74.2	92.6 / 84.5
Ours Few-Shot	98.5 / 92.5	95.8 / 87.4	92.5 / 83.5	95.6 / 87.8

Table 4.12. Results in terms of AUC and accuracy of proposal and SoTA methods on real and synthetic images download from X. The last row reports a few-shot experiment, where we suppose to know in advance 10+10 real/fake images from a specific generator.

4.6 Discussion

We have proposed a simple method based on CLIP features to distinguish real images from synthetic images. Through extensive experimental analysis, we found that:

- CLIP features support a much higher generalization ability than what had been discovered so far. By leveraging just a few examples, not even belonging to the generator under test, a simple CLIP-based detector achieves top performance on a large variety of generators and in the most challenging conditions.
- Maximizing the diversity of reference CLIP features has a positive impact on performance, even when a relatively large number of examples is considered.
- It is known that CLIP’s descriptive power comes from its huge pre-training set. Further increasing this set keeps boosting performance, up to 10%.
- Experimental evidence suggests that CLIP features, even when adapted to forensic applications, are largely independent of low-level forensic traces. This provides some immunity to malicious attacks targeting low-level artifacts and paves the way for further performance improvements through suitable fusion with traditional detectors.

The present study provides a number of interesting insights but leaves much room for future work. We believe that new and better forensic methods can be proposed based on CLIP features. One major area of study is the development of few-shot methods that adapt the detector to the situation of interest on the fly. Future work should also consider interpretability, starting with understanding which forensic features the detector exploits to make its decisions.

Chapter 5

Analysis and detection of synthetic videos

*No man's knowledge here can go
beyond his experience.*

Jonh Locke

In recent years, the field of AI-based video generation has witnessed rapid advancements. Several flexible tools exist, based on diffusion models, that generate high-quality videos from general conditional inputs like text and images [77, 146, 168]. These powerful tools enable professionals to use AI for innovative and creative applications, such as design, marketing, and entertainment. However, the misuse of such technologies also raises ethical and social concerns related to the spread of disinformation, the violation of intellectual property rights and more generally the erosion of trust in digital media [47, 9, 83]. Therefore there is an urgent need to develop effective methods for distinguishing real from AI-generated videos.

Previous works in video forensics focused mostly on facial forgery detection [119] and proposed solutions that are specifically tailored to faces, for example, by analyzing head pose or facial landmarks and appearance [101, 27], inconsistencies in skin texture or lip motion [88, 63], unnatural heart rate variations [30] or identity-based biometrics [1, 113]. However, these approaches are inherently limited to facial content and struggle to generalize to more complex and semantically rich synthetic videos gener-

ated by modern diffusion models. Only a few works have been designed so far for fully generated video detection [24, 87, 91, 5, 80]. Such methods focus primarily on designing novel and often complex architectures, neglecting the critical role of the data. In fact, in forensic applications, the choice of training and test data is essential to ensure that detectors learn generation-related cues, rather than spurious correlations encoded in the data [141, 89, 19]. Even though the presence of dataset biases is a well-known risk in machine learning, it still forms a significant and underestimated problem in forensics research. In [18], it was shown that a popular benchmark dataset included real and manipulated images compressed at different JPEG quality levels, causing classifiers to learn compression artifacts instead of tampering cues, leading to poor generalization. This same issue was recently observed for the detection of AI-generated images [61], while other works have pointed out the presence of other possible biases for the same task, such as content or resolution bias [19, 116, 62].

Our aim is to design a synthetic video detector that is truly based on generation-specific artifacts. This is because methods that rely exclusively on data-driven learning or semantic errors in generation tend to overfit to the training data or to the artifacts introduced by specific generators, leading to poor generalization. We overcome this limitation by identifying and integrating priors into the synthetic video detectors that leverage forensic traces that are consistently present across diverse generative models.

To this end, we address the following questions:

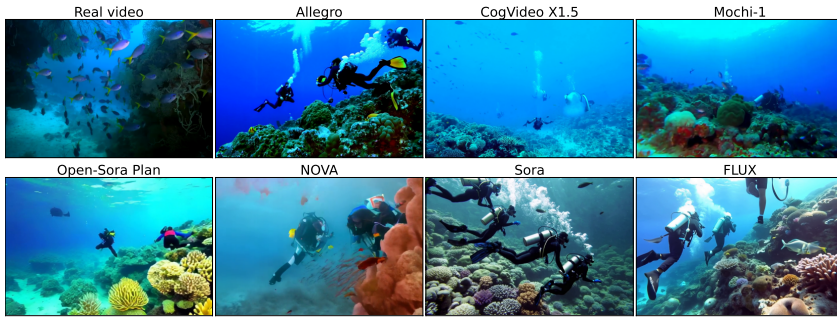


Figure 5.1. Examples of video from our proposed test dataset generated with recent state of the art generators.

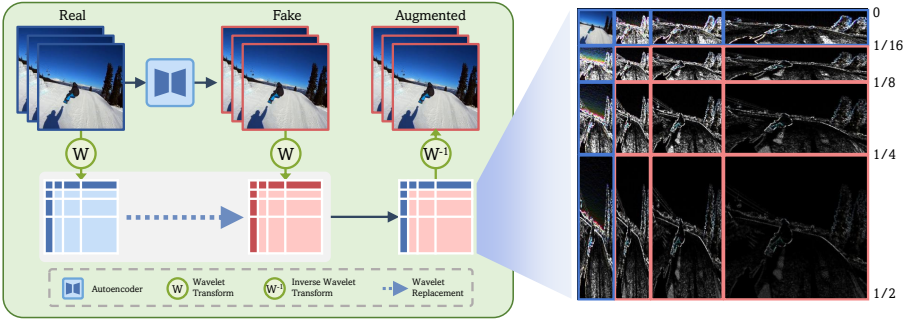


Figure 5.2. Synthetic video generators leave distinct traces, that are observed in the frequency spectrum. We leverage this observation to enhance the generalizability of synthetic video detectors. To this end, we propose a novel training-time data augmentation strategy based on wavelet-bands that forces the model to learn the frequency components that best distinguish real from synthetic content. Fakes are also generated through video autoencoding to avoid semantic bias and to trick the model into exploiting low-level forensic traces left by the modern video generation architectures. Our training paradigm improves the generalizability of the detector without the need for complex algorithms and large datasets that include multiple generators.

- What are good discriminative and robust forensic traces present in AI-generated videos?
- What is a good strategy to exploit them?

As a first step, we identify the hidden forensic traces that are shared by modern video AI-generators and arise from the generative architectures themselves. Prior art has shown that the up-sampling operations inherent in the synthesis network give rise to quasi-periodic patterns that are clearly visible as peaks in the high-frequency portion of the Fourier spectrum [165, 129]. As a consequence, several works exploited high-frequency traces for AI-generated *image* detection [56, 85, 19]. However, high-frequency components are severely degraded by compression, especially by the strong compression usually adopted for videos, with the effect of washing out the most prominent forensic traces (see Fig.5.4). Diffusion-based generated images differ significantly from natural images not only at the high frequencies but also at the medium frequencies, as shown in Chapter 3. This

latter finding is especially important in this context since we show that video compression impacts less the video’s diagonal mid-high frequencies, which therefore turn out to be both discriminative and robust, a good basis to build effective detectors.

To exploit these traces we chose not to work on new detection architectures but, inspired by the recent image-forensics literature [62], to focus on the crucial training phase, in order to emphasize the most meaningful artifacts and thus guide the detector in learning the right features. To this end, we build a dataset of paired real-fake videos with the same caption and carry out two forms of augmentation (see Figure. 5.2): 1) we add controlled fake videos by injecting the architecture-related artifacts on real videos through a video autoencoder. This has already been shown to be effective in avoiding possible semantic biases [116, 62]; 2) we push the model to exploit middle frequencies by a tailored augmentation strategy that works on selected bands of a multiscale wavelet decomposition of the video. This has the advantage of avoiding looking at specific compression cues that affect both real and generated videos.

Overall, we make the following contributions:

- we find that inconsistencies in the middle-high (diagonal) frequency content of synthetic video are discriminative, robust, and common across several different video generators, even more recent models;
- we propose a novel augmentation strategy that works on the wavelet bands to guide the model towards exploiting such cues;
- we show that by including this simple strategy we can outperform current SoTA methods in terms of generalization. Across 15 different generators, our approach yields an average improvement of around 12% in terms of accuracy.

5.1 Dataset

To conduct our experiments, we created a dataset of synthetic videos generated using recent state-of-the-art models. For the real data, we relied on the Panda70M dataset [26], which provides access to YouTube video URLs, precise timestamps for captioned clips, accompanying captions and

a relevance score indicating how well each clip aligns with its caption. We identify clips with favorable characteristics, excluding those that are screen recordings, contain static content, or exhibit minimal camera movement. To ensure a high-quality and diverse dataset, we selected only such type of clips, limiting the choice to a single clip per YouTube video. When multiple clips were available, we chose the one that was at least five seconds long and had the highest alignment score. For our training, we selected 1,700 real videos from the Panda70M dataset and used their captions to generate synthetic samples with one single generator Pyramid Flow [70] (version with 768p miniflux weights). We further augmented the synthetic set by applying the Pyramid Flow Variational Autoencoder (VAE) to the original real videos, resulting in 1,700 more synthetic videos. The full set of 5,100 videos was split into 4,200 for training and 900 for validation. For evaluation, we selected 300 real videos and used their captions to generate synthetic counterparts using five recent video generation models: Allegro [168], Mochi-1 [139], CogVideoX [155] (with CogVideoX1.5-5B weights), NOVA [38], and Open-Sora Plan [82] (with 1.3B weights). To guarantee consistency with the real videos, we compressed all generated videos using exactly the same codec of the real videos, i.e. Advanced Video Coding (H.264). Specifically, we used the Main@L3.1 codec profile and a Constant Rate Factor (CRF) randomly sampled between 16 and 30. Additionally, the same captions were used to produce other synthetic videos via two online platforms: Sora [15] and FLUX [55]. Overall, using 7 generators, we created 2,100 synthetic videos with a frame-rate that goes from 12 fps (NOVA) to 30 fps (Mochi-1, Sora) and a spatial resolution that varies from 640×352 pixels (Open-Sora Plan) to 1360×768 pixels (CogVideoX). Some examples can be seen in Fig.5.1, and at the end of the chapter in Fig.5.9 and Fig.5.10.

5.2 Proposed Method

5.2.1 Video forensic artifacts

A key step in developing effective detectors for synthetic videos is to identify the most discriminative features. Generators can introduce visual and semantic artifacts, such as geometric distortions, layout inconsisten-

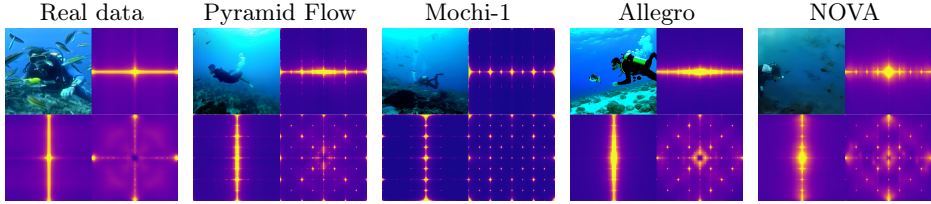


Figure 5.3. From left to right: a real video from the dataset proposed in [26] and synthetic videos generated using the same associated prompt “A scuba diver in the ocean surrounded by fishes” with Pyramid Flow, Mochi-1, Allegro and NOVA. For each of them we show its spatial power spectrum $S_{yx}(u, v)$ (bottom-right), and its temporal-spatial power spectra $S_{tx}(w, v)$ and $S_{yt}(u, w)$ (top-right and bottom-left).

cies, color mismatches, or temporal incoherence. However, these visible errors tend to reduce as generative technologies continue to evolve and it is likely that they may soon disappear altogether. To ensure long-term robustness, our approach focuses on low-level artifacts intrinsic to the generative architecture. In this section we analyze such subtle traces.

Fingerprints in the Fourier domain.

Early research on synthetic images has shown that, much like real cameras, which imprint each photo with a unique device/camera specific signature [90, 34], synthetic architectures also leave a distinct fingerprint in every generated image [94, 158]. Researchers used these artificial fingerprints to develop methods capable of identifying synthetic images and even trace them back to the specific architecture used for their generation [4]. For synthetic images, these artifacts are clearly visible in the frequency domain [129], as we have shown in Chapter 3 and the same holds for frames extracted from synthetic videos [143]. In Fig. 5.3, we show the average power spectra of the residual videos obtained by removing the high-level semantic content of the original videos through a denoiser. More specifically, we compute the residual of the i -th video as

$$r_i(m, n, p) = x_i(m, n, p) - \mathcal{D}(x_i(m, n, p); \sigma), \quad (5.1)$$

where $x_i(m, n, p)$ is the p -th frame of the video with size $M \times N \times P$,

m and n are spatial coordinates and $\mathcal{D}$ denotes the CNN-based denoiser proposed in [163], which is applied to each frame individually with the noise parameter σ set to 1. Then, we evaluate the 3D Fourier transform:

$$R_i(u, v, w) = \sum_{m,n,p=1}^{M,N,P} r_i(m, n, p) e^{-j2\pi(\frac{u}{M}m + \frac{v}{N}n + \frac{w}{P}p)} \quad (5.2)$$

and compute the spatial power spectrum by averaging the spectra of all frames of $I = 150$ videos:

$$S_{yx}(u, v) = \frac{1}{I} \sum_{i=1}^I \frac{1}{P} \sum_{w=1}^P |R_i(u, v, w)|^2. \quad (5.3)$$

The spatial power spectrum accounts for the fraction of the total image power concentrated at a given (vertical, horizontal) frequency pair $(\frac{u}{M}, \frac{v}{N})$. On the other hand, by averaging the power spectra of all videos along the rows we obtain the temporal-spatial power spectrum,

$$S_{tx}(w, v) = \frac{1}{I} \sum_{i=1}^I \frac{1}{M} \sum_{u=1}^M |R_i(u, v, w)|^2, \quad (5.4)$$

which accounts for the fraction of the total image power concentrated at a given (temporal, horizontal) frequency pair $(\frac{w}{P}, \frac{v}{N})$. Similarly, we can compute the temporal-spatial power spectrum $S_{yt}(w, v)$ concentrated at a given (vertical, temporal) frequency pair. The spatial power spectra of Fig. 5.3 exhibit typical spectral peaks (visible as bright spots) caused by the upsampling process in the generative architecture. Interestingly, similar peaks are also visible along the temporal direction. Notice, however, that these peaks are not present in the real image (Fig. 5.3, first column).

Impact of video compression.

In realistic scenarios requiring efficient storage and transmission, synthetically generated videos are usually compressed using standard codecs. These codecs exploit redundancies along both spatial and temporal dimensions, but to achieve higher compression, they also tend to smooth the original signal, thus removing most of the high-frequency components, and valu-

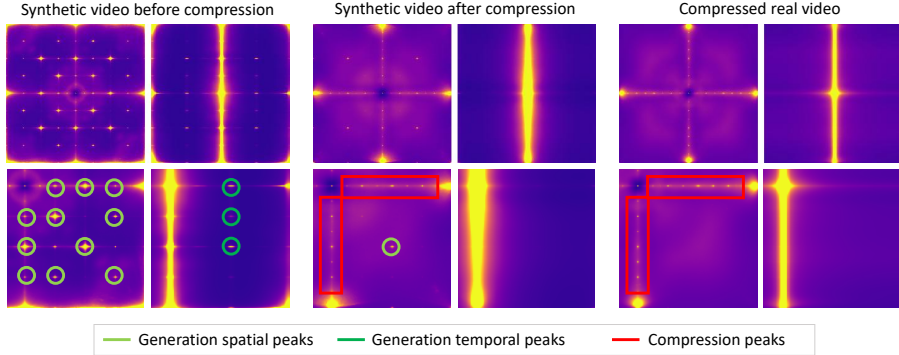


Figure 5.4. Top: Spatial and temporal-spatial power spectra of OpenSora-Plan videos before and after compression, compared to those computed from a real (compressed) video. Bottom: close up of the power spectra presented on the top. Fourier-domain peaks due to video synthesis (forensic artifacts) are highlighted by circles. Peaks originated by compression are highlighted by red boxes. We can notice that after compression most of the peaks are reduced and compression traces (peaks concentrated along the horizontal and vertical directions) are visible in synthetic videos similar to real ones.

able forensic artifacts along with them. Indeed, seeing Fig. 5.4, we can note that, after compression, most peaks vanish from the spatial spectra (left) and all of them from the temporal spectra (right). In addition, we can observe that compression, due to its block-wise processing, also introduces some strong peaks of its own, both in synthetic and real videos. This suggests that looking at vertical and horizontal directions could easily trick a detector into wrong decisions. Instead we can observe that the peak along the diagonal direction is still present, and can provide a more meaningful and robust cue. Notably all synthetic generators exhibit such diagonal peaks which survive even after compression. In particular, in Fig. 5.5, we show the close up of spatial and temporal-spatial power spectra before (top) and after (bottom) compression for additional synthetic generators. In all cases, we observe strong forensic artifacts (even for NOVA that relies on an autoregressive model) that are significantly reduced after compression. However, artifacts remain visible, particularly along the diagonal directions. It is also worth noting that temporal artifacts are less prominent, and the peaks along the vertical and horizontal directions tend to

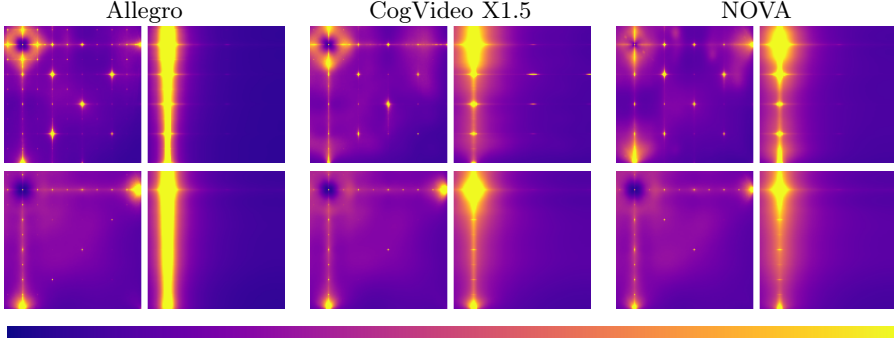


Figure 5.5. For each generator (Allegro, CogVideo X1.5, NOVA), we show the close up of its spatial power spectrum $S_{yx}(u, v)$ (left), and the close up of its temporal-spatial power spectra $S_{yt}(u, w)$ (right), both before compression (top) and after compression (bottom),

overlap with those introduced by the compression algorithm. This observation motivates our focus on these artifacts, that are both shared across different generators and are robust to compression.

To further support our conjecture that the mid-high frequencies along the diagonal directions are more discriminative, we compare real and reconstructed videos in the frequency domain. Reconstructed videos are obtained with the same autoencoder used during generation (as described in the following Section 5.2.2), hence they share the same content as real videos, but embed the traces related to the generation architecture. Such traces can be highlighted by computing the following distance:

$$d(u, v) = \frac{1}{N} \sum_{i=1}^N \frac{\sum_w |X_i(u, v, w) - \hat{X}_i(u, v, w)|^2}{\sum_w |X_i(u, v, w)|^2} \quad (5.5)$$

where $X_i(u, v, w)$ and $\hat{X}_i(u, v, w)$ are the 3d-Fourier transforms of i -th real video and its reconstruction, respectively, while N is the number of videos, that is equal to 100 in this experiment. In this comparison, we take into account the fact that small differences in low-energy regions of the spectrum are more significant than the same differences in high-energy regions. Therefore, we normalize the sum of squared differences by the

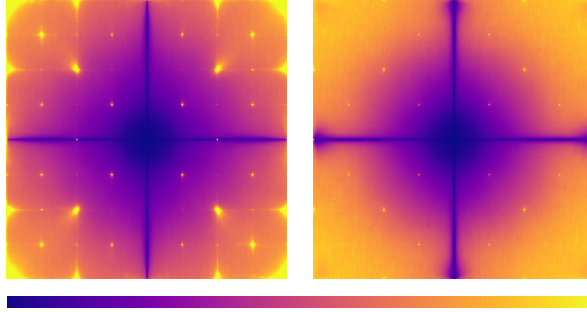


Figure 5.6. Average distance in the frequency domain as defined in eq.5.5, before compression (left) and after compression (right). The distance is computed between the real and the reconstructed video generated by the autoencoder of Pyramid Flow [70]. In this way we can show the artifacts of the model that are not influenced by the semantic content. The distance is nearly zero at the low frequencies and along the vertical and horizontal directions, much larger at mid-high frequencies along diagonal directions, even when videos are compressed.

energy of the real data at each spatial frequency. We show the average distance $d(u, v)$ on the left side of Fig.5.6, while the right side displays the same quantity evaluated after compressing the reconstructed videos using the same codec as for the real ones. Before compression (Fig.5.6, left) the reconstructed video shows a significant deviation from the real video at the mid-high frequencies, while horizontal and vertical frequencies appear to be less useful for detection. After compression (Fig. 5.6, right), the distances are reduced, but the diagonal mid-high frequency components are still discriminative.

5.2.2 Training strategy

The analysis described above has highlighted the need to focus on artifacts in the mid-high frequency range along the diagonal directions. In addition to this, as we have shown in Chapter 3, for Diffusion-based generated images diagonal frequencies were found to be more difficult to synthesize than vertical and horizontal ones, which further supports our conjecture. To implement our training strategy, we include two forms of augmentation: the addition of simulated fake videos in the training dataset by fingerprints

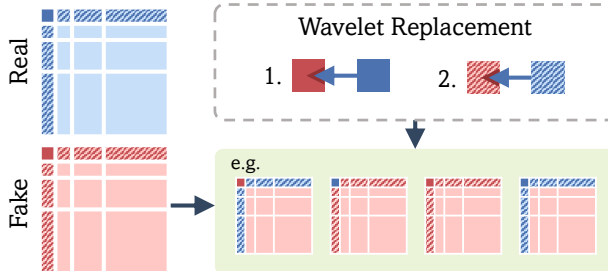


Figure 5.7. For each real video, four corresponding fake versions are generated: one with all fake subbands and others with selective replacements, such as the baseband or specific subbands substituted with their real counterparts. Notably, diagonal mid-to-high frequency subbands are never replaced: this is to teach the detector to do without the traces brought by one or the other low-frequency subbands and hence focus on such traces.

injection and the inclusion of an augmentation strategy during training based on the wavelet transform.

Artificial injection of forensic cues.

To encourage the detector to look at low-level artifacts, we add fake videos in the training dataset that are visually indistinguishable from real ones, but embed the traces related to the generation architecture. To achieve this, real videos are reconstructed with the same autoencoder used during generation. This idea is not new, and is actually gaining ground lately. Several works have simulated forensic artifacts to train the detectors both for GAN-based images [165, 69] and for diffusion models [116, 62]. Building fakes in this way also removes the semantic content bias and encourages the detector to identify intrinsic artifacts and patterns related to the generation process versus relying on differences in the semantic content. We adopt the Variational Autoencoder (VAE) used by Pyramid Flow [70]. This is a 3D convolutional neural network trained on the WebVid-10M dataset [6] with a latent space that compresses videos both spatially and temporally using a downsampling ratio of $8 \times 8 \times 8$. Its structure is similar to MAGVIT-v2 [157], incorporating causal convolution in time and pixel shuffle operations for decoder upsampling.

Augmentation by wavelet-band replacement.

We propose a new form of augmentation that is based on the wavelet decomposition. During training, for fake videos we replace the low-frequency bands (plus those along the horizontal and vertical directions) with those of the real counterpart. In this way, the detector is forced to base the decision on the mid-high frequency diagonal components. The use of wavelet decomposition helps to easily conduct such replacement, since this transform splits the signal into multiple frequency bands preserving spatial relationships within each band (see Fig.5.2). The augmentation process is summarized in Fig.5.7. First, we compute the Wavelet Transform. Then the baseband, the vertical and horizontal frequency bands of the fake frame are substituted with those of the real counterpart. Finally, the inverse Wavelet Transform is applied. We adopt the Fully Separable Wavelet Transform (FSWT), which offers an anisotropic representation that more effectively localizes frequencies relative to horizontal and vertical discontinuities compared to standard wavelets [145, 109], which is a property particularly useful to our task. We leverage the Haar wavelet transform [112] with three decomposition levels, resulting in the image being split into 16 frequency bands. The two-dimensional FSWT can be implemented by applying the multilevel decomposition separately along the rows and columns of each frame of the video. At each decomposition level i , the wavelet transform recursively splits the lowest-frequency band along each dimension into two sub-bands. This is achieved using two convolution operations with a stride of 2:

$$x_i^L = x_{i-1}^L \otimes_s h_L \quad x_i^H = x_{i-1}^L \otimes_s h_H, \quad (5.6)$$

where $\otimes_s$ denotes the stride convolution operation, h_L and h_H are the kernels of two convolutions. At each level i , the low frequency band, x_i^L , represents a low-resolution approximation of the input initialized with the original data at the first iteration. The high frequency band, x_i^H , captures detail information initially present in the higher resolution data. In a similar way, the inverse wavelet decomposition is based on transposed convolutions:

$$x_{i-1}^L = x_i^L \otimes^T h_L + x_i^H \otimes^T h_H, \quad (5.7)$$

where $\otimes^T$ denotes the transposed convolution operation with a stride of 2. Wavelet decomposition also offers computational advantages, as its

structure is well-suited for efficient implementation on GPUs and also has a complexity of $O(n)$ less than the $O(n \cdot \log(n))$ of the Fourier transform. The two kernels for the Haar wavelet are: $h_L = \frac{1}{\sqrt{2}}[1 \ 1]$ $h_H = \frac{1}{\sqrt{2}}[-1 \ 1]$.

5.3 Results

5.3.1 Experimental setup

Architecture.

As already discussed previously, we do not propose a new architecture and leverage the power of large pretrained models, that have already shown their potential for the detection of synthetic media in [104, 91, 132, 24] and in Chapter 4. In particular, we consider a DINOv2 model with registers [106, 36] trained end-to-end and frame-by-frame. The final decision is made by averaging the logit scores of 64 frames.

Training data.

To test for the generalization ability of our method, we train the model using one single synthetic generator. This is a common practice in forensics, reflecting the realistic scenario where generative architectures are unknown at test time. We leverage the Pyramid Flow model since it can generate synthetic videos with no clear visible errors and it was possible to regenerate artificial synthetic data through reconstruction (see Section 5.2.2). Real videos come from the Panda70M dataset [26]. Overall, training / validation datasets include 4,200 / 900 videos. Pairs of real and generated videos are carefully constructed using identical prompts to reduce content bias. To avoid format bias, all generated videos are compressed using the same codec as the real videos, i.e. H.264, as suggested in [61].

Test data.

We benchmark our model on the publicly available GenVideo dataset [24] which, includes videos generated by several models for a total of around 20k real and fake videos. We decided not to use the Diffusion Video Forensics (DVF) dataset [132], since real videos are compressed using MPEG-4 Part2, while fake videos are mostly compressed using H.264. This difference in format introduces a bias that can be exploited if the detector is not

properly trained, leading to an incorrect performance evaluation [61]. We also test on more recent models of the dataset we have created that we have described in Section 5.1. For this reason, it represents an excellent test of a detector’s ability to generalize to different generation strategies. Also for the test data, real and generated video pairs share identical prompts and all synthetic videos are compressed with H.264 (same codec as for real videos) to avoid any format bias.

Metrics.

Following previous works, we evaluate performance using the Area Under the Curve (AUC), which does not need to set a decision threshold, and balanced accuracy, which accounts for both false alarms and missed detections. For balanced accuracy, the decision threshold is set at 0.5. We also report the probability of detection at a 5% false alarm rate ($\text{Pd}@5\%$), the balanced Negative Log-Likelihood (NLL) [114], and binary Expected Calibration Error (ECE) [107]. The former provides insight into the false alarm behavior, which is critical in forensic applications. The latter two measures are calibration metrics and indicate how much the model’s predicted probability scores differ from the true probabilities and quantifies the generalization ability of the model.

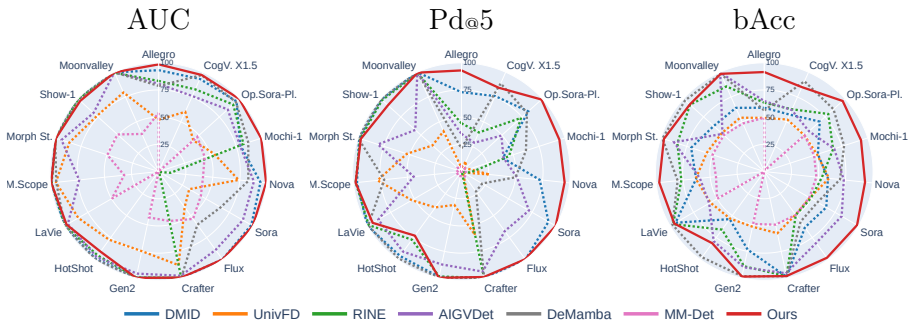


Figure 5.8. Comparison with SoTA methods on 16 generative models across different evaluation metrics.

5.3.2 Comparison with the State-of-The-Art

We compare our proposed method against several SoTA approaches. Among these, DMID [31], UnivFD [104], RINE [79], FreqNet [136], and FIRE [29] were originally developed for the detection of synthetic images. Others, AIGVDet [5], DeMamba [24], and MM-Det [132], were specifically designed for detecting AI-generated videos. To ensure a fair comparison, we include only methods with publicly available code and/or pre-trained models. All models are trained on the datasets used in their respective original publications: ProGAN for UnivFD and FreqNet, Latent Diffusion for DMID and RINE, ADM for FIRE, Stable Video Diffusion for MM-Det, Moonvalley for AIGVDet, and videos from ten generators (ZeroScope, I2VGen-XL, Stable Diffusion, Stable Video Diffusion, VideoCrafter, Pika, DynamiCrafter, SEINE, Latte, OpenSora) for DeMamba.

Analysis with methods trained on their original datasets.

Balanced accuracy results across all tested generators are reported in Table 5.1, while AUC and Pd@5 scores are presented more concisely in Fig.5.8. We begin our analysis with the average results, shown in the last column of Tab.5.1. The proposed method achieves an average balanced accuracy of 94.3%, outperforming DeMamba, the second-best method, by ten percentage points. As expected, methods originally developed for image detection are not competitive in this setting, like UnivFD, trained on older ProGAN images, and FreqNet. However, also the video-oriented AIGVDet lags nearly 20 percentage points behind our proposed method. The per-

Table 5.1. Comparison in terms of balanced accuracy with SoTA methods trained on their original datasets.

Method	Recent Generators (2024-25 years)							GenVideo (2022-23 years)								
	Alleg-ro	CogV. XI.5	OSora Plan	Mochi	Nova	Sora	Flux	Craft-er	Gen2	Hot Shot	LaVie	Model Scope	Morph Stu.	Show1	Moon valley	AVG
DMID	58.7	57.3	68.7	51.8	58.8	65.8	63.5	99.0	73.4	54.1	94.2	62.9	70.7	62.6	64.1	67.0
UnivFD	49.3	52.3	49.7	51.0	59.8	48.3	49.0	58.0	51.9	53.5	57.2	63.3	55.6	52.7	54.2	53.7
RINE	63.3	62.7	79.2	66.5	53.2	49.3	49.3	98.3	89.1	66.0	96.7	76.6	84.0	91.8	85.7	74.1
FreqNet	66.8	45.0	68.8	52.2	40.5	54.2	43.5	47.9	37.2	27.7	47.7	38.0	46.5	46.2	50.9	47.5
FIRE	51.5	49.8	51.7	51.5	51.5	51.3	51.3	49.4	49.1	47.6	48.6	46.6	48.8	47.3	48.9	49.7
AIGVDet	64.8	62.0	55.3	67.7	73.5	82.2	81.2	95.6	90.0	79.1	58.5	59.5	87.2	51.7	99.6	73.9
DeMamba	56.2	90.8	85.7	77.2	70.0	56.5	56.8	98.4	98.3	97.4	97.6	80.1	98.4	97.9	98.3	84.0
MM-Det	49.5	1.3	49.8	48.2	48.2	47.8	49.3	49.9	49.3	0.1	50.1	44.6	50.0	50.1	48.4	42.4
Ours	91.5	85.0	97.0	93.5	93.2	98.5	98.0	98.3	98.8	81.4	95.5	97.1	97.0	92.1	98.4	94.3

formance of MM-Det, falling below random guessing, may seem especially surprising. However, this outcome is primarily due to the biased DVF (Diffusion Video Forensics) dataset used for its original training, which leads to a misleading assessment. Turning to results on the individual generators, a clear divide can be observed between the older ones, included in the GenVideo dataset, and the more recent ones added here. Several detectors exhibit a very good performance on GenVideo (DeMamba is the top detector on those data), which drops substantially on the newer generators, highlighting a limited generalization capability. In contrast, the proposed method performs consistently well across all generators, with isolated dips on CogVideo X1.5 and on Hot Shot.

Analysis with methods re-trained on Pyramid Flow.

To demonstrate that the observed performance gap is not due to the inclusion of the recent Pyramid Flow model in training, we retrained all methods using videos generated by Pyramid Flow. The resulting accuracy metrics are reported in Table 5.2. Under this new training condition, we observe significant changes in performance. However even with the re-training, our method is still better than SoTA methods by a large margin. Results become generally more consistent across both older and newer generators, and in some cases, most notably for MM-Det, substantially better than before. The most striking outcome, however, is that image-oriented methods outperform video-oriented ones and by a notable margin. This is an unexpected result and highlights that the video-based detectors are not able to fully exploit the advantage of operating in 3D. We can observe that other frequency-based approaches, such as FreqNet and FIRE, struggle to achieve good results even if trained on our same dataset. This highlights the importance to properly take into account the influence of the video codec in the proposed strategy. It is also important to note that a few abrupt drops in performance (e.g., AIGVDet falling from 99.6% to 50.0% on Moon Valley, and DeMamba from 97.4% to 62.0% on Hot Shot) have a significant impact on the averages. These anomalies may be attributed to biases introduced during training, however this also requires further investigation.

Our analysis so far has been only focused on the balanced accuracy, we extend it also to other metrics. In Tables 5.3, 5.4, 5.5 and 5.6, we

Table 5.2. Comparison in terms of balanced accuracy with SoTA methods re-trained using Pyramid Flow (*).

Method	Recent Generators (2024-25 years)							GenVideo (2022-23 years)								AVG
	Alleg-ro	CogV. X1.5	OSora Plan	Mochi1	Nova	Sora	Flux	Craft-er	Gen2	Hot Shot	LaVie	Model Scope	Morph Stu.	Show1	Moon valley	
DMID*	77.3	72.0	91.2	75.8	82.0	89.2	89.7	93.7	94.6	75.1	88.3	76.9	81.0	73.6	97.0	83.8
UnivFD*	76.3	63.8	74.5	71.5	75.2	88.7	80.3	90.4	92.8	66.6	84.9	80.6	85.7	79.4	95.0	80.4
RINE*	90.7	64.3	89.8	77.3	83.3	81.8	80.8	94.5	95.2	76.6	89.6	87.2	89.3	80.4	97.2	85.2
FreqNet*	84.3	65.5	86.3	77.5	69.0	64.5	56.5	52.1	58.5	42.4	64.7	42.8	56.5	63.0	65.0	63.3
FIRE*	75.3	65.8	63.3	72.8	74.8	64.0	73.7	83.2	88.0	73.8	75.6	75.0	81.4	76.0	93.1	75.7
AIGVDet*	75.7	60.8	50.8	73.3	70.7	68.5	67.5	87.7	85.5	54.0	50.4	54.1	78.5	50.0	50.0	65.2
DeMamba*	83.3	61.3	91.8	79.2	68.7	73.0	76.0	87.5	92.2	62.0	79.6	68.6	83.8	72.6	92.4	78.1
MM-Det*	56.8	49.8	71.5	70.5	70.5	76.7	72.8	84.5	87.2	40.5	85.8	63.3	76.5	80.8	87.0	71.6
Ours	91.5	85.0	97.0	93.5	93.2	98.5	98.0	98.3	98.8	81.4	95.5	97.1	97.0	92.1	98.4	94.3

Table 5.3. Comparison in terms of AUC with SoTA methods both trained on their original datasets and re-trained using Pyramid Flow (*). Bold underlines the best performance for each column with a margin of 1%.

AUC ↑	Recent Generators (2024-25 years)							GenVideo (2022-23 years)								AVG
	Alleg-ro	CogV. X1.5	OSora Plan	Mochi1	Nova	Sora	Flux	Craft-er	Gen2	Hot Shot	LaVie	Model Scope	Morph Stu.	Show1	Moon valley	
DMID	93.2	93.2	96.4	78.8	94.1	98.3	99.8	100.	99.7	98.9	99.9	99.2	99.6	99.7	99.9	96.7
UnivFD	47.7	60.1	42.8	53.3	73.9	32.4	40.6	88.0	74.8	77.5	85.3	95.0	86.3	76.4	79.6	67.6
RINE	83.3	82.7	92.2	82.0	11.0	0.7	1.0	99.9	99.5	95.4	99.5	99.0	99.3	99.6	99.4	76.3
FreqNet	75.1	43.1	80.5	52.8	37.3	56.9	39.8	57.4	30.8	23.2	55.7	42.5	54.2	48.7	58.1	50.4
FIRE	56.0	51.2	93.0	85.8	79.7	40.8	46.8	36.9	23.0	35.1	62.0	25.7	24.3	40.1	18.6	47.9
DMID*	99.5	98.7	99.9	96.6	99.4	99.3	99.8	98.8	99.8	93.3	97.6	96.9	97.9	97.2	99.6	98.3
UnivFD*	93.5	83.6	92.8	89.3	93.5	97.3	94.4	97.2	98.0	81.7	94.0	91.7	94.6	90.6	98.9	92.7
RINE*	99.7	92.3	99.1	96.8	98.6	97.5	97.1	98.6	98.9	90.7	97.5	97.0	97.5	93.8	99.7	97.0
FreqNet*	91.6	72.3	94.3	84.3	79.2	73.7	69.6	53.5	63.6	41.1	76.0	42.4	59.8	69.2	71.8	69.5
FIRE*	83.9	68.6	66.4	79.8	83.4	70.3	80.8	92.2	95.0	84.3	85.1	84.4	91.0	83.2	97.4	83.0
AIGVDet	80.7	76.0	85.4	78.5	88.0	89.4	89.1	98.0	95.9	97.0	96.4	73.5	94.0	80.6	99.9	88.2
DeMamba	78.7	97.1	95.6	86.6	82.8	65.8	60.5	100.	100.	99.6	99.7	94.7	100.	99.9	100.	90.7
MM-Det	53.9	0.0	47.7	44.3	41.1	44.8	53.8	46.5	43.1	0.0	49.7	31.9	49.6	51.0	37.9	39.7
AIGVDet*	92.3	91.2	42.4	91.4	94.8	88.7	91.4	99.9	100.	98.6	95.0	99.3	99.9	88.2	100.	91.5
DeMamba*	97.9	88.5	99.2	93.8	90.8	94.2	95.3	94.7	97.2	70.0	92.3	78.2	92.9	85.7	96.3	91.1
MM-Det*	89.2	0.0	97.2	96.5	97.3	96.6	97.0	93.3	95.4	0.0	94.3	69.6	84.5	89.1	95.0	79.7
Ours	98.6	97.2	99.8	99.1	99.3	99.9	99.9	99.8	100.	91.0	97.7	99.4	99.3	97.1	99.9	98.5

show the performance using the other four metrics: AUC, Pd@5, NLL and ECE. AUC is generally higher than the balanced Accuracy for most of the methods, which means that a fixed threshold set to 0.5 is not the best choice and proper calibration is needed. However, our approach is instead able to achieve good performance consistently over all the metrics, in particular it gains an improvement on average of around 53% and 68% for NLL and ECE, respectively.

Table 5.4. Comparison in terms of Pd@5 with SoTA methods both trained on their original datasets and re-trained using Pyramid Flow (*). Bold underlines the best performance for each column with a margin of 1%.

Pd@5 $\uparrow$	Recent Generators (2024-25 years)							GenVideo (2022-23 years)								AVG
	Alleg-ro	CogV-X1.5	OSora Plan	Mochi1	Nova	Sora	Flux	Craft-er	Gen2	Hot Shot	LaVie	Model Scope	Morph Stu.	Show1	Moon valley	
DMID	73.0	76.0	82.7	40.0	72.3	92.3	99.9	100.	99.0	96.0	99.9	96.6	99.0	99.1	100.	88.4
UnivFD	3.7	10.3	4.3	7.3	26.0	1.7	3.0	59.6	31.4	40.7	52.7	77.6	52.9	35.9	40.6	29.8
RINE	44.3	39.3	73.7	46.0	8.3	0.7	0.7	99.9	98.7	77.4	98.5	97.7	98.0	98.1	99.0	65.4
FreqNet	16.7	4.8	22.2	5.6	0.9	7.5	3.0	14.8	3.6	4.4	12.8	8.1	12.7	9.6	12.4	9.3
FIRE	6.3	5.7	84.7	70.3	62.3	1.3	2.0	0.0	0.0	0.0	1.6	0.6	0.0	0.0	0.0	15.7
DMID*	97.3	94.3	99.0	87.0	97.3	98.0	98.7	95.9	99.7	81.0	90.9	88.9	90.3	85.9	99.2	93.6
UnivFD*	67.3	45.7	67.7	61.0	68.3	86.3	75.3	85.8	90.9	38.7	75.2	66.4	76.4	64.4	94.9	71.0
RINE*	98.3	67.3	95.7	84.3	93.3	89.0	87.0	94.7	95.4	64.1	87.6	84.9	86.4	73.0	98.6	86.7
FreqNet*	56.7	22.7	71.7	33.7	18.3	11.3	5.7	6.7	13.2	6.6	33.3	3.7	11.0	16.1	13.7	21.6
FIRE*	45.7	26.3	28.7	50.7	51.3	22.7	44.0	67.3	70.0	45.9	48.9	38.9	59.0	49.3	88.5	49.1
AIGVDet	34.7	29.0	49.3	40.7	51.3	71.7	67.7	94.1	87.3	91.1	88.3	44.1	80.7	57.1	99.8	65.8
DeMamba	21.3	88.7	81.0	62.3	50.0	22.0	22.7	100.	99.9	98.7	99.0	71.4	100.	99.4	100.	74.4
MM-Det	5.6	0.0	4.6	4.3	3.8	4.3	5.5	4.6	4.3	0.0	5.0	3.1	5.0	5.1	3.8	3.9
AIGVDet*	71.7	59.0	23.3	68.7	72.0	64.7	65.3	99.4	99.9	94.4	75.8	96.1	99.7	43.9	100.	75.6
DeMamba*	94.0	59.7	97.3	82.0	69.7	78.7	85.7	81.8	90.1	31.1	69.9	45.3	75.3	56.0	89.3	73.7
MM-Det*	52.3	0.0	85.7	83.7	84.7	85.7	82.7	72.2	76.2	0.0	74.1	25.1	39.1	52.6	76.7	59.4
Ours	93.0	84.3	99.0	94.7	95.3	100.	99.7	99.1	99.9	72.9	94.3	97.9	97.3	91.0	99.5	94.5

Table 5.5. Comparison in terms of NLL with SoTA methods both trained on their original datasets and re-trained using Pyramid Flow (*). Bold underlines the best performance for each column with a margin of 1%.

NLL $\downarrow$	Recent Generators (2024-25 years)							GenVideo (2022-23 years)								AVG
	Alleg-ro	CogV-X1.5	OSora Plan	Mochi1	Nova	Sora	Flux	Craft-er	Gen2	Hot Shot	LaVie	Model Scope	Morph Stu.	Show1	Moon valley	
DMID	2.18	2.14	1.53	3.97	2.10	1.10	0.73	0.04	0.92	2.09	0.20	1.82	1.17	1.10	1.08	1.48
UnivFD	2.76	2.16	3.02	2.50	1.45	3.62	3.13	1.76	2.67	2.44	1.96	1.21	1.93	2.56	2.34	2.37
RINE	0.59	0.60	0.45	0.59	6.62	17.2	15.6	0.14	0.30	0.50	0.16	0.36	0.32	0.26	0.34	2.93
FreqNet	10.6	15.6	10.2	13.6	16.8	13.7	16.8	30.5	33.0	35.5	30.6	33.6	30.7	30.7	29.5	23.4
FIRE	3.95	4.01	3.94	3.94	3.95	3.95	3.95	4.84	4.84	5.00	4.86	5.01	4.85	4.96	4.84	4.46
DMID*	0.41	0.51	0.19	0.57	0.35	0.26	0.22	0.20	0.14	0.69	0.32	0.50	0.41	0.54	0.14	0.36
UnivFD*	0.53	1.00	0.56	0.71	0.54	0.27	0.46	0.22	0.19	0.88	0.37	0.48	0.35	0.52	0.14	0.48
RINE*	0.24	2.61	0.42	1.27	0.68	0.95	1.11	0.24	0.19	1.46	0.38	0.45	0.38	0.97	0.07	0.76
FreqNet*	0.65	1.80	0.61	1.02	1.17	1.52	1.66	2.93	2.34	4.66	2.35	4.31	2.76	2.27	1.96	2.13
FIRE*	0.51	0.77	0.82	0.62	0.54	0.70	0.56	0.39	0.29	0.66	0.63	0.65	0.43	0.68	0.20	0.56
AIGVDet	0.95	1.16	0.84	1.07	0.67	0.70	0.69	0.31	0.46	0.48	0.52	1.35	0.46	1.12	0.32	0.74
DeMamba	1.98	0.33	0.48	1.12	1.41	2.57	2.78	0.05	0.05	0.08	0.07	0.64	0.05	0.06	0.05	0.78
MM-Det	38.4	88.3	38.4	38.6	38.5	38.6	38.4	48.3	48.3	98.3	48.3	48.8	48.3	48.3	48.4	50.4
AIGVDet*	0.60	0.72	3.57	0.64	0.56	0.77	0.69	0.44	0.30	2.72	4.38	1.08	0.56	7.55	0.35	1.66
DeMamba*	0.49	1.66	0.22	0.85	1.31	1.00	0.83	0.45	0.26	1.93	0.67	1.47	0.58	1.08	0.30	0.87
MM-Det*	3.60	50.0	1.50	1.71	1.62	1.46	1.55	0.74	0.63	50.5	0.69	2.13	1.13	0.89	0.65	7.92
Ours	0.28	0.53	0.07	0.19	0.18	0.04	0.05	0.06	0.05	0.53	0.16	0.09	0.09	0.22	0.05	0.17

Table 5.6. Comparison in terms of ECE with SoTA methods both trained on their original datasets and re-trained using Pyramid Flow (*). Bold underlines the best performance for each column with a margin of 1%.

ECE ↓	Recent Generators (2024-25 years)							GenVideo (2022-23 years)								AVG
	Alleg-ro	CogV.X1.5	OSora Plan	Mochi1	Nova	Sora	Flux	Craft-er	Gen2	Hot Shot	LaVie	Model Scope	Morph Stu.	Show1	Moon valley	
DMID	.406	.406	.312	.471	.399	.334	.329	.026	.274	.436	.068	.362	.289	.347	.341	.320
UnivFD	.479	.419	.481	.438	.324	.501	.489	.398	.464	.445	.410	.348	.420	.458	.440	.434
RINE	.152	.149	.167	.134	.536	.623	.624	.107	.202	.229	.107	.230	.208	.173	.230	.258
FreqNet	.238	.449	.217	.381	.497	.361	.470	.180	.286	.378	.184	.281	.194	.198	.153	.298
FIRE	.483	.492	.482	.482	.482	.484	.482	.503	.505	.520	.509	.529	.507	.522	.506	.499
DMID*	.227	.254	.125	.237	.189	.137	.140	.078	.086	.241	.126	.220	.193	.248	.080	.172
UnivFD*	.195	.299	.200	.229	.206	.094	.162	.039	.024	.252	.096	.142	.091	.151	.024	.147
RINE*	.098	.339	.110	.225	.172	.190	.198	.032	.023	.202	.075	.103	.079	.162	.010	.134
FreqNet*	.108	.296	.106	.170	.244	.289	.361	.404	.347	.512	.312	.499	.368	.309	.289	.308
FIRE*	.086	.156	.171	.129	.100	.144	.091	.069	.051	.158	.141	.168	.090	.142	.044	.116
AIGVDet	.287	.305	.264	.283	.233	.204	.206	.157	.192	.239	.230	.293	.150	.277	.230	.237
DeMamba	.410	.079	.120	.211	.274	.415	.405	.020	.019	.012	.014	.171	.020	.014	.019	.147
MM-Det	.129	.612	.126	.140	.139	.143	.129	.019	.025	.518	.018	.070	.019	.018	.034	.143
AIGVDet*	.230	.292	.411	.233	.243	.266	.258	.181	.162	.420	.442	.305	.248	.492	.250	.296
DeMamba*	.159	.364	.074	.193	.291	.249	.224	.092	.047	.346	.167	.278	.127	.236	.050	.193
MM-Det*	.421	.502	.272	.284	.292	.239	.268	.111	.089	.593	.101	.305	.170	.134	.091	.258
Ours	.076	.140	.023	.056	.058	.014	.012	.025	.029	.145	.016	.013	.012	.043	.026	.046

5.4 Ablation study

Here we present our ablation study, where we analyze our method with different augmentations, backbones and training data.

Influence of augmentation.

Table 5.7 shows the impact of different augmentation strategies, using always the same backbone, DINOv2 [106, 36] trained in an end-to-end manner. In the first row, only a standard form of augmentation is used in training, including compression, blurring and resizing, operations typically carried out in forensics applications. Then, starting with the second row, reconstructed fake videos paired with real ones are included (marked by a check in the *+Rec.* column), and new forms of augmentation are added: MixUp [162], CutMix [160], and finally our wavelet-band replacement (WaveRep). On average, across all generators, it appears that the inclusion of artificial fake videos in training (second row) ensures a significant gain in performance under all metrics, AUC and bAcc in the top part of the table, and NLL and ECE in the bottom. Then, the computer vision-oriented forms of augmentation, MixUp and CutMix, provide very

Table 5.7. Influence of the augmentation.

		AUC $\uparrow$ / bAcc $\uparrow$				
+Rec.	augment.	Allegro	CogVideo X1.5	Mochi-1	OpenSora-Plan	AVG
	-	94.7 / 68.2	91.3 / 60.2	94.9 / 75.5	98.0 / 80.0	94.7 / 71.0
✓	-	97.3 / 87.8	95.0 / 78.2	98.9 / 90.8	99.4 / 95.8	97.6 / 88.2
✓	MixUp	96.7 / 84.7	94.8 / 79.0	98.6 / 87.8	99.3 / 94.3	97.3 / 86.5
✓	CutMix	97.8 / 87.7	95.8 / 80.3	98.9 / 92.3	99.3 / 94.3	97.9 / 88.7
✓	WaveRep	98.6 / 91.5	97.2 / 85.0	99.1 / 93.5	99.8 / 97.0	98.7 / 91.8
		NLL $\downarrow$ / ECE $\downarrow$				
	-	1.95 / .309	2.73 / .389	1.58 / .244	1.03 / .196	1.82 / .284
✓	-	0.50 / .118	0.88 / .198	0.28 / .084	0.15 / .038	0.45 / .109
✓	MixUp	0.35 / .106	0.71 / .197	0.24 / .094	0.12 / .037	0.35 / .109
✓	CutMix	0.38 / .120	0.63 / .185	0.22 / .073	0.15 / .051	0.35 / .107
✓	WaveRep	0.28 / .076	0.53 / .140	0.19 / .056	0.07 / .023	0.27 / .074

limited additional improvements, if any. On the contrary, the proposed WaveRep provides a further performance boost, driving AUC to almost 99% and bAcc past 90%.

Influence of the backbone.

Next, we fix the training setup, including artificial fake videos and adding the WaveRep augmentation on top of the basic configuration, and analyze performance across different backbone architectures, with results reported in Table 5.8. The best results for each column (considering a margin of 1%) are highlighted in bold. Focusing on the average performance metrics, we observe that the end-to-end training (e2e) always performs significantly better than linear probing (LP) versions of both DINOv2 and Clip [115]. In the former the entire network is fine-tuned, while only a linear layer on the final features is learned for the latter. We hypothesize that the semantics-oriented pretraining of these models is not well suited to our forensic application, highlighting the need for a deeper adaptation. Excluding the LP variants, DINOv2 stands out as the clearly preferable option across all metrics. In general, the variability in AUC across the variants is more limited, likely because scores are already very close to 100%, but DINOv2 maintains a consistent advantage also compared to a

Table 5.8. Influence of the backbone.

	AUC $\uparrow$ / bAcc $\uparrow$					
Network	Allegro	CogVideo X1.5	Mochi-1	OpenSora-Plan	AVG	
DINOv2 e2e	98.6 / 91.5	97.2 / 85.0	99.1 / 93.5	99.8 / 97.0	98.7 / 91.8	
DINOv2 LP	86.6 / 79.2	77.5 / 70.8	88.6 / 81.7	91.6 / 83.5	86.1 / 78.8	
Clip e2e	99.1 / 86.5	97.2 / 67.3	98.6 / 79.3	99.9 / 92.7	98.7 / 81.5	
Clip LP	91.8 / 70.8	89.3 / 67.5	87.8 / 69.8	94.1 / 76.5	90.7 / 71.2	
Hiera video	98.9 / 79.7	93.5 / 65.8	94.7 / 65.3	99.4 / 81.8	96.6 / 73.2	
	NLL $\downarrow$ / ECE $\downarrow$					
DINOv2 e2e	0.28 / .076	0.53 / .140	0.19 / .056	0.07 / .023	0.27 / .074	
DINOv2 LP	0.50 / .110	0.58 / .059	0.46 / .109	0.43 / .126	0.49 / .101	
Clip e2e	0.52 / .141	1.54 / .312	0.92 / .209	0.21 / .073	0.80 / .184	
Clip LP	0.54 / .207	0.60 / .227	0.60 / .215	0.45 / .173	0.55 / .205	
Hiera video	0.66 / .204	1.66 / .339	1.54 / .338	0.49 / .179	1.09 / .265	

Table 5.9. Performance by varying L (number of wavelet decomposition levels).

L	AUC $\uparrow$ / bAcc $\uparrow$					NLL $\downarrow$ / ECE $\downarrow$				
	Allegro	CogVideo X1.5	Mochi-1	OpenSora-Plan	AVG	Allegro	CogVideo X1.5	Mochi-1	OpenSora-Plan	AVG
2	98.6 / 90.0	96.5 / 81.7	98.9 / 91.5	99.8 / 96.0	98.4 / 89.8	0.29 / .094	0.60 / .174	0.24 / .079	0.08 / .031	0.30 / .095
3	98.6 / 91.5	97.2 / 85.0	99.1 / 93.5	99.8 / 97.0	98.7 / 91.8	0.28 / .076	0.53 / .140	0.19 / .056	0.07 / .023	0.27 / .074
4	97.6 / 88.8	95.7 / 81.5	98.2 / 89.3	99.1 / 94.5	97.7 / 88.5	0.39 / .102	0.67 / .168	0.31 / .091	0.18 / .045	0.39 / .101

spatial-temporal backbone based on Hiera [123], that works on clips of 16 frames, and the final decision is obtained by averaging the logits across all clips.

Ablation on the use of diagonal frequencies

To verify that the proposed detector leverages these mid-high frequencies to establish if a video is fake or not, we conducted a suitable toy experiment: at test time we modified the real videos and injected the mid-high frequencies of fully synthetic videos into them. We observe a performance drop on this test set. In particular, the True Negative Rate (TNR), which is the probability that real videos are correctly classified as real, goes from 98.7 to 0.2 (on average). This clearly shows that the detector inverts its decision.

Table 5.10. Performance by varying compression quality (CRF) and the number of wavelet decomposition levels (L).

AUC $\uparrow$ / bAcc $\uparrow$	$L=2$	$L=3$	$L=4$
CRF 20 (high quality)	97.8 / 83.8	97.3 / 84.5	96.6 / 83.2
CRF 24	96.8 / 77.1	96.0 / 78.3	95.2 / 78.2
CRF 28	93.9 / 67.9	92.5 / 69.1	91.1 / 68.5
CRF 32 (low quality)	88.2 / 59.5	85.2 / 59.2	82.6 / 59.3

Impact of the decomposition levels

In Table 5.9, we conducted a study to assess the impact of the number of decomposition levels of the wavelet transform implemented in our augmentation strategy. Specifically, we evaluated configurations with 2, 3, and 4 levels of decomposition. The results indicate that the choice of 3 decomposition levels yields the best performance across multiple metrics. The difference is more marked if we look at the Accuracy than the AUC. Even NLL and especially ECE show an advantage in this situation, which confirms that this choice helps to achieve more calibrated results.

Compression robustness.

We compressed the test dataset used in our ablation study with H.264 at different quality levels by varying the Constant Rate Factor (CRF). Table 5.10 reports results across decomposition levels, showing robustness to compression (AUC always above 80%). Accuracy drops under severe compression (around 60% at CRF 32) motivating a threshold calibration procedure for low-quality compressed videos (calibrating with only 10 compressed validation videos raises accuracy to 76%).

Training data.

In this paragraph, we evaluate the influence of training data. In Table 5.11, the proposed augmentation yields clear gains even with limited data (20% of the training set) and when training on a different generator (CogVideoX instead of Pyramid Flow). With only 20% of the data, augmentation still improves balanced accuracy (+3.2%). Likewise, when replacing Pyramid Flow with CogVideoX, we again observe a substantial boost from the proposed augmentation strategy (+6.3%).

Table 5.11. Performance by varying the type of generator and the size of the dataset during training.

		AUC $\uparrow$ / bAcc $\uparrow$			
training data	aug.	Allegro	Mochi-1	OpenSora-Plan	AVG
100% Pyr. Flow		94.7 / 68.2	94.9 / 75.5	98.0 / 80.0	95.9 / 74.6
100% Pyr. Flow	✓	98.6 / 91.5	99.1 / 93.5	99.8 / 97.0	99.1 / 94.0
20% Pyr. Flow		94.3 / 70.8	93.9 / 77.8	98.4 / 87.3	95.5 / 78.7
20% Pyr. Flow	✓	97.6 / 88.3	98.8 / 92.7	99.8 / 96.5	98.7 / 92.5
100% CogV. X1.5		90.7 / 69.3	90.7 / 72.0	94.9 / 78.0	92.1 / 73.1
100% CogV. X1.5	✓	98.5 / 88.8	97.6 / 85.2	99.2 / 92.8	98.4 / 88.9

Influence of number of generators in training.

In this section, we evaluate the influence of the number of generators included during training. More specifically, beyond Pyramid Flow, we also add CogVideo X1.5 (two generators) and another setting with four models by adding Allegro and OpenSora Plan. Note that for all these models we considered fake videos reconstructed with the same autoencoder used during generation. Results of our method with and without the wavelet-based augmentation on recent generative models are presented in Table 5.12, where we crossed-out the results for the generators included during training. We can observe a consistent improvement with the increase of the number of generators. Our proposed augmentation also guarantees a significant improvement especially when few models are considered, though even with four generators there is an increase of around 5% in terms of balanced accuracy. Finally, we want to highlight that our approach ranked first on the SAFE Synthetic Video Detection Challenge 2025 [124]. This challenge simulates a realistic scenario where no information about the test models are known in advance and videos vary in resolution, bitrate, frame rate, and length. This confirms the advantage of our training data and augmentation strategy to ensure good generalization ability.

Table 5.12. Performance by varying the number of generators during training in terms of AUC and balanced Accuracy. For each row, the results for the generators used in training are crossed-out. The average values do not include crossed-out results. We compare DINOv2 network with and without our approach.

		Recent Generators (2024-25 years)								
#	aug.	Pyramid Flow	CogV. X1.5	Allegro	OpenSora-Plan	Mochi-1	Nova	Sora	Flux	AVG
1		100 / 99.7	91.3 / 60.2	94.7 / 68.2	98.0 / 80.0	94.9 / 75.5	99.0 / 87.7	95.7 / 69.5	96.8 / 72.0	95.8 / 73.3
1	✓	100 / 99.3	97.2 / 85.0	98.6 / 91.5	99.8 / 97.0	99.1 / 93.5	99.3 / 93.2	99.9 / 98.5	99.9 / 98.0	99.1 / 93.8
2		99.9 / 98.5	99.8 / 98.3	94.9 / 80.5	98.7 / 94.7	96.5 / 86.5	99.4 / 96.5	94.4 / 77.5	96.8 / 88.2	96.8 / 87.3
2	✓	100 / 99.3	100 / 99.2	99.1 / 94.3	99.8 / 96.5	99.6 / 93.5	99.9 / 98.7	99.9 / 97.8	99.9 / 98.8	99.7 / 96.6
4		100 / 98.3	99.8 / 97.7	99.9 / 98.0	100 / 98.3	98.0 / 89.5	99.9 / 98.2	98.1 / 89.8	99.4 / 95.2	98.8 / 93.2
4	✓	100 / 99.2	100 / 98.7	100 / 99.2	100 / 99.2	99.8 / 97.3	100 / 99.0	100 / 99.0	100 / 99.2	99.9 / 98.6

5.5 Conclusions

We have proposed a novel augmentation strategy for the detection of AI-generated videos. We begin by identifying key forensic cues that are robust to compression artifacts and consistently present across different generative models. Our approach is motivated by the observation that vertical and horizontal frequency components are particularly susceptible to degradation due to block-based compression algorithms commonly employed in video codecs. Through extensive experiments, we demonstrate that this simple augmentation improves detection performance even if only one single generator is included during training. Our findings highlight the importance of working on the training paradigm and of injecting carefully curated domain specific knowledge, which can yield a greater benefit than the design of more complex detection architectures.

The method proposed incorporates a forensic-oriented augmentation strategy that encourages the detector to focus on the most relevant spatial traces. However, it is not explicitly designed to capture temporal artifacts. Even if our approach outperforms existing methods that leverage temporal information, we believe it is essential to systematically investigate and characterize the most discriminative forensic cues present in the temporal domain. A deeper understanding of such artifacts could lead to detectors that more effectively exploit spatial and temporal traces. Furthermore, our method was trained using fake videos generated by the Pyramid Flow model, selected for its ability to produce high-quality synthetic content

with minimal perceptible visual artifacts. This design choice allowed us to focus on subtle, low-level traces rather than easily detectable flaws. However, as new generative models emerge, potentially relying on entirely different synthesis paradigms, the current approach may struggle to generalize. Although the core principles may still be applicable, the method itself may require adaptation or retraining.

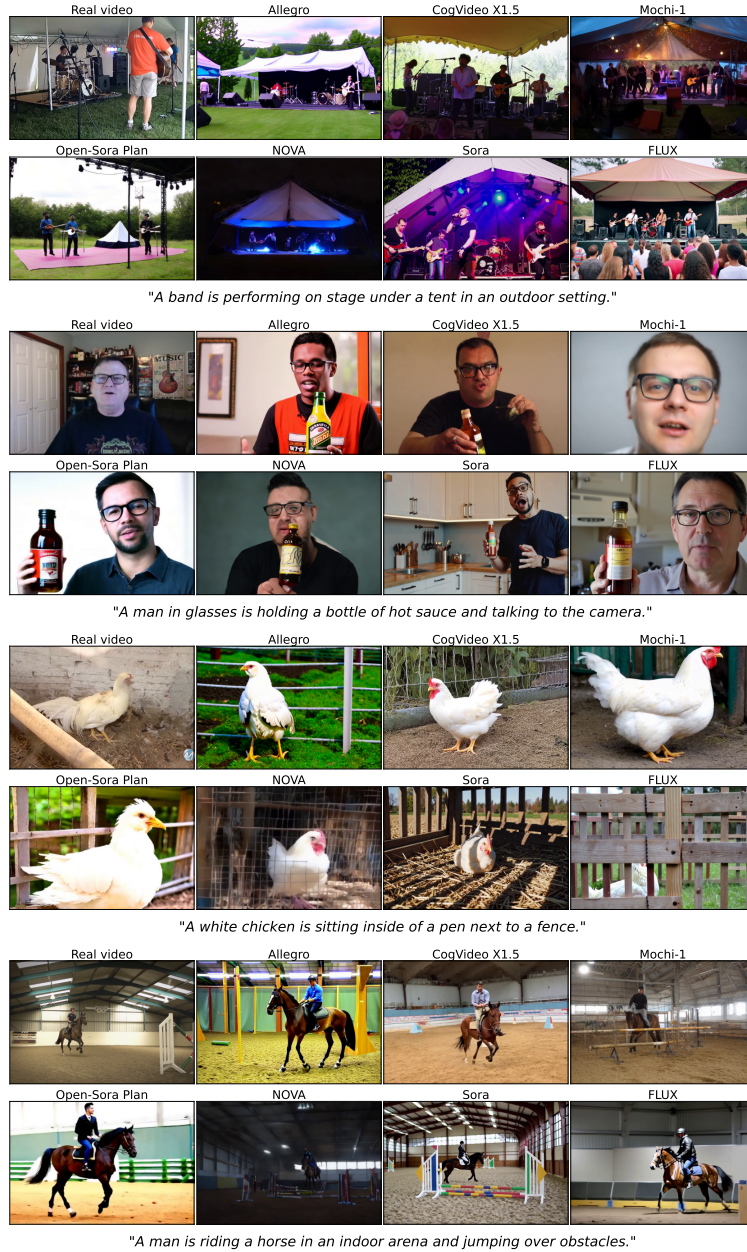


Figure 5.9. Some examples of videos from our dataset that we created with recent state of the art generators.



Figure 5.10. Some examples of videos from our dataset that we created with recent state of the art generators.

Conclusions

In this thesis, we have conducted a thorough analysis of various types of synthetic contents, both image and video, and we have proposed methods for their detection. The proposed approaches have moved the field forward and improved over the state-of-the-art providing detectors which are more robust and general. First, we conducted a detailed study of a large number of generators and the artifacts present in their generated images. We leverage second-order statistics to conduct an analysis both in the spatial domain, with the autocorrelation, and in the frequency domain, with the power spectra. We showed that even the most sophisticated generators still leave detectable artifacts in their images. We then examined how common post-processing operations affect these features and found they can be highly disruptive. Finally, we identified two notable properties of current generators: (i) they inherit and reproduce biases present in their training data; and (ii) although their outputs may appear indistinguishable to humans, they fail to match the mid-to-high-frequency spectral statistics of real images—differences that are particularly evident in the radial and angular spectra.

We then developed a simple yet effective detector that uses CLIP features to distinguish real from synthetic images. Our experiments show these features generalize far better than previously reported: using only a few samples from a single generator, a lightweight classifier can reach state-of-the-art performance and generalize broadly, even in challenging settings. Dataset diversity matters more than size. Larger CLIP pretrain-

ing yields more descriptive features and better results. We also find CLIP features are largely orthogonal to the low-level cues used by traditional detectors, making CLIP-based methods robust to attacks that target low-level artifacts and enabling further gains through fusion with conventional detectors.

We next turned to synthetic-video detection. Like images, generated videos exhibit artifacts in both space and time. We find that horizontal and vertical frequency components are especially fragile (degraded by block-based video compression), while diagonal components are more resilient and consistently present across models. Rather than introducing a new architecture, we propose an augmentation-based training strategy: first, remove semantic bias by creating fake videos via artifact injection into real footage; then, use wavelet-based augmentations that swap the low, horizontal, and vertical sub-bands with those from real videos, driving detectors to focus on diagonal frequencies. Experiments show this simple training paradigm outperforms state-of-the-art methods, even when trained on a single generator, highlighting the value of better training over added model complexity.

Bibliography

- [1] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.
- [2] Michael Albright and Scott McCloskey. Source Generator Attribution via Inversion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.
- [3] Roberto Amoroso, Davide Morelli, Marcella Cornia, Lorenzo Baraldi, Alberto Del Bimbo, and Rita Cucchiara. Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.
- [4] Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. Reverse engineering of generative models: Inferring model hyperparameters from generated images. *IEEE transactions on pattern analysis and machine intelligence*, 45(12):15477–15493, 2023.
- [5] Jianfa Bai, Man Lin, Gang Cao, and Zijie Lou. Ai-generated video detection via spatial-temporal anomaly learning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 2024.
- [6] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1728–1738, 2021.
- [7] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-Image

- Diffusion Models with an Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324*, 2022.
- [8] Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 2023.
- [9] Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, et al. Identifying and Mitigating the Security Risks of Generative AI. *Foundations and Trends® in Privacy and Security*, 6(1):1–52, 2023.
- [10] Jordan J Bird and Ahmad Lotfi. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *IEEE Access*, 2024.
- [11] Matyáš Boháček and Hany Farid. A geometric and photometric exploration of gan and diffusion synthesized faces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 874–883, 2023.
- [12] Matyáš Boháček and Hany Farid. Human Action CLIPS: Detecting AI-generated Human Motion. *arXiv preprint arXiv:2412.00526v1*, 2024.
- [13] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations (ICLR)*, 2018.
- [14] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2023.
- [15] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. <https://openai.com/research/video-generation-models-as-world-simulators>, 2024.
- [16] Tu Bui, Ning Yu, and John Collomosse. Repmix: Representation mixing for robust attribution of synthesized images. In *European Conference on Computer Vision (ECCV)*, pages 146–163, 2022.
- [17] Geoffrey J Burton and Ian R Moorhead. Color and spatial structure in natural scenes. *Applied Optics*, 26:157–170, 1987.
- [18] Giuseppe Cattaneo and Gianluca Roscigno. A possible pitfall in the experimental analysis of tampering detection algorithms. In *International Conference on Network-Based Information Systems*, 2014.
-

-
- [19] George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. FakeInversion: Learning to Detect Images from Unseen Text-to-Image Models by Inverting Stable Diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10759–10769, June 2024.
 - [20] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What Makes Fake Images Detectable? Understanding Properties that Generalize. In *European Conference on Computer Vision (ECCV)*, pages 103–120, 2020.
 - [21] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, 2022.
 - [22] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Alexander Binder, and Ngai-Man Cheung. Discovering Transferable Forensic Features for CNN-Generated Images Detection. In *European Conference on Computer Vision (ECCV)*, pages 671–689, 2022.
 - [23] Chirui Chang, Zhengzhe Liu, Xiaoyang Lyu, and Xiaojuan Qi. What matters in detecting ai-generated videos like sora? *arXiv preprint arXiv:2406.19568*, 2024.
 - [24] Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. DeMamba: AI-Generated Video Detection on Million-Scale GenVideo Benchmark. *arXiv preprint arXiv:2405.19707*, 2024.
 - [25] Mo Chen, Jessica Fridrich, Miroslav Goljan, and Jan Lukás. Determining image origin and integrity using sensor noise. *IEEE Transactions on information forensics and security*, 3(1):74–90, 2008.
 - [26] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
 - [27] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake video detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1133–1143, 2024.
-

-
- [28] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797, 2018.
 - [29] Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. FIRE: Robust Detection of Diffusion-Generated Images via Frequency-Guided Reconstruction Error. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12830–12839, 2025.
 - [30] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE transactions on pattern analysis and machine intelligence*, 2020.
 - [31] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.
 - [32] Davide Cozzolino, Justus Thies, Andreas Rossler, Matthias Nießner, and Luisa Verdoliva. SpoC: Spoofing Camera Fingerprints. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 990–1000, 2021.
 - [33] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018.
 - [34] Davide Cozzolino and Luisa Verdoliva. Noiseprint: A CNN-based camera model fingerprint. *IEEE Transactions on information forensics and security*, 15(1):14–27, 2020.
 - [35] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. RAISE: A Raw Images Dataset for Digital Image Forensics. In *ACM multimedia systems conference*, page 219–224, 2015.
 - [36] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. In *International Conference on Learning Representations (ICLR)*, 2024.
 - [37] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khắc, Luke Melas, and Ritobrata Ghosh. DALL-E Mini. <https://github.com/borisdyma/dalle-mini>, 2021.
 - [38] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *International Conference on Learning Representations (ICLR)*, 2025.
-

-
- [39] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.
 - [40] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems (NeurIPS)*, 34:8780–8794, 2021.
 - [41] Yuzhen Ding, Nupur Thakur, and Baoxin Li. Does a GAN leave distinct model-specific fingerprints? In *BMVC*, 2021.
 - [42] Pantelis Dogoulis, Giorgos Kordopatis-Zilos, Ioannis Kompatsiaris, and Symeon Papadopoulos. Improving Synthetically Generated Image Detection in Cross-Concept Settings. In *ACM International Workshop on Multimedia AI against Disinformation (MAD)*, page 28–35, 2023.
 - [43] Mengnan Du, Shiva Pentiyala, Yuening Li, and Xia Hu. Towards Generalizable Deepfake Detection with Locality-Aware AutoEncoder. In *ACM international conference on information knowledge management (CIKM)*, page 325–334, 2020.
 - [44] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7890–7899, 2020.
 - [45] Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. In *Advances in neural information processing systems (NeurIPS)*, 2020.
 - [46] David C. Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online Detection of AI-Generated Images. In *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 382–392, 2023.
 - [47] Ziv Epstein, Aaron Hertzmann, Laura Herman, Robert Mahari, Morgan R Frank, Matthew Groh, et al. Art and the science of generative AI: A deeper dive. *Science*, 2023.
 - [48] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pretrained model clip. In *AAAI Conference on Artificial Intelligence*, pages 6568–6576, 2022.
 - [49] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International conference on machine learning (ICML)*, 2024.
-

-
- [50] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021.
 - [51] Shengbang Fang, Tai D Nguyen, and Matthew C Stamm. Open set synthetic image source attribution. In *BMVC*, 2023.
 - [52] Hany Farid. Lighting (in)consistency of paint by text. *arXiv preprint arXiv:2207.13744*, 2022.
 - [53] Hany Farid. Perspective (in)consistency of paint by text. *arXiv preprint arXiv:2206.14617*, 2022.
 - [54] Adobe Firefly. <https://www.adobe.com/sensei/generative-ai/firefly.html>, 2023.
 - [55] Flux-AI. Flux. <https://flux-ai.io/flux-video-ai>, 2024.
 - [56] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *International conference on machine learning (ICML)*, pages 3247–3258, 2020.
 - [57] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. In *Advances in neural information processing systems (NeurIPS)*, volume 36, pages 27092–27112, 2023.
 - [58] Sharath Girish, Saksham Suri, Sai Saketh Rambhatla, and Abhinav Shrivastava. Towards discovery and attribution of open-world gan generated images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14094–14103, 2021.
 - [59] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. *Advances in neural information processing systems (NIPS)*, 27, 2014.
 - [60] Diego Gagnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021.
 - [61] Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or JPEG? Revealing Common Biases in Generated Image Detection Datasets. 2024.
-

-
- [62] Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A Bias-Free Training Paradigm for More General AI-generated Image Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
 - [63] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
 - [64] Syou Hirofumi, Kazuto Fukuchi, Yohei Akimoto, and Jun Sakuma. Did you use my GAN to generate fake? post-hoc attribution of GAN images via latent recovery. In *International Joint Conference on Neural Networks (IJCNN)*, 2022.
 - [65] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems (NeurIPS)*, 33:6840–6851, 2020.
 - [66] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
 - [67] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. https://github.com/mlfoundations/open_clip, 2021.
 - [68] Hyeonseong Jeon, Young Oh Bang, Junyaup Kim, and Simon Woo. T-GD: Transferable GAN-generated Images Detection Framework. In *International conference on machine learning (ICML)*, volume 119, pages 4746–4761, 2020.
 - [69] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, Pyounggeon Kim, and Jongwon Choi. FingerprintNet: Synthesized Fingerprints for Generated Image Detection. In *European Conference on Computer Vision (ECCV)*, pages 76–94, 2022.
 - [70] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations (ICLR)*, 2025.
-

-
- [71] Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing Global and Local Features for Generalized AI-Synthesized Image Detection. In *IEEE International Conference on Image Processing (ICIP)*, pages 3465–3469, 2022.
 - [72] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10124–10134, 2023.
 - [73] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations (ICLR)*, 2018.
 - [74] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Advances in neural information processing systems (NeurIPS)*, volume 34, pages 852–863, 2021.
 - [75] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.
 - [76] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
 - [77] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
 - [78] Misha Konstantinov, Alex Shonenkov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. <https://www.deepfloyd.ai/deepfloyd-if>, 2023.
 - [79] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision (ECCV)*, 2024.
 - [80] Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, and Amit K Roy-Chowdhury. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated con-
-

- tent. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [81] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International conference on machine learning (ICML)*, pages 12888–12900, 2022.
- [82] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.
- [83] Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. Detecting multimedia generated by large ai models: A survey. *arXiv preprint arXiv:2204.06125*, 2024.
- [84] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.
- [85] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European Conference on Computer Vision (ECCV)*, pages 95–110, 2022.
- [86] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [87] Qingyuan Liu, Pengyuan Shi, Yun-Yun Tsai, Chengzhi Mao, and Junfeng Yang. Turns Out I’m Not Real: Towards Robust Detection of AI-Generated Videos. *arXiv preprint arXiv:2406.09601*, 2024.
- [88] Zhengzhe Liu, Xiaojuan Qi, and Philip H. S. Torr. Global texture enhancement for fake face detection in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [89] Zhuang Liu and Kaiming He. A Decade’s Battle on Dataset Bias: Are We There Yet? 2025.
- [90] Jan Lukáš, Jessica Fridrich, and Miroslav Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on information forensics and security*, 1(2):205–214, 2006.
-

-
- [91] Long Ma, Jiajia Zhang, Hongping Deng, Ningyu Zhang, Qinglang Guo, Haiyang Yu, Yong Liao, and Pengyuan Zhou. DeCoF: Generated Video Detection via Frame Consistency. *arXiv preprint arXiv:2402.02085v1*, 2024.
 - [92] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. Detecting GAN-generated Images by Orthogonal Training of Multiple CNNs. In *IEEE International Conference on Image Processing (ICIP)*, 2022.
 - [93] Francesco Marra, Diego Gagnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of GAN-generated fake images over social networks. In *IEEE conference on multimedia information processing and retrieval (MIPR)*, 2018.
 - [94] Francesco Marra, Diego Gagnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs Leave Artificial Fingerprints? In *IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 506–511, 2019.
 - [95] Francesco Marra, Cristiano Saltori, Giulia Boato, and Luisa Verdoliva. Incremental learning for the detection and classification of GAN-generated images. In *IEEE international workshop on information forensics and security (WIFS)*, pages 1–6, 2019.
 - [96] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2019.
 - [97] Scott McCloskey and Michael Albright. Detecting GAN-Generated Imagery using Saturation Cues. In *IEEE International Conference on Image Processing (ICIP)*, 2019.
 - [98] Midjourney. <https://www.midjourney.com/home>, 2023.
 - [99] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyu Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in neural information processing systems (NIPS)*, pages 35087–35102, 2022.
 - [100] Harvey B Mitchell and Paul A Schaefer. A “soft” K-nearest neighbor voting scheme. *International journal of intelligent systems*, 16(4):459–468, 2001.
 - [101] Daniel Mas Montserrat, Hanxiang Hao, Sri K. Yarlagadda, Sriram Baireddy, Ruiting Shao, Janos Horvath, Emily Bartusiak, Justin Yang, David Gùera, Fengqing Zhu, and Edward J. Delp. Deepfakes detection with automatic face weighting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.
-

-
- [102] John Mullan, Duncan Crawbuck, and Aakash Sastry. Hotshot-XL. <https://github.com/hotshotco/hotshot-xl>, 2023.
 - [103] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diff. Models. In *International conference on machine learning (ICML)*, pages 16784–16804, 2022.
 - [104] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24480–24489, 2023.
 - [105] OpenAI. <https://openai.com/dall-e-3>, 2023.
 - [106] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research Journal*, 2024.
 - [107] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *AAAI Conference on Artificial Intelligence*, 29(1), February 2015.
 - [108] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346, 2019.
 - [109] Carolyn Pe Rosiene and Truong Q. Nguyen. Tensor-product wavelet vs. Mallat decomposition: a comparative analysis. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, volume 3, pages 431–434, 1999.
 - [110] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023.
 - [111] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *International Conference on Learning Representations (ICLR)*, 2024.
 - [112] Piotr Porwik and Agnieszka Lisowska. The Haar-wavelet transform in digital image processing: its status and achievements. *Machine graphics and vision*, 13(1/2):79–98, 2004.
-

-
- [113] Ekta Prashnani, Koki Nagano, Shalini De Mello, David Luebke, and Orazio Gallo. Avatar Fingerprinting for Authorized Use of Synthetic Talking-Head Videos. In *European Conference on Computer Vision (ECCV)*, pages 209–228, 2024.
 - [114] Joaquin Quiñonero-Candela, Carl Edward Rasmussen, Fabian Sinz, Olivier Bousquet, and Bernhard Schölkopf. Evaluating predictive uncertainty challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d’Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 1–27. Springer Berlin Heidelberg, 2006.
 - [115] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pages 8748–8763, 2021.
 - [116] Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. On the effectiveness of dataset alignment for fake image detection. In *International Conference on Learning Representations (ICLR)*, 2025.
 - [117] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
 - [118] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. In *VISAPP*, 2024.
 - [119] Andreas Roessler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
 - [120] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.
 - [121] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. Stable diffusion. <https://github.com/CompVis/stable-diffusion>, 2022.
 - [122] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. <https://github.com/Stability-AI/stablediffusion>, 2022.
-

-
- [123] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles. In *International conference on machine learning (ICML)*, pages 29441–29454, 2023.
- [124] SAFE. Synthetic Video Detection Challenge. <https://safe-video-2025.dsri.org/>, 2025.
- [125] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad. Shadows Don’t Lie and Lines Can’t Bend! Generative Models don’t know Projective Geometry... for now. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [126] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis. In *International conference on machine learning (ICML)*, 2023.
- [127] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, and Mitchell Wortsman et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in neural information processing systems (NIPS)*, volume 35, pages 25278–25294, 2022.
- [128] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *Advances in neural information processing systems (NeurIPS) Workshops*, 2021.
- [129] Katja Schwarz, Yiyi Liao, and Andreas Geiger. On the Frequency Bias of Generative Models. In *Advances in neural information processing systems (NeurIPS)*, 2021.
- [130] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models. In *ACM SIGSAC conference on computer and communications security*, page 3418–3432, 2023.
- [131] Sergey Sinitisa and Ohad Fried. Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.
-

-
- [132] Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. On learning multi-modal forgery representation for diffusion generated video detection. In *Advances in neural information processing systems (NeurIPS)*, 2024.
 - [133] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.
 - [134] Eugene Switkes, Melanie J Mayer, and Jeffrey A Sloan. Spatial frequency analysis of the visual environment: anisotropy and the carpentered environment hypothesis. *Vision Research*, 18:1393–1399, 1978.
 - [135] Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
 - [136] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Domain Learning. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 5052–5060, 2024.
 - [137] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12105–12114, June 2023.
 - [138] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. Galip: Generative adversarial clips for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14214–14223, 2023.
 - [139] Genmo Team. Mochi 1. <https://github.com/genmoai/models>, 2024.
 - [140] David J Tolhurst, Yoav Tadmor, and Tang Chao. The amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12:229–232, 1992.
 - [141] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
 - [142] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14:391–412, 2003.
-

-
- [143] Danial Samadi Vahdati, Tai D Nguyen, Aref Azizpour, and Matthew C Stamm. Beyond deepfake images: Detecting ai-generated videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4397–4408, 2024.
 - [144] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *International Conference on Learning Representations (ICLR)*, 2022.
 - [145] Vladan Velisavljevic, Baltasar Beferull-Lozano, Martin Vetterli, and Pier Luigi Dragotti. Directionlets: anisotropic multidirectional representation with separable filtering. *IEEE transactions on image processing*, 15(7):1916–1933, 2006.
 - [146] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.
 - [147] Jun Wang, Omran Alamayreh, Benedetta Tondi, and Mauro Barni. Open set classification of gan-based image manipulations via a vit-based hybrid architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2023.
 - [148] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot... for now. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8692–8701, 2020.
 - [149] Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models. *arXiv preprint arXiv:2410.02757*, 2024.
 - [150] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for Diffusion-Generated Image Detection. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22445–22455, 2023.
 - [151] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. In *International Conference on Learning Representations (ICLR)*, 2023.
 - [152] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision (ECCV)*, pages 736–753, 2022.
-

-
- [153] Tianyun Yang, Ziyao Huang, Juan Cao, Lei Li, and Xirong Li. Deepfake network architecture attribution. In *AAAI Conference on Artificial Intelligence*, pages 4662–4670, 2022.
 - [154] Tianyun Yang, Danding Wang, Fan Tang, Xinying Zhao, Juan Cao, and Sheng Tang. Progressive open space expansion for open-set model attribution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15856–15865, 2023.
 - [155] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. 2025.
 - [156] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.
 - [157] Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, et al. Language Model Beats Diffusion - Tokenizer is key to visual generation. In *International Conference on Learning Representations (ICLR)*, 2024.
 - [158] Ning Yu, Larry Davis, and Mario Fritz. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7555–7565, 2019.
 - [159] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
 - [160] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019.
 - [161] Baiwu Zhang, Jin Peng Zhou, Ilia Shumailov, and Nicolas Papernot. On attribution of deepfakes. *arXiv preprint arXiv:2008.09194*, 2020.
 - [162] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations (ICLR)*, 2018.
-

-
- [163] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017.
 - [164] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free Adaption of CLIP for Few-shot Classification. In *European Conference on Computer Vision (ECCV)*, pages 493–510, 2022.
 - [165] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *IEEE international workshop on information forensics and security (WIFS)*, pages 1–6, 2019.
 - [166] Yichi Zhang and Xiaogang Xu. Diffusion noise feature: Accurate and fast generated image detection. *arXiv preprint arXiv:2312.02625v1*, 2023.
 - [167] Nan Zhong, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. *arXiv preprint arXiv:2311.12397v1*, 2023.
 - [168] Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model. *arXiv preprint arXiv:2410.15458*, 2024.
 - [169] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zeg-CLIP: Towards Adapting CLIP for Zero-shot Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11175–11185, 2023.
 - [170] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.
 - [171] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in neural information processing systems (NeurIPS)*, 36, 2024.
-

Author's publications

1. R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, L. Verdoliva. "Intriguing properties of synthetic images: from generative adversarial networks to diffusion models" In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Workshop on Media Forensics at CVPR 2023 Vancouver, BC, Canada, 2023, pp. 973-982, DOI: 10.1109/CVPRW59228.2023.00104
2. R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, L. Verdoliva. "On the detection of synthetic images generated by diffusion models", In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023 pp. 1-5, IEEE, Rhodes Island, Greece, 2023, DOI: 10.1109/ICASSP49357.2023.10095167
3. G. Zingarini, D. Cozzolino, R. Corvi, G. Poggi, L. Verdoliva "M3Dsynth: A dataset of medical 3D images with AI-generated local manipulations", In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2024 Seoul, Republic of Korea, 2024 pp. 13176-13180, DOI: 10.1109/ICASSP48485.2024.10446605
4. D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, L. Verdoliva "Raising the Bar of AI-generated Image Detection with CLIP" In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Workshop on Media Forensics at CVPR 2024 Seattle, WA, USA, pp. 4356-4366, DOI: 10.1109/CVPRW63382.2024.00439
5. D. Tariang, R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, L. Verdoliva "Synthetic Image Verification in the Era of Generative Artificial Intelligence: What Works and What Isn't There yet" In: IEEE Security & Privacy, Volume: 22, Issue: 3, May-June 2024, DOI: 10.1109/MSEC.2024.3376637

6. R. Corvi, D. Cozzolino, E. Prashnani, S. De Mello, K. Nagano, L. Verdoliva “Seeing What Matters: Generalizable AI-generated Video Detection with Forensic-Oriented Augmentation” In: Annual Conference on Neural Information Processing Systems (NeurIPS), San Diego, CA, United States of America December 2025
 7. O. Papadopoulou, M. Schinas, R. Corvi, D. Karageorgiou, C. Koutlis, F. Guillaro, E. Gavves, H. Mareen, L. Verdoliva, S. Papadopoulos “Synthetic Images at MediaEval 2025: Advancing detection of generative AI in real-world online images” In: MediaEval 2025 Workshop, Dublin Ireland, 25-26 October 2025
-